Prévision à court terme des flux de voyageurs : une approche par les réseaux bayésiens

Jérémy Roos

▶ **To cite this version:**

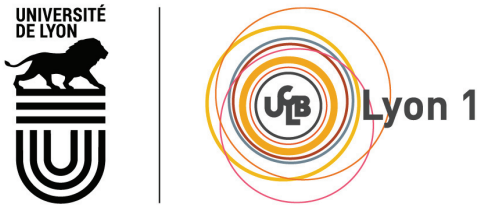
Jérémy Roos. Prévision à court terme des flux de voyageurs : une approche par les réseaux bayésiens. Algorithme et structure de données [cs.DS]. Université de Lyon, 2018. Français. NNT : 2018LYSE1170 . tel-01943718

HAL Id: tel-01943718
https://theses.hal.science/tel-01943718

Submitted on 4 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



N° d'ordre NNT : 2018LYSE1170

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale ED 512
Informatique et Mathématiques de Lyon

Spécialité de doctorat : Informatique

Soutenue publiquement le 28/09/2018, par :
Jérémy ROOS

Prévision à court terme des flux de voyageurs : une approche par les réseaux bayésiens

---

Devant le jury composé de :

Alexandre AUSSEM, Professeur des universités, Université Lyon 1	Président
Philippe LERAY, Professeur des universités, Université de Nantes	Rapporteur
Latifa OUKHELLOU, Directrice de recherche IFSTTAR, Université Paris-Est	Rapporteure
Shadi SADEGHIAN, Responsable du management de l'innovation, RATP	Examinatrice
Stéphane BONNEVAY, Maître de conférences HDR, Université Lyon 1	Directeur de thèse
Gérald GAVIN, Maître de conférences, Université Lyon 1	Co-directeur de thèse
Nathalie LAURENT, Responsable d'unité spécialisée, RATP	Encadrante industrielle

Remerciements

En premier lieu, je souhaite exprimer ma profonde reconnaissance à M. Stéphane Bonnevay pour avoir accepté de diriger mes recherches. Durant ces quatre années de thèse, il a su m'enseigner la rigueur scientifique et le mode de réflexion indispensables à l'exercice du métier de chercheur. Sa disponibilité et son engagement indéfectibles m'ont été précieux tout au long de ce projet.

Je tiens également à exprimer ma gratitude à mon co-directeur de thèse, M. Gérald Gavin, pour son soutien et son implication dans ces travaux. Ses relectures attentives, ses remarques constructives ainsi que les débats passionnés (et passionnants !) qui ont animé nos réunions de travail ont grandement contribué à l'amélioration de ce manuscrit.

Je remercie ensuite chaleureusement Mme Véronique Fontalirant-Baptiste, qui m'a autorisé à monter ce projet et n'a eu de cesse de le valoriser au sein de la RATP. Son exigence et la confiance qu'elle m'a accordée dès mes débuts dans l'entreprise ont été extrêmement formatrices. Elles m'ont permis d'évoluer peu à peu du statut d'« étudiant » à celui de « professionnel ».

Mes sincères remerciements s'adressent également à Mme Nathalie Laurent, qui a accepté de reprendre l'encadrement industriel de ce projet malgré sa complexité et l'investissement qu'il demandait. L'intérêt qu'elle a porté à mes travaux, le temps et les moyens qu'elle leur a consacrés m'ont permis de poursuivre cette thèse dans les meilleures conditions.

Je tiens ensuite à remercier M. Philippe Leray et Mme Latifa Oukhellou pour avoir accepté d'être les rapporteurs de ce mémoire et pour m'avoir fait part, au cours de nos échanges, de leurs conseils avisés. Mes remerciements vont également à M. Alexandre Aussem et Mme Shadi Sadeghian, qui m'ont fait l'honneur de participer à mon jury de thèse.

Ces travaux n'auraient pu aboutir sans la contribution et le soutien d'un grand nombre de personnes. Mes remerciements chaleureux s'adressent notamment à...

... Mme Claire Spitzmuller, qui m'a recruté dans son équipe il y a cinq ans et

qui, par sa disponibilité et son accueil toujours bienveillants, a grandement facilité mon intégration à la RATP.

... M. Vincent Leblond pour son appui scientifique et les réflexions passionnantes que nous avons eues autour de la donnée.

... MM. Denis Sochon, Antonio Galante et Mme Xiaoning Yang, qui ont supervisé l'avancement de cette thèse et m'ont épaulé administrativement tout au long du projet.

... MM. Stéphane Amirouche, Christian Durieux et François Heitzmann, fins connaisseurs et précieux pourvoyeurs de données.

... M. Patrick Grascœur pour sa maîtrise des procédures d'inscription en conférence, mais aussi et surtout pour son infinie patience.

... M. Mathieu Bordes pour sa maîtrise de l'anglais et, désormais, des relectures d'articles scientifiques.

... Aurèle, Bérénice, Camille, Cécile, David, Dione, Élise, Fabien, Françoise, Mélina, Nicolas, Olivier M., Olivier R., Pierre, Pierrette, Rhariba, Risienne... et tous les collègues du département CML, avec qui j'ai partagé tant de bons moments et qui m'ont permis, ne serait-ce qu'un instant, d'oublier les difficultés de mon travail.

... MM. Saïd Lahboub, Laurent Drouin et toute l'équipe du Datalab pour leur accueil chaleureux et leurs précieux conseils lors de la préparation de ma soutenance.

Parce qu'un doctorat est une aventure parfois solitaire, je remercie les personnes de mon entourage qui m'ont accompagné et encouragé durant ces quatre années. Merci notamment à Supakanya qui, en plus de sa propre thèse, a réussi à supporter la mienne ! Merci également à mes parents et à mon frère pour leur hébergement lors de mes nombreux séjours à Lyon et, surtout, pour l'affection qu'ils m'ont toujours témoignée.

Enfin, je ne saurais conclure ces remerciements sans rendre hommage à MM. Massal et Thomasset, qui m'ont si bien transmis leur passion des mathématiques lorsque j'étais encore tout petit. :-)

Jérémy Roos

Résumé

La Régie Autonome des Transports Parisiens (RATP) est le principal opérateur de transport public de la région Île-de-France. Son objectif est d'assurer une qualité et une régularité de service constantes, tout en préservant la sécurité et le confort des voyageurs qui fréquentent son réseau. Cette excellence opérationnelle passe par une nécessaire maîtrise des risques de congestion, qui se traduit par la capacité à connaître et anticiper la demande tout au long du service.

Dans ces travaux de thèse, nous proposons un modèle de prévision à court terme des flux de voyageurs basé sur les réseaux bayésiens. Ce modèle est destiné à répondre à des besoins opérationnels divers liés à l'information voyageurs, la régulation des flux ou encore la planification de l'offre de transport. Conçu pour s'adapter à tout type de configuration spatiale, il permet de combiner des sources de données hétérogènes et fournit une représentation intuitive des relations de causalité spatio-temporelles entre les flux. Sa capacité à gérer les données manquantes lui permet de réaliser des prédictions en temps réel même en cas de défaillances techniques ou d'absences de systèmes de collecte.

Après un état de l'art des méthodes de prévision des flux à court terme, nous présentons les réseaux bayésiens et leur extension temporelle (les réseaux bayésiens dynamiques), ainsi que les algorithmes d'apprentissage et d'inférence qui leur sont associés. Nous nous intéressons plus particulièrement aux réseaux bayésiens à mélanges gaussiens, dont la flexibilité permet d'approximer une grande variété de distributions et de représenter le comportement non linéaire des flux. Dans ce contexte, nous développons un algorithme d'apprentissage permettant d'optimiser automatiquement le nombre de composantes des modèles de mélanges gaussiens.

Collectées sur le réseau ferré de la RATP, les données utilisées dans notre étude sont issues de trois sources différentes : les validations des titres de transport, les comptages par pesée des voyageurs à bord des trains et l'offre de transport. Ces données sont combinées par le biais d'un référentiel spatial commun défini à partir d'une description topologique du réseau de transport. La structure de notre modèle

dérive alors directement de leurs relations spatiales. Issue de la connaissance experte du fonctionnement du réseau, cette structure repose sur les dépendances causales entre les flux de voyageurs et leur voisinage spatio-temporel. La prise en compte de l'offre de transport, par l'intermédiaire des intervalles de départ des trains, constitue l'un des aspects novateurs de notre méthode.

En appliquant notre approche par les réseaux bayésiens à la ligne 2 du métro de Paris, les résultats obtenus surpassent ceux fournis par d'autres méthodes de prédiction. Ils témoignent notamment de la supériorité des réseaux bayésiens à mélanges gaussiens par rapport à l'utilisation de simples distributions gaussiennes. Ils illustrent également l'intérêt d'exploiter le voisinage spatio-temporel des flux, mais aussi et surtout le rôle fondamental de l'offre de transport.

Abstract

The Régie Autonome des Transports Parisiens (RATP) is the main public transport operator of the Île-de-France region. Its objective is to ensure consistent quality and regularity of service, while preserving the safety and comfort of the passengers who use its network. This operational excellence requires controlling the risks of congestion, which consists in knowing and anticipating demand all along service.

In this thesis, we propose a Bayesian network model for short-term passenger flow forecasting. This model is intended to cater for various operational needs related to passenger information, passenger flow regulation or operation planning. As well as adapting to any spatial configuration, it is designed to combine heterogeneous data sources and provides an intuitive representation of the causal spatio-temporal relationships between flows. Its ability to deal with missing data allows to make real-time predictions even in case of technical failures or absences of collection systems.

After a state of the art of short-term passenger flow forecasting, we present Bayesian networks and their temporal extension (dynamic Bayesian networks), with their associated learning and inference algorithms. We especially focus on Gaussian mixture Bayesian networks, whose flexibility allows to approximate a wide range of distributions and represent the nonlinear behaviour of flows. In this context, we develop a learning algorithm that automatically optimizes the number of components of Gaussian mixture models.

The data we use in our study are collected on the RATP urban rail network and come from three different sources: ticket validation, on-board counts by weighing systems, and transport service. These data are combined through a common spatial repository defined from a topological description of the transport network. The structure of our model directly derives from their spatial relationships. Using expert knowledge of the network, this structure is based on the causal dependencies between passenger flows and their spatio-temporal neighbourhood. The incorporation of transport service, through train departure intervals, is one of the innovative aspects of our method.

Applying our Bayesian network approach to Paris metro line 2, the obtained results outperform those provided by other prediction methods. They evidence the superiority of Gaussian mixture Bayesian networks compared to the use of simple Gaussian distributions. They also illustrate the interest of exploiting the spatio-temporal neighbourhood of flows, as well as the fundamental role of transport service.

Table des matières

Introduction

Contexte et problématique

Dans un contexte d'accroissement de la demande de déplacement en milieu urbain, les pouvoirs publics sont confrontés à une saturation des infrastructures de transport, impliquant de nombreux défis sociaux, économiques et environnementaux. De plus en plus présentes dans notre quotidien, les nouvelles technologies de l'information et de la communication jouent un rôle majeur dans la réponse à ces défis. Avec l'avènement des systèmes de transport intelligents, il est aujourd'hui possible de mieux maîtriser cette demande, en optimisant notamment la gestion et l'utilisation des infrastructures, en développant de nouveaux services destinés aux usagers ou en favorisant les reports vers des modes de transport durables.

La Régie Autonome des Transports Parisiens (RATP) est directement concernée par ces enjeux. Principal opérateur de transport public de la région Île-de-France, elle exploite les 16 lignes du métro de Paris, la majeure partie des lignes A et B du réseau express régional (RER) d'Île-de-France, 8 lignes de tramway et plus de 350 lignes de bus. Malgré la complexité de son réseau, elle doit être en mesure d'assurer une qualité et une régularité de service constantes, tout en préservant la sécurité et le confort des millions de voyageurs qui empruntent ses lignes chaque jour. Cette excellence opérationnelle passe par une nécessaire maîtrise des risques de congestion, qui se traduit par la capacité à connaître, mais aussi et surtout à anticiper la demande tout au long du service.

La RATP collecte une grande diversité de données sur la mobilité des voyageurs au sein de son réseau (validations des titres de transport, comptages, enquêtes origine-destination, etc.). Cependant, cette diversité est encore mal exploitée. Chaque source de données est généralement traitée de manière indépendante pour répondre à un besoin métier précis. Bien souvent, elle possède son propre référentiel, ce qui complique sa mise en relation avec les autres sources. Cette organisation des données « en silos », par ailleurs fréquente dans de nombreuses entreprises, est

préjudiciable à une connaissance fine de la mobilité. Dans un réseau de transport multimodal aussi vaste que celui d'Île-de-France, la demande est soumise à l'influence de nombreux facteurs internes et externes, d'où l'intérêt d'exploiter toute la richesse de l'information disponible.

Afin de prédire la demande de déplacement en Île-de-France, la RATP développe un certain nombre d'outils de modélisation, dont les plus connus sont sans doute les modèles GLOBAL et IMPACT (Leblond et Garcia Castello, 2016). Basés sur des données issues d'enquêtes de mobilité, ces outils sont conçus pour évaluer les effets à long terme de changements d'infrastructures ou de politiques de transport. N'étant pas destinés à la prévision en temps réel, ils ne prennent pas compte les conditions courantes du réseau. Par conséquent, ils demeurent insensibles aux phénomènes imprévus ou non récurrents qui sont susceptibles d'impacter la demande (aléas de l'offre de transport, indisponibilité temporaire de certaines infrastructures, événements générateurs d'affluence, etc.).

De ces observations émerge la question suivante : comment prédire la demande à court terme en tenant compte de la diversité des données collectées en temps réel sur le réseau ? Cette problématique peut être abordée sous plusieurs angles différents, selon l'objectif recherché et les sources de données disponibles. Ainsi, une première catégorie de travaux consiste à prédire des flux de voyageurs, c'est-à-dire le nombre de voyageurs transitant par des endroits précis du réseau (les entrées des stations, les départs des points d'arrêt, etc.). D'autres travaux ont pour but d'estimer des matrices origine-destination, c'est-à-dire le nombre de déplacements réalisés entre des couples de lieux distincts (Toqué *et al.*, 2016).

Dans le cadre de cette thèse, notre objectif est d'élaborer un modèle d'apprentissage automatique permettant de prédire à court terme les flux de voyageurs du réseau de la RATP. Pour que ce modèle soit applicable à grande échelle, il doit être capable de s'adapter à tout type de configuration spatiale, tout en proposant un formalisme suffisamment flexible pour intégrer des sources de données hétérogènes. Par ailleurs, il doit être robuste aux données manquantes, afin que les défaillances techniques ou les absences de systèmes de collecte ne compromettent pas la prédiction en temps réel.

Ces travaux de recherche ouvrent la voie à de nombreuses applications industrielles relatives à l'information voyageurs, la régulation des flux ou encore la planification de l'offre de transport. À titre d'exemple, les prédictions réalisées en temps réel peuvent être utilisées pour avertir les voyageurs des risques de congestion sur leur parcours et améliorer ainsi leurs conditions de confort. Elles peuvent également

permettre aux personnels en station de mieux anticiper la fréquentation de leurs espaces et de déployer des dispositifs de régulation adéquats. Enfin, dans le cadre de l'automatisation des lignes de transport, il est possible d'imaginer de futurs systèmes de pilotage capables d'adapter l'offre de transport à la demande en temps réel.

Démarche scientifique

Dans le chapitre 1 de cette thèse, nous présentons la démarche générale de prévision des flux à court terme, avant de réaliser un état de l'art des méthodes de prédiction existantes. Bien que nos travaux s'appliquent à un réseau de transport public, ce chapitre accorde une place importante aux méthodes développées dans le contexte routier, dont la littérature est beaucoup plus étendue.

Nous proposons d'aborder la problématique de la prévision à court terme des flux de voyageurs à travers une approche par les réseaux bayésiens. Issus du mariage de la théorie des graphes et de la théorie des probabilités, ces modèles permettent de représenter intuitivement des systèmes qui évoluent de manière non déterministe. Leur intérêt réside dans leur grande flexibilité, qui permet à l'expert de combiner des données de sources diverses, mais aussi d'incorporer sa propre connaissance du système étudié. De plus, leur mécanisme de transmission de l'information favorise le développement d'algorithmes d'inférence capables de prédire en présence de données incomplètes. Les réseaux bayésiens constituent donc des outils puissants pour la modélisation de problèmes industriels, comme en témoignent les travaux de Donat (2009) appliqués à la maintenance préventive du réseau de la RATP.

Les réseaux bayésiens sont présentés dans le chapitre 2, avec les algorithmes d'apprentissage et d'inférence qui leur sont associés. Nous nous intéressons notamment à deux types de modèles :

— les réseaux bayésiens gaussiens, qui reposent sur l'hypothèse de normalité des données et de linéarité des relations entre les variables ;
— les réseaux bayésiens à mélanges gaussiens, plus complexes à manipuler mais capables de modéliser des processus non linéaires.

Une partie importante de ce chapitre est également consacrée aux réseaux bayésiens dynamiques, qui étendent le formalisme des réseaux bayésiens à la représentation des relations temporelles entre les variables.

Les données utilisées dans le cadre de notre étude sont collectées sur le réseau ferré de la RATP (métro et RER) et issues de trois sources distinctes : les validations des titres de transport, les comptages par pesée des voyageurs à bord des

trains et l'offre de transport. Dans le chapitre 3, nous procédons tout d'abord à une description de ces données, ainsi qu'à la définition d'un référentiel spatial permettant de les associer. Dans un second temps, nous proposons un modèle de prévision à court terme des flux de voyageurs basé sur les réseaux bayésiens dynamiques. Déterminée à partir de considérations expertes sur le fonctionnement du réseau de transport, la structure de ce modèle repose sur les relations de causalité spatio-temporelles entre les flux. La prise en compte de l'offre de transport, dont l'impact sur les flux est mis en évidence dans une courte expérimentation sur la gare RER de Nanterre-Préfecture, constitue l'un des aspects novateurs de notre approche. Bien que le terrain d'étude se limite au réseau ferré, notre démarche de modélisation s'appuie sur des principes topologiques suffisamment génériques pour être transposés à d'autres types de réseaux.

Dans le chapitre 4, nous appliquons notre méthodologie à la prévision à court terme des flux de la ligne 2 du métro de Paris. Au cours de cette expérimentation, les distributions de probabilité locales du réseau bayésien dynamique sont d'abord décrites par de simples gaussiennes, avant d'être étendues aux modèles de mélanges gaussiens. Les performances de notre approche sont également comparées à celles obtenues par d'autres méthodes de prédiction, afin de mieux évaluer les bénéfices des principes de construction développés dans cette thèse.

Ces travaux sont réalisés dans le cadre d'une collaboration entre la RATP et le laboratoire Entrepôts, Représentation et Ingénierie des Connaissances (ERIC), rattaché à l'Université de Lyon. Ils font l'objet de plusieurs publications dans des actes de conférences en informatique (Roos *et al.*, 2016, 2017a,b), ainsi que dans une revue scientifique sur les transports (Roos *et al.*, 2017c).

Chapitre 1

Prévision des flux à court terme

La notion de prévision à court terme peut être définie comme la réalisation de prédictions à un horizon restreint (en principe de quelques secondes à quelques heures), à partir d'informations passées et actuelles. Dans le domaine des transports, cette question est le plus souvent abordée dans le contexte routier, pour la prévision de flux de véhicules. Couramment désigné sous le terme de prévision du « trafic » à court terme, c'est dans ce contexte que sont introduites la plupart des méthodes de prédiction. Par comparaison, il existe peu de travaux consacrés à la prévision à court terme des flux de voyageurs d'un réseau de transport public. Une revue rapide de ces travaux est effectuée dans la section 1.5. Dans le reste de ce chapitre, nous nous référons essentiellement à la littérature sur la prévision du trafic.

1.1 Démarche de prévision des flux à court terme

La démarche de prévision des flux à court terme est décrite dans la figure 1.1. En règle générale, le modèle est construit par apprentissage à partir d'un historique de données. Cette étape peut être réalisée en exploitant les caractéristiques spatiales du réseau de transport, mais aussi la connaissance que possède l'expert sur le fonctionnement de ce réseau. Une fois le modèle construit, la prédiction est opérée à partir des données collectées en temps réel. Ces données sont ensuite historisées afin de permettre la mise à jour ultérieure du modèle. À noter que si la prédiction doit obligatoirement être opérée en temps réel, ce n'est pas nécessairement le cas de l'apprentissage.

Dans les travaux sur la prévision du trafic à court terme, le niveau d'agrégation des données varie généralement de quelques secondes à une heure. L'horizon de prédiction est souvent limité à un seul pas de temps, bien que certains modèles

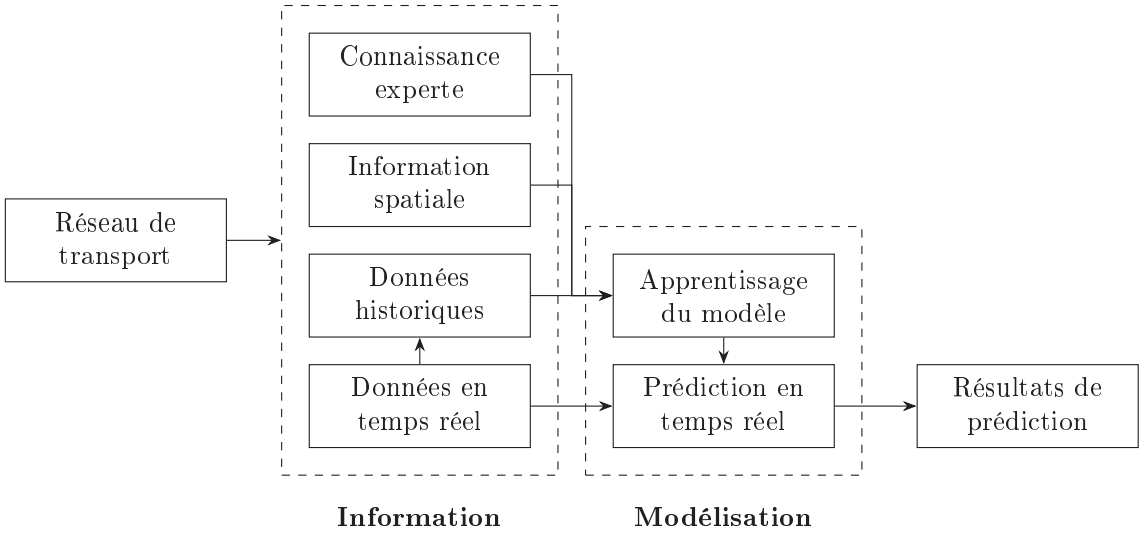


FIGURE 1.1 – Démarche de prévision des flux à court terme.

soient conçus pour prédire à un horizon plus lointain (Vlahogianni *et al.*, 2005 ; Min et Wynter, 2011). Si la notion de trafic fait instinctivement référence aux flux de véhicules, de plus en plus de réseaux sont équipés de systèmes permettant de mesurer d'autres types de variables (BITRE, 2014). Ainsi, la prévision de caractéristiques telles que le temps de parcours (van Lint, 2008 ; Fei *et al.*, 2011 ; Haworth et Cheng, 2012 ; Zheng et van Zuylen, 2013), la vitesse (Ishak et Alecsandru, 2004 ; Chandra et Al-Deek, 2009) ou la densité de véhicules (Quek *et al.*, 2006 ; Celikoglu, 2013) suscite un intérêt croissant. Par analogie avec l'objet de notre étude, nous nous concentrons principalement sur les méthodes de prévision des flux dans la suite de ce chapitre.

1.2 Panorama des méthodes de prédiction

Les premières recherches sur la prévision du trafic à court terme remontent à la fin des années 1970 (Ahmed et Cook, 1979). Aujourd'hui, la littérature du domaine est si vaste qu'il est difficile de recenser l'ensemble des travaux existants. C'est pourquoi notre objectif n'est pas de présenter une liste exhaustive de ces travaux, mais de citer les plus représentatifs d'entre eux. Pour des états de l'art plus complets, nous invitons le lecteur à se référer aux articles de Vlahogianni *et al.* (2004, 2014) ou à la thèse de Haworth (2014).

Bien qu'aucune typologie des méthodes de prédiction ne semble faire consensus, la plupart des auteurs tendent à distinguer les approches paramétriques des approches non paramétriques (Smith *et al.*, 2002 ; Vlahogianni *et al.*, 2004 ; van Hinsbergen *et al.*, 2007 ; van Lint et van Hinsbergen, 2012 ; Haworth, 2014). Un

modèle est paramétrique si sa structure (autrement dit son nombre de paramètres) est prédéfinie et que ses paramètres sont estimés à partir des données. À l'inverse, il est non paramétrique si sa structure doit être, au même titre que ses paramètres, déterminée à partir des données (van Hinsbergen *et al.*, 2007). Cette distinction n'est pas toujours évidente, étant donné qu'un même type de méthode peut être associé aux deux catégories. À titre d'exemple, un réseau bayésien dont la structure graphique et le type de distribution de probabilité sont fixés à l'avance constitue un modèle paramétrique. En revanche, s'il n'existe aucun apriori sur sa structure, alors il peut être assimilé à un modèle non paramétrique. Dans cette section, et afin de faciliter la lecture, nous traitons les réseaux bayésiens comme des modèles paramétriques, dans la mesure où des hypothèses fortes (liées à la nature du trafic) régissent la construction de leur structure. Au préalable, nous introduisons brièvement les méthodes naïves, pour lesquelles aucune structure ni paramètres ne sont déduits des données.

1.2.1 Méthodes naïves

Les méthodes naïves sont fréquemment utilisées en raison de leur facilité d'implémentation et de leur faible complexité en temps de calcul. Cependant, leurs performances de prédiction demeurent généralement limitées. Parmi ces méthodes, la marche aléatoire consiste à prédire les flux en se basant sur l'observation la plus récente. Exploitant uniquement les données collectées en temps réel, sa principale faiblesse réside dans son incapacité à prendre en compte les patterns historiques du trafic (Williams et Hoel, 2003).

La moyenne historique consiste à prédire les flux en calculant la moyenne de leurs valeurs passées aux pas de temps correspondants. À l'inverse de la marche aléatoire, cette méthode repose donc uniquement sur des données historiques. Son incapacité à prendre en compte les conditions courantes du trafic ne lui permet pas de réagir à des événements inattendus, tels que les incidents qui surviennent sur le réseau (Smith et Demetsky, 1997). Afin de pallier cette lacune, des données collectées en temps réel sont parfois introduites dans la prédiction, comme dans le cas du lissage exponentiel simple (Williams et Hoel, 2003).

1.2.2 Méthodes paramétriques

La structure d'un modèle paramétrique est définie à partir de considérations théoriques sur le trafic (van Lint et van Hinsbergen, 2012), incluant par exemple

de la connaissance experte sur les processus dynamiques qui régissent les flux. Une attention particulière doit cependant être accordée à la validité des hypothèses de départ, afin de ne pas dégrader la qualité de la prédiction.

Basés sur la nature stochastique du trafic (Vlahogianni *et al.*, 2004), les modèles autorégressifs à moyenne mobile intégrée (ARIMA) font partie des méthodes paramétriques les plus fréquemment appliquées à la prévision du trafic à court terme. Introduits par Ahmed et Cook (1979) et Levin et Tsao (1980), ils connaissent un essor important dans les années 1990 (Hamed *et al.*, 1995 ; Kirby *et al.*, 1997 ; Lee et Fambro, 1999). Cependant, Davis *et al.* (1990) mettent en évidence le fait que les prédictions se concentrent autour des valeurs moyennes et tendent à ignorer les valeurs extrêmes, ce qui représente un inconvénient majeur au regard des larges fluctuations du trafic. Afin d'améliorer les performances, un certain nombre de variantes sont proposées, telles que les modèles ARIMA saisonniers (SARIMA) (Williams *et al.*, 1998), ARIMA avec variables exogènes (ARIMAX) (Williams, 2001), ARMA vectoriels (VARMA), ARIMA spatio-temporels (STARIMA) (Kamarianakis et Prastacos, 2005), ou encore les modèles autorégressifs généralisés conditionnellement hétéroscédastiques (GARCH) (Kamarianakis *et al.*, 2005).

Une partie significative de la littérature est également consacrée aux modèles espace-état. Estimés à l'aide du filtre de Kalman (Kalman, 1960), leur capacité à représenter des processus physiques les rend particulièrement adaptés à la prévision du trafic à court terme. Initialement utilisés par Okutani et Stephanedes (1984), leur efficacité en contexte multivarié est mise en évidence par Whittaker *et al.* (1997), tandis que Stathopoulos et Karlaftis (2003) démontrent leur supériorité par rapport à de simples modèles ARIMA. Dans les travaux de Wang et Papageorgiou (2005), une extension du filtre de Kalman est proposée afin de permettre la modélisation de systèmes non linéaires.

Depuis quelques années, les réseaux bayésiens connaissent une popularité croissante dans la prévision du trafic à court terme. La structure graphique de ces modèles constitue un outil naturel pour la représentation des relations de dépendance (et d'indépendance) spatio-temporelles entre les flux. Leur intérêt réside également dans leur capacité à gérer les données manquantes (Whitlock et Queen, 2000 ; Sun *et al.*, 2006). Dans le cadre de leurs travaux, Whitlock et Queen (2000) et Queen et Albers (2009) développent un modèle dynamique multirégressif (Queen et Smith, 1993) qui permet de prendre en compte les changements soudains de l'état du trafic en procédant à des interventions externes. Dans l'approche de Sun *et al.* (2005, 2006), les relations statistiques entre les flux sont décrites par des modèles de mé-

langes gaussiens, dont la flexibilité permet d'approximer une grande variété de distributions. Zhu *et al.* (2016) utilisent quant à eux des réseaux bayésiens gaussiens et ajoutent des variables discrètes pour améliorer les performances de prédiction.

1.2.3 Méthodes non paramétriques

Dans le cas des méthodes non paramétriques, la structure du modèle est, au même titre que les paramètres, déterminée à partir des données. C'est pourquoi ces méthodes requièrent souvent davantage de données que les méthodes paramétriques. Leur flexibilité leur confère une meilleure capacité à modéliser les processus dynamiques non linéaires, sans qu'il soit nécessaire de connaître préalablement ces processus. En revanche, le fait que la structure dérive des données peut s'avérer problématique en cas d'événements non observés (van Hinsbergen *et al.*, 2007). Par ailleurs, la recherche du juste équilibre entre la complexité de la structure et la qualité de la prédiction constitue également un défi de l'apprentissage (van Lint et van Hinsbergen, 2012).

Basées sur la reconnaissance de formes, les méthodes de régression non paramétrique permettent de modéliser le comportement chaotique du trafic (Smith *et al.*, 2002). Elles ont l'avantage d'être rapides à implémenter et facilement compréhensibles par les praticiens (Clark, 2003). Appliquée depuis le début des années 1990 (Davis et Nihan, 1991), la méthode des k plus proches voisins offre de bonnes performances de prédiction, surpassant par exemple la moyenne historique, les modèles ARIMA et les réseaux de neurones dans les travaux de Smith et Demetsky (1997). Sa capacité à exploiter la nature multivariée du trafic est illustrée par Clark (2003). Plus récemment, certains auteurs s'intéressent également à la méthode de régression à noyau, avec des résultats prometteurs (Sun et Chen, 2008 ; Huang et Sun, 2013). Haworth et Cheng (2012) montrent toutefois que ce type d'approche tend à produire des résultats similaires à ceux des k plus proches voisins.

Les réseaux de neurones artificiels sont utilisés dans de nombreuses applications du domaine des transports (Dougherty, 1995). Ils constituent un outil puissant pour la modélisation des relations non linéaires en contexte multivarié, bien que leur aspect « boîte noire » rende leur interprétation difficile (Zhang *et al.*, 1998). Dans le cadre de la prévision du trafic à court terme, les perceptrons multicouches proposés à l'origine produisent parfois des résultats mitigés (Smith et Demetsky, 1994 ; Dougherty et Cobbett, 1997 ; Kirby *et al.*, 1997). Toutefois, la diversification des architectures et des méthodes d'apprentissage permet d'aboutir à des modèles beau-

coup plus performants, tels que les réseaux de neurones à fonction de base radiale (Park *et al.*, 1998), à logique floue (Yin *et al.*, 2002), à ondelettes (Jiang et Adeli, 2005), génétiquement optimisés (Vlahogianni *et al.*, 2005) ou combinés bayésiens (Zheng *et al.*, 2006). Avec l'avènement de l'apprentissage profond, ces modèles bénéficient aujourd'hui d'une grande popularité (Huang *et al.*, 2014 ; Lv *et al.*, 2015).

Si la régression non paramétrique et les réseaux de neurones sont largement appliqués à la prévision du trafic à court terme, une dernière méthode mérite d'être évoquée : la régression à vecteurs de support. Issue des travaux de Cortes et Vapnik (1995) sur les machines à vecteurs de support, cette méthode permet de pallier les problèmes de surapprentissage et de convergence vers des minimums locaux auxquels sont notamment confrontés les réseaux de neurones (Zhang et Xie, 2008). C'est pourquoi elle tend à surpasser ces derniers dans un certain nombre d'expérimentations (Zhang et Xie, 2008 ; Castro-Neto *et al.*, 2009 ; Hong, 2011).

1.2.4 Comparaison des méthodes de prédiction

Nous avons présenté un panorama des méthodes naïves, paramétriques et non paramétriques destinées à la prévision du trafic à court terme. Or comme le souligne Haworth (2014), il n'existe pas de consensus sur la méthode la plus adaptée. Ce constat vient du fait que les modèles élaborés sont de plus en plus sophistiqués et impliquent donc des temps d'implémentation plus longs. Par conséquent, les auteurs préfèrent comparer leur approche à des méthodes plus simples à mettre en œuvre. Karlaftis et Vlahogianni (2011) suggèrent ainsi que de nombreux travaux de comparaison sont biaisés, en particulier lorsqu'ils confrontent des modèles non linéaires complexes à des modèles linéaires basiques.

La difficulté à comparer les méthodes de prédiction existantes s'explique également par la disparité des jeux de données utilisés dans les différentes études. Chacun d'entre eux provient en effet d'un terrain d'application spécifique et possède ses propres caractéristiques (résolution spatiale et temporelle, niveau de bruit, etc.) (Haworth, 2014). En outre, les critères d'évaluation des performances ne sont pas nécessairement les mêmes d'une étude à l'autre (Vlahogianni *et al.*, 2004).

Haworth (2014) établit une matrice de confusion indiquant le nombre d'études comparatives dans lesquelles chaque type de méthode obtient de meilleures performances que les autres. Il en ressort une légère tendance des méthodes non linéaires à surpasser les méthodes linéaires. En outre, les réseaux de neurones obtiennent de très bonnes performances générales, bien que les méthodes à noyaux semblent

produire de meilleurs résultats.

Les travaux de Stathopoulos et Karlaftis (2003) et Vlahogianni *et al.* (2005, 2007) font partie des rares études à être conduites dans des conditions identiques, avec le même jeu de données issu du réseau routier d'Athènes. Les résultats de ces études montrent que le réseau de neurones modulaire génétiquement optimisé de Vlahogianni *et al.* (2007) obtient de meilleures performances que les modèles ARIMA et espace-état de Stathopoulos et Karlaftis (2003), ainsi que des autres réseaux de neurones développés par Vlahogianni *et al.* (2005).

Dans une expérimentation plus récente, Chen *et al.* (2012) comparent une grande variété de méthodes pour différents niveaux d'agrégation des données (de 3 à 15 minutes). Aux échelles temporelles les plus fines, les modèles ARIMA se révèlent supérieurs à la plupart des méthodes non linéaires. Lorsque l'échelle augmente, les méthodes des k plus proches voisins, des réseaux bayésiens ou encore de lissage exponentiel obtiennent les meilleurs résultats. Contrairement à beaucoup d'autres travaux, les réseaux de neurones et la régression à vecteurs de support affichent ici des performances limitées. Ces résultats illustrent bien l'absence d'existence d'une méthode de prédiction « idéale » et que le choix de celle-ci est étroitement lié à la nature des données utilisées.

1.3 Voisinage spatio-temporel

La notion de voisinage spatio-temporel fait référence à l'information spatialement et temporellement voisine utilisée pour prédire une quantité à un endroit et un instant donnés. La pertinence de cette information peut être déterminée à partir d'hypothèses sur les processus physiques du trafic ou à l'aide de mesures statistiques de dépendance (Haworth, 2014).

La plupart des approches de prévision du trafic à court terme traitent les flux comme des processus exclusivement temporels. Bien que la structure spatiale des données joue un rôle capital dans la prédiction, une minorité de méthodes prennent en compte cette information. Dans le modèle STARIMA, Kamarianakis et Prastacos (2005) utilisent une matrice de pondération spatiale estimée à partir des distances entre les différents points de collecte. Dans les modèles espace-état, la dynamique spatio-temporelle du trafic est traduite par les équations de transition d'état (Whittaker *et al.*, 1997). Dans les approches par les réseaux de neurones, le voisinage spatio-temporel peut être introduit en entrée du modèle (Yin *et al.*, 2002 ; Vlahogianni *et al.*, 2005, 2007 ; Sun *et al.*, 2012), voire explicitement représenté dans sa

structure interne (van Lint *et al.*, 2005).

La structure des réseaux bayésiens permet de décrire de manière intuitive l'évolution spatio-temporelle du trafic. Les relations de dépendance et d'indépendance conditionnelles entre les flux sont représentées par un graphe orienté sans circuit. Dans les travaux de Whitlock et Queen (2000) et Queen et Albers (2009), ce graphe est construit à partir des relations de causalité entre les flux adjacents. Sun *et al.* (2006) et Zhu *et al.* (2016) reprennent ce principe en introduisant un décalage temporel entre les flux. L'avantage de ces approches est que la structure du modèle dérive directement de la topologie du réseau de transport. Sun *et al.* (2005) proposent une méthodologie légèrement différente en sélectionnant les flux utilisés dans la prédiction à l'aide du coefficient de corrélation de Pearson. Au final, les réseaux bayésiens développés par Sun *et al.* (2005, 2006) se révèlent plus performants que les chaînes de Markov, qui exploitent uniquement les valeurs historiques des flux à prédire (Yu *et al.*, 2003).

1.4 Données manquantes

En situation réelle, les systèmes de collecte des données peuvent être sujets à des périodes d'inactivité, du fait par exemple de défaillances techniques. Les données recueillies sont donc souvent incomplètes, ce qui complexifie l'apprentissage du modèle, mais aussi et surtout son application en temps réel. Les données manquantes sont habituellement classées en trois catégories (Rubin, 1976 ; Little et Rubin, 1987) :

- les données manquantes complètement aléatoires (MCAR), dont la probabilité d'absence est indépendante des valeurs observées ou manquantes ;
- les données manquantes aléatoires (MAR), dont la probabilité d'absence dépend uniquement des valeurs observées ;
- les données manquantes non aléatoires (NMAR), dont la probabilité d'absence dépend des valeurs manquantes.

Les données MCAR sont aléatoirement dispersées dans le jeu de données et ne biaisent donc pas la prédiction. Les données MAR sont également faciles à traiter, dans la mesure où l'information observée est suffisante pour estimer leur distribution. En revanche, le cas des données NMAR est plus compliqué à appréhender car il entraîne la sous-représentation de certaines situations. Il est alors nécessaire d'incorporer de l'information externe pour revenir au cas MCAR ou MAR (Naïm *et al.*, 2011). En pratique, il est souvent difficile de prouver le caractère aléatoire ou non des données manquantes (celles-ci étant, par définition, non observées). Par souci

de simplification, de nombreux auteurs partent de l'hypothèse que ces données sont MCAR ou MAR.

Dans le contexte de la prévision du trafic, la majorité des méthodes d'imputation des données sont univariées. En d'autres termes, les valeurs manquantes d'un flux sont estimées uniquement à partir de ses valeurs observées. Zhong *et al.* (2004) comparent plusieurs approches basées sur des modèles factoriels, ARIMA, la régression localement pondérée et les réseaux de neurones à délais temporels. Qu *et al.* (2009) proposent quant à eux une méthode d'analyse en composantes principales probabiliste. Afin d'améliorer la robustesse des estimations, Ni *et al.* (2005) présentent une méthode d'imputation multiple consistant à réaliser plusieurs estimations pour chaque valeur manquante. Comme le soulignent Haworth et Cheng (2012), les méthodes univariées sont difficilement applicables aux longues séries de données manquantes. Une manière de résoudre ce problème est d'utiliser des approches multivariées, prenant notamment en compte l'information spatiale (Zhang et Liu, 2009 ; Li *et al.*, 2013a).

Malgré quelques exceptions (Chen *et al.*, 2003, 2012), peu de méthodes d'imputation sont conçues pour opérer en temps réel. De fait, une très grande majorité de modèles de prévision du trafic reposent sur l'hypothèse que les données sont complètes et sont donc mal équipés pour fonctionner en présence de données manquantes (Haworth, 2014). Certains modèles sont toutefois capables de gérer cette situation. Par exemple, l'approche par les réseaux de neurones espace-état de van Lint *et al.* (2005) intègre une couche de prétraitement combinant des méthodes d'interpolation spatiale et de lissage exponentiel. Haworth et Cheng (2012) utilisent un modèle de régression à noyau exploitant le voisinage spatio-temporel et permettant de prédire en présence de données NMAR. À noter que ces deux études ne visent pas à prédire des flux de trafic, mais des temps de parcours.

Les réseaux bayésiens se démarquent des autres modèles par leur capacité intrinsèque à gérer les données incomplètes. Dans le modèle dynamique multirégressif de Whitlock et Queen (2000), les valeurs manquantes sont estimées par inférence approchée à l'aide d'une méthode de Monte-Carlo par chaînes de Markov. Dans le réseau bayésien étendu de Sun *et al.* (2006), les flux manquants sont directement remplacés par leurs parents dans le graphe. Si cette méthode permet de conserver les relations de causalité entre les flux amont et aval, les algorithmes utilisés nécessitent l'observation complète des parents des flux écartés. Dans la mesure où les flux partiellement observés ne peuvent pas être intégrés dans la modélisation, cette contrainte peut potentiellement engendrer une perte significative d'information.

1.5 Flux de voyageurs

Tout au long de ce chapitre, la problématique de la prévision des flux à court terme a été exclusivement abordée dans le contexte routier. Or dans le cadre de ces travaux de thèse, nous nous intéressons aux flux de voyageurs d'un réseau de transport public. Si la plupart des approches que nous avons présentées peuvent être transposées à notre terrain d'étude, il convient de se pencher sur les méthodes spécifiquement développées pour ce type de réseau.

Il existe peu de travaux consacrés à la prévision à court terme des flux de voyageurs dans la littérature. Comme le soulignent Ma *et al.* (2014), les méthodes de prévision de la demande de transport public sont en majorité conçues pour la planification à long terme. Une présentation de ces méthodes peut être trouvée dans l'ouvrage de Ortúzar et Willumsen (2011). Le nombre limité de recherches sur l'aspect court terme s'explique peut-être par deux raisons. Tout d'abord, ce domaine d'application est relativement récent, la plupart des travaux ayant été réalisés au cours de la dernière décennie. Ensuite, de nombreux réseaux sont encore insuffisamment équipés de systèmes permettant de mesurer les flux de voyageurs en temps réel, ce qui entrave la mise en œuvre de tels modèles. Notons toutefois que ce champ de recherche est particulièrement actif en Chine, où plusieurs études sont notamment réalisées sur le métro de Pékin (Li *et al.*, 2013b ; Sun *et al.*, 2014, 2015 ; Jiao *et al.*, 2016). Sur le réseau de transport public d'Île-de-France, nous pouvons citer les travaux de Toqué *et al.* (2017), qui comparent les performances de réseaux de neurones récurrents à celles de forêts aléatoires pour la prévision de flux multimodaux.

Bien que peu répandues, les méthodes de prévision à court terme élaborées dans le contexte des transports publics connaissent une diversité comparable aux méthodes développées dans le contexte routier. Il est par exemple possible de trouver des approches basées sur les séries temporelles (Xue *et al.*, 2015), les modèles espace-état (Jiao *et al.*, 2016), la régression non paramétrique (Sun *et al.*, 2014) ou encore les machines à vecteurs de support (Chen *et al.*, 2011 ; Sun *et al.*, 2015). Toutefois, une certaine préférence semble être accordée aux approches basées sur les réseaux de neurones (Celikoglu et Cigizoglu, 2007 ; Wei et Chen, 2012 ; Li *et al.*, 2013b ; Zhang *et al.*, 2013 ; Toqué *et al.*, 2017).

Dans la plupart des travaux, les flux de voyageurs sont mesurés par le biais des validations des titres de transport. Outre l'information relative au flux à prédire, certains modèles prennent en compte le voisinage spatio-temporel, soit en intégrant directement les flux adjacents (Li *et al.*, 2013b), soit en sélectionnant les flux perti-

nents à l'aide de mesures de corrélation (Jiao *et al.*, 2016). Des données externes au réseau sont également exploitées, telles que les facteurs calendaires (Wei et Chen, 2012 ; Li *et al.*, 2013b ; Toqué *et al.*, 2017), les conditions météorologiques (Li *et al.*, 2013b) ou même les événements extraits des réseaux sociaux (Ni *et al.*, 2017). En l'état actuel de nos recherches, aucune méthode ne semble tenir compte de l'offre de transport fournie par l'opérateur de transport public. Pourtant, cette information joue un rôle capital dans la prédiction, comme nous allons le voir dans la suite de cette thèse.

Chapitre 2

Réseaux bayésiens

Dans le chapitre 1, nous avons mis en évidence la diversité des méthodes de prévision des flux à court terme. S'il n'existe pas de consensus sur la méthode la plus appropriée, nous avons vu que les réseaux bayésiens suscitent un intérêt croissant en raison de leur représentation intuitive des relations de causalité spatio-temporelles entre les flux, ainsi que de leur capacité intrinsèque à gérer les données manquantes. Dans ce chapitre, nous présentons les algorithmes d'apprentissage et d'inférence associés à ce formalisme. Notre étude se concentre sur deux types de réseaux bayésiens continus : les réseaux bayésiens gaussiens et les réseaux bayésiens à mélanges gaussiens. À partir de la section 2.6, nous nous intéressons plus spécifiquement aux réseaux bayésiens dynamiques, qui permettent de représenter des systèmes qui évoluent au cours du temps.

2.1 Introduction aux réseaux bayésiens

2.1.1 Modèles graphiques probabilistes

Dans la plupart des applications du monde réel, nous ne disposons que d'observations partielles du système que nous cherchons à modéliser. Notre connaissance limitée de ce système engendre une part d'incertitude qui doit être prise en compte dans le raisonnement. Situés au croisement de la théorie des graphes et de la théorie des probabilités, les modèles graphiques probabilistes constituent un ensemble d'outils permettant de raisonner dans l'incertain (Koller et Friedman, 2009). Ces modèles représentent les relations de dépendance et d'indépendance conditionnelle entre des variables aléatoires par un graphe, et traduisent ces relations de manière quantitative par une distribution de probabilité jointe entre les nœuds. Leur structure graphique fournit une interface intuitive qui facilite leur interprétation par l'utilisateur. Elle

est à la base d'une représentation compacte de la distribution jointe, sur laquelle reposent de nombreux algorithmes d'apprentissage et d'inférence. Enfin, la flexibilité de ce formalisme permet à un expert de combiner des informations de sources diverses, provenant aussi bien des données que de ses propres connaissances.

Il existe plusieurs classes de modèles graphiques probabilistes dans la littérature. Les plus étudiées sont sans doute les champs aléatoires de Markov, dont la structure est non orientée (Kindermann et Snell, 1980), et les réseaux bayésiens, que nous présentons dans ce chapitre. Introduits par Judea Pearl à la fin des années 1980 (Pearl, 1988), les réseaux bayésiens sont des modèles graphiques probabilistes dont la structure se caractérise par un graphe orienté sans circuit. De ce fait, ils s'avèrent particulièrement adaptés à la représentation des relations de causalité entre les variables. Ces modèles se retrouvent dans de nombreux domaines d'application, tels que le diagnostic médical (Lucas *et al.*, 2004), la bioinformatique (Friedman *et al.*, 2000), l'analyse des risques (Weber *et al.*, 2012), la sécurité informatique (Kruegel *et al.*, 2003) et bien d'autres (Pourret *et al.*, 2008).

2.1.2 Notions de théorie des graphes

Avant de définir formellement les réseaux bayésiens, il est nécessaire d'expliciter un certain nombre de notions issues de la théorie des graphes :

Graphe orienté : Un graphe orienté est un couple $G = (V, E)$, où V est un ensemble de nœuds et $E \subseteq V \times V$ un ensemble d'arcs orientés reliant chacun un couple de nœuds. Si $(u, v) \in E$, alors on dit que u est un « parent » de v et v un « enfant » de u.

Chemin : Un chemin est une suite de nœuds $(v_1, \ldots, v_n)$ $(n \geq 2)$ telle que, pour tout $i \in \{1, \ldots, n-1\}$, $(v_i, v_{i+1}) \in E$. Pour deux nœuds de ce chemin v_i et v_j tels que $i < j$, alors on dit que v_i est un « ascendant » de v_j et v_j un « descendant » de v_i.

Circuit : Un circuit est un chemin $(v_1, \ldots, v_n)$ tel que $v_1 = v_n$.

Chaîne : Une chaîne est une suite de nœuds $(v_1, \ldots, v_n)$ $(n \geq 2)$ telle que, pour tout $i \in \{1, \ldots, n-1\}$, $(v_i, v_{i+1}) \in E$ ou $(v_{i+1}, v_i) \in E$. Contrairement à un chemin, une chaîne ne tient donc pas compte de l'orientation des arcs.

D-séparation : Soient A, B et Z trois sous-ensembles de nœuds disjoints. A et B sont d-séparés par Z si, pour toute chaîne c entre un nœud de A et un nœud de B, au moins l'une des conditions suivantes est satisfaite :

— c contient une chaîne de type $u \to z \to v$ ou $u \leftarrow z \leftarrow v$ telle que $z \in Z$;

— c contient une chaîne de type $u \leftarrow z \rightarrow v$ telle que $z \in Z$;
— c contient une chaîne de type $u \rightarrow w \leftarrow v$ telle que $w \notin Z$ et, pour tout x descendant de w, $x \notin Z$.

2.1.3 Définition des réseaux bayésiens

Un réseau bayésien peut être défini comme un couple $\mathcal{B} = (G, \Theta)$ tel que :
— $G = (V, E)$ est un graphe orienté sans circuit dont l'ensemble des nœuds V est associé de manière bijective à un ensemble de variables aléatoires $\{X_1, \ldots, X_n\}$ [1] ;
— $\Theta = \left\{\theta_{X_1|\mathrm{Pa}_{X_1}}, \ldots, \theta_{X_n|\mathrm{Pa}_{X_n}}\right\}$ est un ensemble de paramètres tel que $\theta_{X_i|\mathrm{Pa}_{X_i}}$ permet de décrire la distribution de probabilité de X_i conditionnellement à ses parents Pa_{X_i} dans G.

D'après la propriété de Markov locale, toute variable d'un réseau bayésien est indépendante de ses non-descendants conditionnellement à ses parents. De ce fait, la distribution jointe sur l'ensemble des variables peut être représentée de manière compacte (Pearl, 1988) :

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\mathrm{Pa}_{X_i}) \,. \tag{2.1}$$

Cette décomposition de la distribution jointe globale en un produit de distributions conditionnelles locales est à la base d'algorithmes d'inférence permettant de calculer la probabilité de n'importe quelle variable du modèle à partir de l'observation, même partielle, des autres variables. Ces algorithmes sont assimilables à des méthodes de propagation d'information dans un graphe (Leray, 2006). Nous les abordons plus en détail dans la sous-section 2.6.5, dans le cadre des réseaux bayésiens dynamiques.

Les travaux de Verma et Pearl (1988) montrent que les indépendances conditionnelles encodées dans un réseau bayésien sont étroitement liées à la notion de d-séparation. Si deux sous-ensembles de variables A et B sont d-séparés par un troisième sous-ensemble Z, alors A est indépendant de B conditionnellement à Z. En d'autres termes, l'information circulant entre A et B est bloquée par Z. Ainsi, dans le réseau bayésien donné en exemple dans la figure 2.1, la variable X_5 est indépendante de X_1, X_4 et X_7 conditionnellement à X_2. Elle est également indépendante de X_6 conditionnellement à X_3. Nous retrouvons bien au passage la propriété de Mar-

1. Du fait de cette bijection, les notions de variable et de nœud peuvent être employées de manière interchangeable.

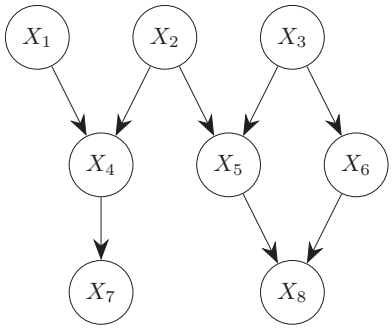


FIGURE 2.1 – Exemple de réseau bayésien.

kov locale, puisque X_5 est indépendante de ses non-descendants conditionnellement à ses parents.

Comme le relèvent Naïm *et al.* (2011), la plupart des travaux sur les réseaux bayésiens supposent que les variables utilisées sont discrètes. En présence de variables continues, deux types de stratégie peuvent être mis en place :

— la discrétisation des variables ;
— l'utilisation de distributions de probabilité continues, dont les paramètres sont déterminés de manière experte ou appris à partir des données.

Bien que la première option soit souvent privilégiée (Naïm *et al.*, 2011), la perte d'information qu'elle engendre peut s'avérer problématique dans certaines situations. C'est pourquoi nous nous intéressons dans notre étude aux réseaux bayésiens dont les variables sont continues et ne subissent pas de discrétisation préalable. Le cas discret étant traité dans de nombreux ouvrages, nous référons le lecteur à Koller et Friedman (2009), Neapolitan (2004) ou encore, dans la littérature francophone, à Naïm *et al.* (2011) et à Leray (2006). Le cas hybride (variables continues et discrètes) est également abordé dans l'ouvrage de Koller et Friedman (2009). Notons enfin que si la plupart des algorithmes d'apprentissage et d'inférence sont développés dans le cadre des réseaux bayésiens discrets, la majorité d'entre eux sont aisément transposables aux réseaux bayésiens continus.

2.2 Réseaux bayésiens gaussiens

2.2.1 Définition

D'après les travaux de Shachter et Kenley (1989), un réseau bayésien est gaussien si la distribution de probabilité jointe sur l'ensemble des variables est une distribution

gaussienne multivariée de densité :

$$p(x) = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}\det(\Sigma)^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}, \tag{2.2}$$

où μ est le vecteur moyenne et Σ la matrice de covariance. $\det(\Sigma)$ et Σ^{-1} désignent respectivement le déterminant et la matrice inverse de Σ, tandis que $(x-\mu)^\top$ désigne la transposée de $(x-\mu)$.

Dans un réseau bayésien gaussien, la densité jointe globale se décompose en un produit de densités conditionnelles locales décrites par des modèles linéaires gaussiens univariés :

$$p(X_i|\mathrm{Pa}_{X_i}) = \mathcal{N}\big(X_i\big|\beta_{0i} + \beta_i{}^\top \mathrm{Pa}_{X_i}, {\sigma_i}^2\big), \tag{2.3}$$

où β_{0i}, β_i et σ_i sont les paramètres de $p(X_i|\mathrm{Pa}_{X_i})$ (β_{0i} étant un scalaire et β_i un vecteur de même taille que Pa_{X_i}). En d'autres termes, chaque variable est une fonction linéaire de ses parents. L'équivalence entre la représentation multivariée globale et la représentation linéaire locale, ainsi que les transformations permettant de passer de l'une à l'autre, sont mises en évidence dans les travaux de Wermuth (1980).

2.2.2 Apprentissage des paramètres

Nous souhaitons déterminer les paramètres d'un réseau bayésien gaussien dont la structure graphique est connue. Pour ce faire, nous disposons d'un ensemble de données composé de N observations complètes $\mathcal{X} = \{x^1, \ldots, x^N\}$. Notre objectif est de trouver les paramètres qui expliquent le mieux ces données, autrement dit de réaliser un « apprentissage » des paramètres à partir des données.

La méthode d'apprentissage la plus couramment utilisée est celle du maximum de vraisemblance. Elle consiste à estimer les paramètres :

$$\Theta^* = \arg\max_{\Theta} \mathcal{L}(\Theta|\mathcal{X}), \tag{2.4}$$

où $\mathcal{L}(\Theta|\mathcal{X})$ est la vraisemblance de Θ sachant $\mathcal{X}$:

$$\mathcal{L}(\Theta|\mathcal{X}) = p(\mathcal{X}|\Theta) = \prod_{m=1}^{N} p(x^m|\Theta). \tag{2.5}$$

D'un point de vue analytique, il est souvent plus facile de maximiser la log-vraisemblance plutôt que la vraisemblance elle-même :

$$\ell(\Theta|\mathcal{X}) = \log \mathcal{L}(\Theta|\mathcal{X}) = \sum_{m=1}^{N} \log p(x^m|\Theta). \tag{2.6}$$

Soient x_i^m et $\mathrm{pa}_{X_i}^m$ les valeurs respectives de X_i et Pa_{X_i} pour l'observation x^m. En appliquant la propriété de décomposition des réseaux bayésiens, la log-vraisemblance se décompose en une somme de termes locaux indépendants :

$$\begin{aligned}\ell(\Theta|\mathcal{X}) &= \sum_{m=1}^{N} \log\left(\prod_{i=1}^{n} p\big(x_i^m \big| \mathrm{pa}_{X_i}^m, \Theta\big)\right) \\ &= \sum_{m=1}^{N}\sum_{i=1}^{n} \log p\big(x_i^m \big| \mathrm{pa}_{X_i}^m, \Theta\big) \\ &= \sum_{i=1}^{n} \ell\big(\theta_{X_i|\mathrm{Pa}_{X_i}} \big| \mathcal{X}\big)\,.\end{aligned} \tag{2.7}$$

Maximiser cette log-vraisemblance globale revient donc à maximiser chaque log-vraisemblance conditionnelle locale. Dans un réseau bayésien gaussien, il est donc possible d'estimer les paramètres de chaque modèle linéaire gaussien de manière indépendante, ce qui simplifie les calculs.

Considérons le modèle linéaire gaussien décrivant $p(X_i|\mathrm{Pa}_{X_i})$. D'après la démarche détaillée par Koller et Friedman (2009), l'estimation des paramètres de cette distribution s'effectue en deux temps. Elle consiste tout d'abord à estimer les paramètres μ_i et Σ_i de la distribution jointe gaussienne $p(X_i, \mathrm{Pa}_{X_i})$ (ce qui est trivial), puis à en déduire les paramètres de $p(X_i|\mathrm{Pa}_{X_i})$ à l'aide la distribution marginale $p(\mathrm{Pa}_{X_i})$:

$$p(X_i|\mathrm{Pa}_{X_i}) = \frac{p(X_i, \mathrm{Pa}_{X_i})}{p(\mathrm{Pa}_{X_i})}\,. \tag{2.8}$$

En partitionnant $\mu_i = \begin{pmatrix}\mu_{i1}\\ \mu_{i2}\end{pmatrix}$ et $\Sigma_i = \begin{pmatrix}\Sigma_{i1,1} & \Sigma_{i1,2}\\ \Sigma_{i2,1} & \Sigma_{i2,2}\end{pmatrix}$ tel que μ_{i1} et $\Sigma_{i1,1}$ sont les paramètres de $p(X_i)$, μ_{i2} et $\Sigma_{i2,2}$ les paramètres de $p(\mathrm{Pa}_{X_i})$ et $\Sigma_{i1,2} = {\Sigma_{i2,1}}^\top$ un vecteur ligne de même taille que Pa_{X_i}, les paramètres de $p(X_i|\mathrm{Pa}_{X_i})$ s'obtiennent de la manière suivante :

$$\begin{aligned}\beta_{0i} &= \mu_{i1} - \Sigma_{i1,2}{\Sigma_{i2,2}}^{-1}\mu_{i2} \\ {\beta_i}^\top &= \Sigma_{i1,2}{\Sigma_{i2,2}}^{-1} \\ {\sigma_i}^2 &= \Sigma_{i1,1} - \Sigma_{i1,2}{\Sigma_{i2,2}}^{-1}\Sigma_{i2,1}\,.\end{aligned} \tag{2.9}$$

Une méthode d'apprentissage alternative est proposée par Geiger et Heckerman (1994). Elle consiste à estimer les paramètres du réseau bayésien gaussien de manière globale, à partir de la distribution gaussienne multivariée qu'il représente. Une description détaillée de cette méthode est fournie dans l'ouvrage de Neapolitan (2004).

2.3 Réseaux bayésiens à mélanges gaussiens

2.3.1 Définition

L'utilisation des réseaux bayésiens gaussiens repose sur l'hypothèse que les données suivent une distribution gaussienne et que les relations entre les variables sont linéaires. Or dans bon nombre de situations, il arrive que cette hypothèse se révèle trop restrictive. C'est pourquoi certains auteurs choisissent de se tourner vers un type de distribution plus flexible : les modèles de mélanges gaussiens. Ces derniers se caractérisent par une somme pondérée de distributions gaussiennes :

$$p(x) = \sum_{j=1}^{M} \alpha_j \mathcal{N}(x|\mu_j, \Sigma_j)\,, \tag{2.10}$$

où les proportions de mélange $\alpha_1, \dots \alpha_M$ sont telles que $\alpha_j \geq 0$ pour tout j et $\sum_{j=1}^{M} \alpha_j = 1$. Plusieurs raisons expliquent l'intérêt de ce type de distribution (Sun *et al.*, 2006) :

- de nombreux processus naturels suivent une distribution gaussienne ;
- les distributions gaussiennes sont relativement simples à manipuler d'un point de vue mathématique ;
- il est possible d'approximer une distribution de probabilité arbitraire en combinant un nombre suffisant de gaussiennes.

Dans les réseaux bayésiens, l'utilisation de sommes pondérées de gaussiennes pour approximer des distributions arbitraires remonte aux travaux de Driver et Morrell (1995). D'après la représentation proposée par Davies et Moore (2000) et couramment adoptée depuis, un réseau bayésien à mélanges gaussiens est un réseau bayésien dont chaque distribution jointe locale $p(X_i, \mathrm{Pa}_{X_i})$ est décrite par un modèle de mélange gaussien. En notant $\alpha_{i,j}$, $\mu_{i,j}$ et $\Sigma_{i,j}$ les paramètres associés à la j-ème composante de $p(X_i, \mathrm{Pa}_{X_i})$ et en partitionnant $\mu_{i,j}$ et $\Sigma_{i,j}$ de manière analogue à la sous-section 2.2.2, la distribution conditionnelle locale s'exprime (Sun *et al.*, 2006) :

$$p(X_i|\mathrm{Pa}_{X_i}) = \sum_{j=1}^{M} \gamma_{i,j} \mathcal{N}\big(X_i \big| {\beta_{0}}_{i,j} + {\beta_{i,j}}^{\top} \mathrm{Pa}_{X_i}, {\sigma_{i,j}}^2\big)\,, \tag{2.11}$$

où :

$$\begin{aligned}
\gamma_{i,j} &= \frac{\alpha_{i,j}\mathcal{N}(\mathrm{Pa}_{X_i}|\mu_{i,j_2}, \Sigma_{i,j_{2,2}})}{\sum_{l=1}^{M} \alpha_{i,l}\mathcal{N}(\mathrm{Pa}_{X_i}|\mu_{i,l_2}, \Sigma_{i,l_{2,2}})} \\
{\beta_0}_{i,j} &= \mu_{i,j_1} - \Sigma_{i,j_{1,2}}{\Sigma_{i,j_{2,2}}}^{-1}\mu_{i,j_2} \\
{\beta_{i,j}}^{\top} &= \Sigma_{i,j_{1,2}}{\Sigma_{i,j_{2,2}}}^{-1} \\
{\sigma_{i,j}}^2 &= \Sigma_{i,j_{1,1}} - \Sigma_{i,j_{1,2}}{\Sigma_{i,j_{2,2}}}^{-1}\Sigma_{i,j_{2,1}}\,.
\end{aligned} \tag{2.12}$$

$p(X_i|\mathrm{Pa}_{X_i})$ est donc caractérisée par une somme pondérée de modèles linéaires gaussiens ayant le même nombre de composantes que $p(X_i, \mathrm{Pa}_{X_i})$.

À noter que nous nous intéressons ici exclusivement aux modèles de mélanges composés d'un nombre fini de distributions gaussiennes. Pour une étude des réseaux bayésiens à mélanges infinis, nous référons le lecteur aux travaux de Jarraya Siala (2013).

2.3.2 Apprentissage des paramètres

À l'instar des réseaux bayésiens gaussiens, la méthode la plus largement utilisée pour apprendre les paramètres d'un réseau bayésien à mélanges gaussiens est celle du maximum de vraisemblance. En présence d'un jeu de données complet, la log-vraisemblance se décompose de la même manière que dans l'équation (2.7). Maximiser la log-vraisemblance du réseau bayésien revient donc à maximiser indépendamment la log-vraisemblance de chaque modèle de mélange conditionnel. Comme pour les modèles linéaires gaussiens, l'approche classique consiste d'abord à estimer les paramètres de $p(X_i, \mathrm{Pa}_{X_i})$, puis à en déduire les paramètres de $p(X_i|\mathrm{Pa}_{X_i})$.

Supposons que nous souhaitons estimer les paramètres du maximum de vraisemblance d'un modèle de mélange gaussien à partir d'un ensemble de données $\mathcal{X} = \{x^1, \ldots, x^N\}$ de dimension d. Pour un ensemble de paramètres θ donné, la log-vraisemblance s'exprime :

$$\ell(\theta|\mathcal{X}) = \sum_{m=1}^{N} \log\left(\sum_{j=1}^{M} \alpha_j \mathcal{N}(x^m|\mu_j, \Sigma_j)\right). \tag{2.13}$$

La présence du logarithme d'une somme nous empêche ici de trouver une solution analytique au problème. Par conséquent, il s'avère nécessaire de recourir à des techniques plus sophistiquées.

Introduit par Dempster *et al.* (1977), l'algorithme espérance-maximisation (EM) permet d'estimer itérativement les paramètres du maximum de vraisemblance d'une

distribution lorsque les données observées $\mathcal{X}$ sont incomplètes. Cette méthode suppose l'existence d'un ensemble de données non observées $\mathcal{Y}$ tel que $(\mathcal{X}, \mathcal{Y})$ constitue les données complètes. En partant de paramètres initiaux θ^0, l'algorithme EM opère, à chaque itération k, deux étapes successives :

— l'étape d'espérance (E), où est calculée l'espérance de la log-vraisemblance des données complètes compte tenu des données observées et des paramètres θ^{k-1} estimés lors de l'itération précédente :

$$Q(\theta|\theta^{k-1}) = \mathbb{E}[\ell(\theta|\mathcal{X}, \mathcal{Y})|\mathcal{X}, \theta^{k-1}]\,; \tag{2.14}$$

— l'étape de maximisation (M), où sont estimés les paramètres θ^k qui maximisent cette espérance :

$$\theta^k = \arg\max_{\theta} Q(\theta|\theta^{k-1})\,. \tag{2.15}$$

Comme le montrent Dempster *et al.* (1977), chaque itération augmente la log-vraisemblance jusqu'à ce que celle-ci converge vers un maximum local.

Dans le cas d'un modèle de mélange gaussien, nous supposons que les données $\mathcal{X}$ sont incomplètes et qu'il existe un ensemble de données non observées $\mathcal{Y} = \{y^1, \ldots, y^N\}$ tel que y^m indique quelle composante a généré l'observation x^m. La log-vraisemblance des données complètes s'écrit alors :

$$\begin{aligned} \ell(\theta|\mathcal{X}, \mathcal{Y}) &= \sum_{m=1}^{N} \log p(x^m, y^m|\theta) \\ &= \sum_{m=1}^{N} \log(p(x^m|y^m, \theta)p(y^m|\theta)) \\ &= \sum_{m=1}^{N} \log(\alpha_{y^m}\mathcal{N}(x^m|\mu_{y^m}, \Sigma_{y^m}))\,, \end{aligned} \tag{2.16}$$

et peut désormais être maximisée de manière analytique. D'après ce résultat, il est donc possible d'estimer les paramètres du maximum de vraisemblance du modèle de mélange gaussien à l'aide de l'algorithme EM. À chaque itération k, l'étape E revient à calculer les probabilités a posteriori :

$$p(j|x^m, \theta^{k-1}) = \frac{\alpha_j^{k-1}\mathcal{N}\left(x^m\middle|\mu_j^{k-1}, \Sigma_j^{k-1}\right)}{\sum_{l=1}^{M} \alpha_l^{k-1}\mathcal{N}\left(x^m\middle|\mu_l^{k-1}, \Sigma_l^{k-1}\right)}\,. \tag{2.17}$$

Les paramètres sont ensuite mis à jour lors de l'étape M (Bilmes, 1998) :

$$\begin{aligned}
\alpha_j^k &= \frac{1}{N}\sum_{m=1}^{N} p(j|x^m,\theta^{k-1}) \\
\mu_j^k &= \frac{\sum_{m=1}^{N} p(j|x^m,\theta^{k-1})x^m}{\sum_{m=1}^{N} p(j|x^m,\theta^{k-1})} \\
\Sigma_j^k &= \frac{\sum_{m=1}^{N} p(j|x^m,\theta^{k-1})\left(x^m-\mu_j^k\right)\left(x^m-\mu_j^k\right)^\top}{\sum_{m=1}^{N} p(j|x^m,\theta^{k-1})}\,.
\end{aligned} \tag{2.18}$$

Au cours de la réalisation de l'algorithme EM, il peut arriver qu'une composante se réduise à un seul point et que sa matrice de covariance tende vers zéro. La log-vraisemblance devient alors infinie, provoquant un arrêt prématuré de l'algorithme et une situation de surapprentissage. Ce type de problème, appelé « singularité », peut être évité en procédant à une régularisation bayésienne du modèle. D'après la démarche de Ormoneit et Tresp (1996), cette régularisation consiste à remplacer l'estimation du maximum de vraisemblance par celle du maximum a posteriori, en maximisant la log-vraisemblance pénalisée :

$$\ell_{\mathrm{p}}(\theta|\mathcal{X}) = \ell(\theta|\mathcal{X}) + \sum_{j=1}^{M} \log\left(\det(\Sigma_j)^{-\frac{1}{2}} e^{-\frac{\lambda}{2}\operatorname{tr}\left({\Sigma_j}^{-1}\right)}\right), \tag{2.19}$$

où λ est un paramètre défini expérimentalement et $\operatorname{tr}\left({\Sigma_j}^{-1}\right)$ désigne la trace de ${\Sigma_j}^{-1}$. L'étape E de l'algorithme EM demeure inchangée. Dans l'étape M, seule l'estimation des matrices de covariances est légèrement modifiée :

$$\Sigma_j^k = \frac{\sum_{m=1}^{N} p(j|x^m,\theta^{k-1})\left(x^m-\mu_j^k\right)\left(x^m-\mu_j^k\right)^\top + \lambda I_d}{\sum_{m=1}^{N} p(j|x^m,\theta^{k-1}) + 1}\,, \tag{2.20}$$

où I_d la matrice identité d'ordre d. Ainsi, les éléments diagonaux des matrices de covariance sont bornés inférieurement, empêchant ces dernières de tendre vers zéro et prévenant ainsi les risques de singularité. En dehors de ce type de situation, les estimations obtenues sont similaires à celles du maximum de vraisemblance (Fraley et Raftery, 2007).

De par sa facilité d'implémentation et ses avantages par rapport à d'autres approches telles que les méthodes de gradient (Xu et Jordan, 1996), l'algorithme EM se révèle particulièrement attractif. C'est généralement sur cette méthode que reposent les algorithmes d'apprentissage des paramètres des réseaux bayésiens à mélanges gaussiens (Davies et Moore, 2000 ; Sun *et al.*, 2006 ; Cansado et Soto, 2008 ; Ko

et al., 2009 ; Liu, 2012). Dans la littérature, de nombreuses variantes sont proposées afin d'améliorer la vitesse la convergence (Neal et Hinton, 1998 ; Thiesson *et al.*, 2001). Davies et Moore (2000) utilisent par exemple des arbres k-d à multirésolution pour réduire le coût de chaque itération. Cependant, l'efficacité de cette méthode décroît significativement lorsque le nombre de dimensions augmente (Moore, 1999). L'algorithme de condensation des données utilisé par Cansado et Soto (2008) se révèle beaucoup plus rapide, notamment en présence d'un nombre élevé de variables (Soto *et al.*, 2007). Dans les travaux de Sun *et al.* (2006), une réduction de la dimension des données est opérée préalablement à l'algorithme EM par le biais d'une analyse en composantes principales.

2.3.3 Log-vraisemblance conditionnelle

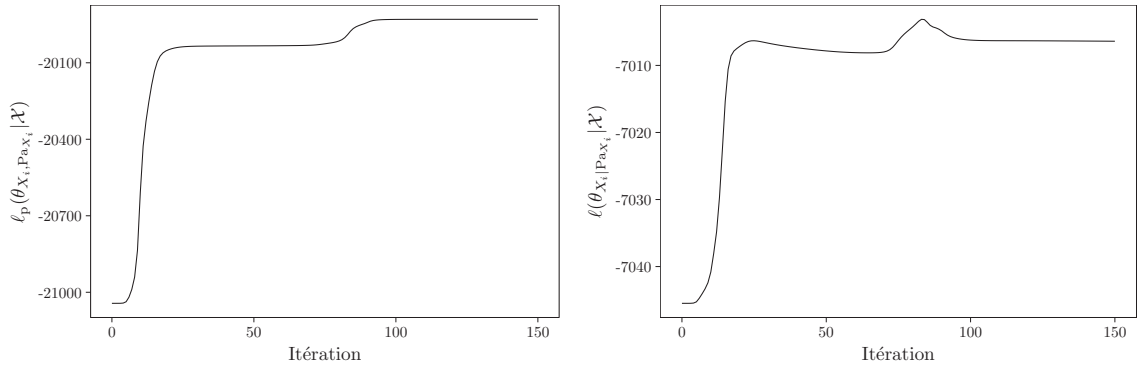
Lors de l'apprentissage des paramètres d'un réseau bayésien à mélanges gaussiens, l'approche classique consiste à estimer les paramètres de chaque distribution jointe locale $p(X_i, \mathrm{Pa}_{X_i})$, puis à en déduire ceux de la distribution conditionnelle associée $p(X_i|\mathrm{Pa}_{X_i})$. Or si les estimations issues de l'algorithme EM maximisent localement $\ell(\theta_{X_i,\mathrm{Pa}_{X_i}}|\mathcal{X})$ (ou $\ell_{\mathrm{p}}(\theta_{X_i,\mathrm{Pa}_{X_i}}|\mathcal{X})$ en cas de régularisation bayésienne), elles ne maximisent pas nécessairement $\ell(\theta_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X})$. En effet, l'inégalité de Jensen permet de garantir la croissance monotone de la log-vraisemblance jointe (Dempster *et al.*, 1977) :

$$\begin{aligned}\Delta\big(\theta^k_{X_i,\mathrm{Pa}_{X_i}}, \theta^{k-1}_{X_i,\mathrm{Pa}_{X_i}}\big) &= \ell\big(\theta^k_{X_i,\mathrm{Pa}_{X_i}}\big|\mathcal{X}\big) - \ell\big(\theta^{k-1}_{X_i,\mathrm{Pa}_{X_i}}\big|\mathcal{X}\big) \\ &\geq Q\big(\theta^k_{X_i,\mathrm{Pa}_{X_i}}|\theta^{k-1}_{X_i,\mathrm{Pa}_{X_i}}\big) - Q\big(\theta^{k-1}_{X_i,\mathrm{Pa}_{X_i}}|\theta^{k-1}_{X_i,\mathrm{Pa}_{X_i}}\big) \\ &\geq 0\,.\end{aligned} \tag{2.21}$$

Elle ne permet pas en revanche d'assurer la monotonie de la log-vraisemblance conditionnelle :

$$\begin{aligned}\Delta\big(\theta^k_{X_i|\mathrm{Pa}_{X_i}}, \theta^{k-1}_{X_i|\mathrm{Pa}_{X_i}}\big) &= \sum_{m=1}^{N} \log \frac{p\big(x_i^m, \mathrm{pa}_{X_i}^m|\theta^k_{X_i,\mathrm{Pa}_{X_i}}\big)}{p\big(\mathrm{pa}_{X_i}^m|\theta^k_{\mathrm{Pa}_{X_i}}\big)} - \sum_{m=1}^{N} \log \frac{p\big(x_i^m, \mathrm{pa}_{X_i}^m|\theta^{k-1}_{X_i,\mathrm{Pa}_{X_i}}\big)}{p\big(\mathrm{pa}_{X_i}^m|\theta^{k-1}_{\mathrm{Pa}_{X_i}}\big)} \\ &= \Delta\big(\theta^k_{X_i,\mathrm{Pa}_{X_i}}, \theta^{k-1}_{X_i,\mathrm{Pa}_{X_i}}\big) - \Delta\big(\theta^k_{\mathrm{Pa}_{X_i}}, \theta^{k-1}_{\mathrm{Pa}_{X_i}}\big) \\ &\geq -\Delta\big(\theta^k_{\mathrm{Pa}_{X_i}}, \theta^{k-1}_{\mathrm{Pa}_{X_i}}\big)\,.\end{aligned} \tag{2.22}$$

En cas de maximisation de la log-vraisemblance jointe pénalisée, un raisonnement analogue permet d'aboutir à la même conclusion.



(a) (b)

FIGURE 2.2 – Application de l'algorithme EM pour l'estimation des paramètres d'un modèle de mélange gaussien $p(X_i, \mathrm{Pa}_{X_i})$: (a) évolution de $\ell_{\mathrm{p}}(\theta_{X_i,\mathrm{Pa}_{X_i}}|\mathcal{X})$; (b) évolution de $\ell(\theta_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X})$.

Dans le cas des modèles de mélanges gaussiens, l'absence de monotonie de la log-vraisemblance conditionnelle est bien illustrée par Jebara et Pentland (1999). Bien que la dernière itération $k_{\max}$ de l'algorithme EM soit celle pour laquelle la log-vraisemblance jointe est la plus élevée (voir la figure 2.2a), la log-vraisemblance conditionnelle finale peut se révéler plus basse que lors de précédentes itérations (voir la figure 2.2b). Les paramètres $\theta^{k_{\max}}_{X_i|\mathrm{Pa}_{X_i}}$ obtenus à cette itération ne sont donc pas forcément optimaux.

Dès lors, il peut être intéressant de se pencher sur des méthodes d'estimation permettant de maximiser directement la log-vraisemblance conditionnelle sans passer par la log-vraisemblance jointe. De telles méthodes sont notamment développées à travers le formalisme des mélanges d'experts (Jacobs *et al.*, 1991). Jebara et Pentland (1999) proposent une version de l'algorithme EM qui, en substituant l'inégalité de Jensen par une autre borne de la fonction logarithme, garantit la convergence monotone de la log-vraisemblance conditionnelle. Les résultats obtenus se révèlent supérieurs à ceux de l'algorithme EM classique. Cependant, la complexité du processus de mise à jour des paramètres et la lenteur de la convergence entravent l'usage pratique de cette méthode (Salojärvi *et al.*, 2005).

Jordan et Jacobs (1994) développent une version de l'algorithme EM dans laquelle l'étape M met à jour les paramètres des mélanges d'experts à l'aide de la méthode itérative des moindres carrés repondérés (IRLS). Afin de pallier l'instabilité de convergence de cet algorithme et améliorer ses performances, Chen *et al.* (1999) suggèrent de remplacer la méthode IRLS par la méthode de Newton-Raphson. Ng

et McLachlan (2004) utilisent un algorithme espérance-maximisation conditionnelle (ECM) pour travailler sur des espaces paramétriques de dimension plus petite et simplifier ainsi l'étape M. Malheureusement, ces approches impliquent la présence d'une nouvelle boucle itérative à l'intérieur de chaque itération de l'algorithme EM, ce qui augmente considérablement les temps de calcul.

Au final, l'approche classique de maximisation de la log-vraisemblance jointe demeure la solution la plus simple et la moins coûteuse à mettre en œuvre. Dans les travaux de Xu *et al.* (1995), elle se révèle près de quatre fois plus rapide que la méthode de Jordan et Jacobs (1994), tout en produisant des résultats similaires. Contrairement à cette dernière, elle ne requiert aucun ajustement externe pour garantir la convergence de l'algorithme.

Il convient néanmoins de s'assurer que l'algorithme EM ne dégrade pas la log-vraisemblance conditionnelle. En effet, de par l'absence de monotonie de cette dernière, il est possible que $\ell\big(\theta^{k_{\max}}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\big)$ soit plus faible que $\ell\big(\theta^{0}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\big)$, notamment lorsque $\theta^{0}_{X_i|\mathrm{Pa}_{X_i}}$ est déjà issu d'une précédente optimisation. Bien que ce problème apparaisse relativement rarement, il peut être préjudiciable à l'optimisation de la log-vraisemblance du réseau bayésien. Un moyen simple de le contourner est de mettre en place un test d'acceptation ou de rejet des paramètres. Si $\ell\big(\theta^{k_{\max}}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\big) < \ell\big(\theta^{0}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\big)$, alors les nouveaux paramètres sont rejetés. Dans le cas contraire, ils sont acceptés.

2.3.4 Limites de l'algorithme EM

L'une des principales limites de l'algorithme EM est sa sensibilité aux valeurs initiales des paramètres. En cas de « mauvaise » initialisation, il peut converger vers un maximum local indésirable et produire des estimations de faible qualité. Dans le cas d'un modèle de mélange gaussien, cette situation se traduit par une mauvaise répartition des composantes dans l'espace des données. Ces dernières ne peuvent plus se déplacer vers leur position optimale (correspondant au maximum global) sans passer par des régions de l'espace où la log-vraisemblance est plus basse (Ueda *et al.*, 2000). Afin d'améliorer les chances d'atteindre le maximum global, l'algorithme EM est souvent exécuté plusieurs fois avec des initialisations différentes. Cependant, cette méthode implique une charge de calcul importante et s'applique donc difficilement à de grands jeux de données (Soto *et al.*, 2007).

Ueda *et al.* (2000) traitent le problème des maximums locaux à l'aide d'un algorithme EM de « division-fusion ». Lors de la convergence vers un maximum local,

cet algorithme opère simultanément la division d'une composante en deux et la fusion de deux autres composantes en une (le nombre total de composantes n'étant pas modifié). Les mouvements discrets engendrés par ces opérations permettent aux composantes de s'extraire du maximum local et de poursuivre leur déplacement vers une meilleure position dans l'espace des données. Zhang *et al.* (2003) reprennent ce principe en introduisant des méthodes de division basées sur la décomposition de matrices.

Une autre limite de l'algorithme EM est qu'il ne permet pas de déterminer automatiquement le nombre optimal de composantes. Ce nombre doit donc être défini a priori, ce qui constitue une difficulté supplémentaire de la modélisation. En effet, si un modèle de mélange comporte trop peu de composantes, alors il n'est pas suffisamment flexible pour capturer convenablement la structure des données. À l'inverse, un nombre trop important de composantes augmente inutilement sa complexité et favorise les risques de surapprentissage.

Figueiredo et Jain (2002) proposent une variante de l'algorithme EM permettant d'optimiser le nombre de composantes d'un modèle de mélange gaussien. Le modèle est initialisé avec un grand nombre de composantes, que l'algorithme annihile successivement quand leur poids descend en dessous d'un certain seuil. À l'inverse, Vlassis et Likas (2002) puis Verbeek *et al.* (2003) développent des algorithmes gloutons permettant, à partir d'un modèle à une seule gaussienne, d'ajouter de nouvelles composantes de manière itérative. Les paramètres d'une distribution gaussienne étant estimés analytiquement, ces deux dernières méthodes ont l'avantage de ne pas requérir d'initialisation.

2.3.5 Algorithme EM de division-fusion

L'algorithme EM de division-fusion de Ueda *et al.* (2000) peut être modifié afin d'élaborer une méthode gloutonne capable de gérer à la fois le problème des maximums locaux et celui du nombre de composantes des modèles de mélanges gaussiens. Lorsque l'algorithme EM converge vers un maximum local, la méthode que nous proposons de mettre en œuvre ne consiste plus à effectuer une division et une fusion simultanées, mais à choisir l'une ou l'autre de ces opérations selon un critère de sélection à maximiser. Cette procédure est réitérée tant que ce critère peut être amélioré. Contrairement aux méthodes évoquées précédemment, chaque itération de cet algorithme est donc susceptible d'augmenter ou de diminuer le nombre de composantes, ce qui apporte davantage de flexibilité. C'est sur cette idée que repose

par exemple l'algorithme EM compétitif de Zhang *et al.* (2004), utilisé dans le cadre de l'apprentissage du réseau bayésien à mélanges gaussiens de Sun *et al.* (2006).

Supposons que nous souhaitons diviser une composante de paramètres $(\alpha_j, \mu_j, \Sigma_j)$ en deux composantes de paramètres respectifs $(\alpha_{j_1}, \mu_{j_1}, \Sigma_{j_1})$ et $(\alpha_{j_2}, \mu_{j_2}, \Sigma_{j_2})$. Ueda *et al.* (2000) recommandent d'initialiser les nouveaux paramètres de la manière suivante :

$$\begin{aligned} \alpha_{j_1} = \alpha_{j_2} &= \frac{\alpha_j}{2} \\ \mu_{j_1} &= \mu_j + \epsilon_1 \\ \mu_{j_2} &= \mu_j + \epsilon_2 \\ \Sigma_{j_1} = \Sigma_{j_2} &= \det(\Sigma_j)^{\frac{1}{d}} I_d \,, \end{aligned} \tag{2.23}$$

où ϵ_1 et ϵ_2 sont de petites perturbations aléatoires permettant de dissocier les deux composantes créées. Si cette initialisation garantit la définie positivité des nouvelles matrices de covariance, elle ne conserve pas l'anisotropie de la matrice d'origine (Zhang *et al.*, 2003). Pour pallier ce problème, nous pouvons simplement initialiser les matrices ainsi :

$$\Sigma_{j_1} = \Sigma_{j_2} = \Sigma_j \,. \tag{2.24}$$

Supposons à présent que nous souhaitons fusionner deux composantes de paramètres respectifs $(\alpha_{j_1}, \mu_{j_1}, \Sigma_{j_1})$ et $(\alpha_{j_2}, \mu_{j_2}, \Sigma_{j_2})$. D'après Ueda *et al.* (2000), les paramètres de la nouvelle composante doivent être initialisés comme suit :

$$\begin{aligned} \alpha_j &= \alpha_{j_1} + \alpha_{j_2} \\ \mu_j &= \frac{\alpha_{j_1}\mu_{j_1} + \alpha_{j_2}\mu_{j_2}}{\alpha_j} \\ \Sigma_j &= \frac{\alpha_{j_1}\Sigma_{j_1} + \alpha_{j_2}\Sigma_{j_2}}{\alpha_j} \,. \end{aligned} \tag{2.25}$$

Toutefois, Zhang *et al.* (2003) proposent une autre initialisation de la matrice de covariance :

$$\Sigma_j = \frac{\alpha_{j_1}\left(\Sigma_{j_1} + \mu_{j_1}{\mu_{j_1}}^\top\right) + \alpha_{j_2}\left(\Sigma_{j_2} + \mu_{j_2}{\mu_{j_2}}^\top\right)}{\alpha_j} - \mu_j {\mu_j}^\top \,, \tag{2.26}$$

laquelle permet de garantir que :

$$\alpha_j \mathcal{N}(x|\mu_j, \Sigma_j) = \alpha_{j_1}\mathcal{N}(x|\mu_{j_1}, \Sigma_{j_1}) + \alpha_{j_2}\mathcal{N}(x|\mu_{j_2}, \Sigma_{j_2}) \,. \tag{2.27}$$

Après l'initialisation d'une division ou d'une fusion de composantes, l'algorithme EM classique est mis en œuvre afin de réestimer les paramètres du modèle de mélange

gaussien. Dans le cas d'une division, il peut être préalablement judicieux de réajuster uniquement les paramètres des deux composantes créées. Cette étape est réalisée à l'aide de l'algorithme EM partiel décrit par Ueda *et al.* (2000). À chaque itération k de cet algorithme, seuls les paramètres des nouvelles composantes sont mis à jour. Afin de conserver la cohérence avec les composantes existantes, les probabilités a posteriori calculées lors de l'étape E (voir l'équation (2.17)) se calculent désormais, pour $u \in \{1, 2\}$:

$$p(j_u|x^m, \theta^{k-1}) = \frac{\alpha_{j_u}^{k-1} \mathcal{N}\left(x^m \middle| \mu_{j_u}^{k-1}, \Sigma_{j_u}^{k-1}\right)}{\sum_{l=1}^{2} \alpha_{j_l}^{k-1} \mathcal{N}\left(x^m \middle| \mu_{j_l}^{k-1}, \Sigma_{j_l}^{k-1}\right)} p(j|x^m, \theta^*)\,, \tag{2.28}$$

où θ^* désigne les paramètres du modèle avant la division. Cette modification de l'étape E permet de garantir que :

$$p(j_1|x^m, \theta^{k-1}) + p(j_2|x^m, \theta^{k-1}) = p(j|x^m, \theta^*)\,. \tag{2.29}$$

Dans le cas d'une fusion, la mise en œuvre de l'algorithme EM partiel n'est pas nécessaire, dans la mesure où l'initialisation proposée par Zhang *et al.* (2003) (en définissant les matrices comme dans l'équation (2.26)) optimise déjà la position de la nouvelle composante.

Pour un modèle de mélange gaussien à M composantes, il existe M possibilités de division et $\frac{M(M-1)}{2}$ possibilités de fusion. Dès lors, la question est de savoir comment sélectionner l'opération la plus pertinente. Une première solution consiste à tester l'ensemble des divisions et des fusions possibles, puis à retenir l'opération qui maximise le critère de sélection. Cependant, réaliser l'ensemble des $\frac{M(M+1)}{2}$ opérations peut s'avérer coûteux si M est élevé. Afin de limiter le nombre de tests, il est possible de définir deux critères rapides à calculer permettant d'ordonner respectivement les divisions et les fusions candidates. Le principe est de tester uniquement la première division et la première fusion de ces classements, puis de retenir la meilleure des deux opérations. Si cette dernière améliore le critère de sélection, alors les paramètres sont mis à jour et l'algorithme se poursuit. Dans le cas contraire, la même procédure est réitérée avec la deuxième division et la deuxième fusion, et ainsi de suite. Si pour tout $c \in \{1, \ldots, C_{\max}\}$ ($C_{\max}$ étant un paramètre défini expérimentalement), ni la c-ème division ni la c-ème fusion n'améliorent le critère de sélection, alors l'algorithme se termine en conservant les paramètres actuels. De cette manière, le nombre d'opérations testées est limité à $2C_{\max}$.

Ueda *et al.* (2000) définissent deux critères permettant de classer respectivement les divisions et les fusions candidates. Le critère de division utilisé est la divergence

locale de Kullback :

$$J_{\text{div}}(j|\theta) = \int f_j(x|\theta) \log \frac{f_j(x|\theta)}{\mathcal{N}(x|\mu_j, \Sigma_j)} \, dx \,, \tag{2.30}$$

qui représente la distance entre la j-ème composante du modèle de mélange gaussien et la densité locale des données autour de cette composante :

$$f_j(x|\theta) = \frac{\sum_{m=1}^{N} \delta(x - x^m) p(j|x^m, \theta)}{\sum_{m=1}^{N} p(j|x^m, \theta)} \,, \tag{2.31}$$

δ étant la fonction delta de Dirac. Plus $J_{\text{div}}(j|\theta)$ est élevé, plus cette densité locale est mal estimée par la j-ème composante et plus cette dernière est donc une bonne candidate à la division.

Le critère de fusion utilisé par Ueda *et al.* (2000) est défini de la manière suivante :

$$J_{\text{fus}}(j_1, j_2|\theta) = \frac{p_{j_1}(\theta)^\top p_{j_2}(\theta)}{\|p_{j_1}(\theta)\| \|p_{j_2}(\theta)\|} \,, \tag{2.32}$$

où $p_{j_1}(\theta) = (p(j_1|x^1, \theta), \ldots, p(j_1|x^N, \theta))^\top$, $p_{j_2}(\theta) = (p(j_2|x^1, \theta), \ldots, p(j_2|x^N, \theta))^\top$, de normes euclidiennes respectives $\|p_{j_1}(\theta)\|$ et $\|p_{j_2}(\theta)\|$. Plus $J_{\text{fus}}(j_1, j_2|\theta)$ est élevé, plus les probabilités a posteriori des j_1-ème et j_2-ème composantes sont semblables et plus ces dernières sont donc de bonnes candidates à la fusion.

L'algorithme EM de division-fusion présenté dans cette sous-section est décrit dans l'algorithme 2.1, en notant $S(\theta|\mathcal{X})$ le critère de sélection du modèle à maximiser. Si de nombreux critères sont proposés dans la littérature (Štepánová et Vavrečka, 2016), le plus utilisé est le critère d'information bayésien (BIC). Proposé par Schwarz (1978), le BIC représente un bon compromis entre la qualité d'ajustement du modèle et la complexité de celui-ci. Il s'exprime comme une fonction de la log-vraisemblance, à laquelle s'ajoute un terme de pénalité :

$$\text{BIC}(\theta|\mathcal{X}) = \ell(\theta|\mathcal{X}) - \frac{\log N}{2} \dim(\theta) \,, \tag{2.33}$$

où $\dim(\theta)$ est le nombre de paramètres libres du modèle. Ce critère est couramment utilisé pour l'optimisation des modèles de mélanges gaussiens, notamment dans le cadre des réseaux bayésiens (Davies et Moore, 2000 ; Ko *et al.*, 2009 ; Liu, 2012). Dans les travaux de Liu (2012), il se révèle plus précis que la log-vraisemblance et le critère d'information d'Akaike (AIC) pour déterminer le nombre adéquat de composantes. En cas de régularisation bayésienne du modèle, la log-vraisemblance du premier terme est remplacée par la log-vraisemblance pénalisée explicitée dans l'équation (2.19) (Fraley et Raftery, 2007) :

$$\text{BIC}_{\text{p}}(\theta|\mathcal{X}) = \ell_{\text{p}}(\theta|\mathcal{X}) - \frac{\log N}{2} \dim(\theta) \,. \tag{2.34}$$

ALGORITHME 2.1 – Algorithme EM de division-fusion pour l'estimation du nombre de composantes et des paramètres d'un modèle de mélange gaussien

Entrée : $\mathcal{X}$
θ^0
$C_{\max}$

1. $\theta^{**} \leftarrow$ réalisation de l'algorithme EM classique à partir de θ^0
2. **répéter**
3. $\quad\theta^* \leftarrow \theta^{**}$
4. $\quad M \leftarrow$ nombre de composantes de θ^*
5. $\quad(j^1, \ldots, j^M) \leftarrow$ classement des composantes de θ^* selon J_{div}
6. $\quad$**si** $M \geq 2$ **alors**
7. $\quad\quad\left(\left\{j_1^1, j_2^1\right\}, \ldots, \left\{j_1^{M(M-1)/2}, j_2^{M(M-1)/2}\right\}\right) \leftarrow$ classement des paires de composantes de θ^* selon J_{fus}
8. $\quad c \leftarrow 0$
9. $\quad$**tant que** $\theta^{**} = \theta^*$ et $c < C_{\max}$ **faire**
10. $\quad\quad c \leftarrow c + 1$
11. $\quad\quad$**si** $c \leq M$ **alors**
12. $\quad\quad\quad\theta^{***} \leftarrow$ initialisation de la division de la j^c-ème composante de θ^*
13. $\quad\quad\quad\theta^{***} \leftarrow$ réalisation de l'algorithme EM partiel sur les nouvelles composantes à partir de θ^{***}
14. $\quad\quad\quad\theta^{***} \leftarrow$ réalisation de l'algorithme EM classique à partir de θ^{***}
15. $\quad\quad\quad$**si** $S(\theta^{***}|\mathcal{X}) > S(\theta^{**}|\mathcal{X})$ **alors**
16. $\quad\quad\quad\quad\theta^{**} \leftarrow \theta^{***}$
17. $\quad\quad$**si** $c \leq \frac{M(M-1)}{2}$ **alors**
18. $\quad\quad\quad\theta^{***} \leftarrow$ initialisation de la fusion des j_1^c-ème et j_2^c-ème composantes de θ^*
19. $\quad\quad\quad\theta^{***} \leftarrow$ réalisation de l'algorithme EM classique à partir de θ^{***}
20. $\quad\quad\quad$**si** $S(\theta^{***}|\mathcal{X}) > S(\theta^{**}|\mathcal{X})$ **alors**
21. $\quad\quad\quad\quad\theta^{**} \leftarrow \theta^{***}$
22. **jusqu'à** $\theta^{**} = \theta^*$
23. **retourner** θ^*

La figure 2.3 permet de visualiser le processus d'apprentissage d'un modèle de mélange gaussien bivarié par l'algorithme EM de division-fusion ($C_{\max} = 3$). Dans cet exemple, les composantes gaussiennes sont représentées sous forme d'ellipses dont le centre et la forme dépendent respectivement des paramètres des vecteurs moyenne et des matrices de covariance (avec un niveau de confiance de 95 %). Les données d'apprentissage sont composées de 1000 observations générées aléatoirement à partir des quatre composantes de la figure 2.3a. Le critère de sélection utilisé est le BIC de la distribution jointe dont la log-vraisemblance est pénalisée ($\lambda = 0.01$). Partant d'un modèle de mélange gaussien à une composante (voir la figure 2.3b), l'algorithme EM

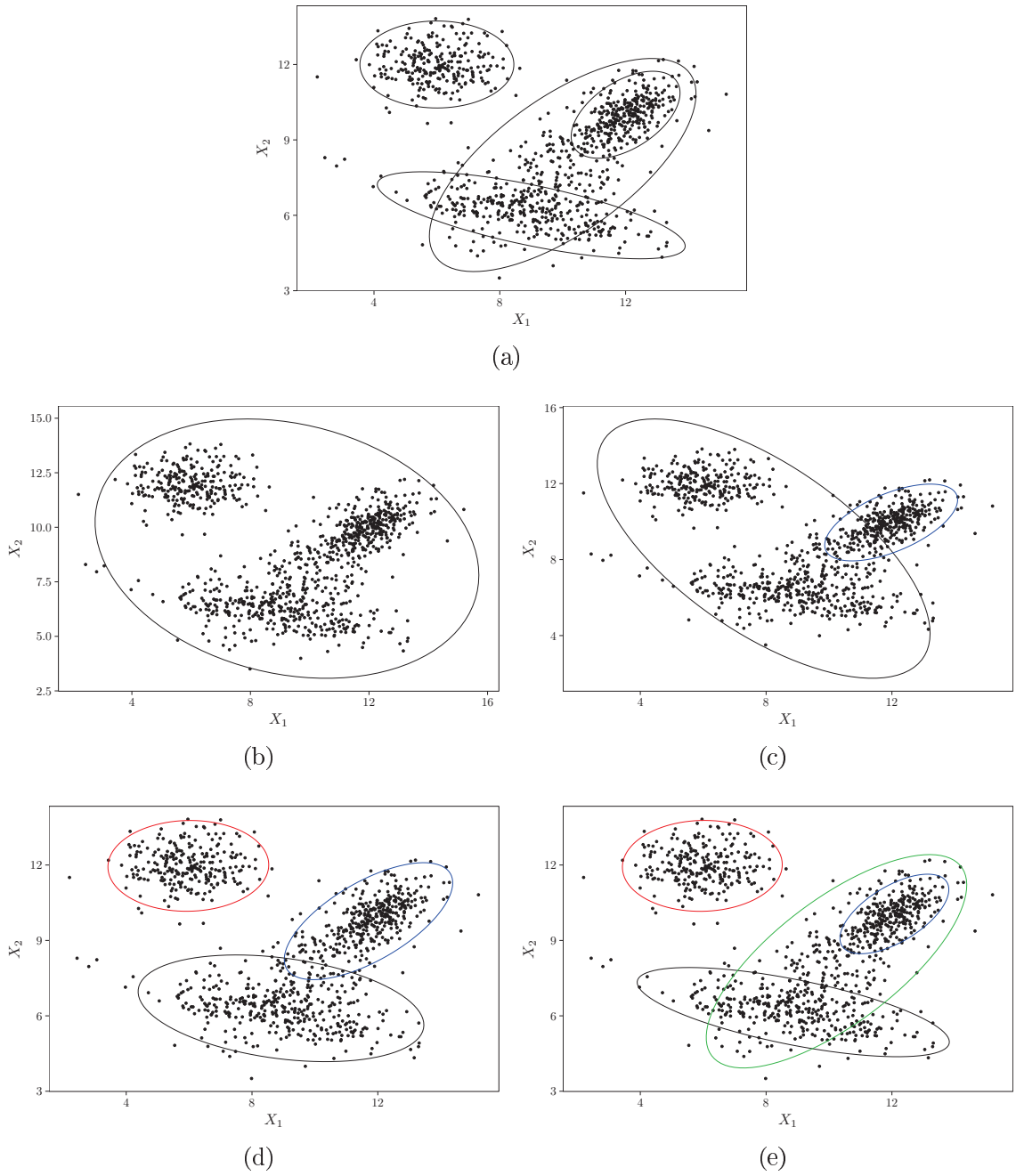


FIGURE 2.3 – Optimisation du nombre de composantes et estimation des paramètres d'un modèle de mélange gaussien bivarié par l'algorithme EM de division-fusion ($C_{\max} = 3$) : (a) composantes génératrices des données d'apprentissage ; (b) optimisation de la composante initiale ; (c) division de la composante noire en deux composantes noire et bleue ; (d) division de la composante noire en deux composantes noire et rouge ; (e) division de la composante noire en deux composantes noire et verte (convergence de l'algorithme).

de division-fusion opère trois divisions successives, jusqu'à converger vers le modèle de la figure 2.3e. Nous pouvons voir que ce modèle final est similaire à celui ayant servi à générer les données, ce qui illustre la qualité de l'apprentissage.

Dans le cadre de notre étude, les modèles de mélanges gaussiens sont utilisés pour décrire chaque distribution jointe locale du réseau bayésien. Aussi, plutôt que d'optimiser le BIC de ces distributions jointes, il peut être judicieux de raisonner de manière globale en cherchant à optimiser directement le BIC du réseau bayésien. Ce dernier possède la même propriété fondamentale que la log-vraisemblance, à savoir la capacité à se décomposer en une somme de termes locaux :

$$\begin{aligned} \mathrm{BIC}(\Theta|\mathcal{X}) &= \ell(\Theta|\mathcal{X}) - \frac{\log N}{2}\dim(\Theta) \\ &= \sum_{i=1}^{n} \ell\big(\theta_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\big) - \frac{\log N}{2}\sum_{i=1}^{n}\dim\left(\theta_{X_i|\mathrm{Pa}_{X_i}}\right) \\ &= \sum_{i=1}^{n} \mathrm{BIC}\big(\theta_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\big)\,. \end{aligned} \tag{2.35}$$

Maximiser ce BIC global revient donc à maximiser indépendamment le BIC de chaque distribution conditionnelle locale. Ainsi, lors de l'application de l'algorithme EM de division-fusion, il paraît intéressant d'utiliser le BIC de la distribution conditionnelle locale comme critère de sélection, plutôt que le BIC de la distribution jointe locale. Une comparaison de ces deux critères est réalisée au cours de l'expérimentation du chapitre 4.

2.4 Apprentissage de la structure

2.4.1 Typologie des méthodes d'apprentissage

Dans les sections précédentes, nous avons considéré que la structure graphique du réseau bayésien était déjà connue. À présent, nous nous intéressons plus spécifiquement aux algorithmes permettant de déterminer quelle structure représente le mieux notre problème. Les approches existantes peuvent être classées en trois catégories : les méthodes de recherche sous contraintes, les méthodes d'optimisation de score et les méthodes hybrides (Naïm *et al.*, 2011 ; Trabelsi, 2013).

Une première catégorie de méthodes assimile l'apprentissage de la structure à un problème de recherche sous contraintes. Ces dernières consistent à identifier les indépendances conditionnelles entre les variables par le biais de tests statistiques, puis à construire la structure du réseau bayésien selon ces indépendances. Les algorithmes

les plus populaires comprennent l'algorithme de causalité inductive (IC) (Pearl et Verma, 1991), l'algorithme de Spirtes, Glymour et Scheines (SGS) ou encore l'algorithme de Peter-Clark (PC) (Spirtes *et al.*, 1993). L'inconvénient de ces approches est qu'elles ne font pas intervenir de fonction objectif explicite. Aussi, leur efficacité dépend étroitement de la fiabilité des tests d'indépendance réalisés (Trabelsi, 2013).

Une seconde catégorie de méthodes assimile l'apprentissage de la structure à un problème d'optimisation. Ces méthodes consistent à définir une fonction de score statistiquement pertinente et à rechercher la structure qui maximise cette dernière. En pratique, le score choisi doit être localement décomposable afin de traiter chaque structure locale (c'est-à-dire une variable et ses parents) de manière indépendante. Un aperçu des scores possédant cette propriété peut être trouvé dans l'ouvrage de Naïm *et al.* (2011). L'inconvénient de ces approches est que la recherche de la structure optimale demeure un problème NP-complet (Chickering, 1996). Tester l'ensemble des structures possibles est souvent irréalisable en un temps raisonnable, d'où la nécessité de recourir à des heuristiques. Il est par exemple possible de limiter la recherche à l'espace des arbres en déterminant l'arbre de recouvrement optimal (Chow et Liu, 1968), d'ordonner les variables afin de réduire le nombre d'arcs candidats (Cooper et Herskovits, 1992) ou encore de procéder à une recherche gloutonne dans l'espace des réseaux bayésiens (Heckerman *et al.*, 1995).

Une troisième catégorie de méthodes d'apprentissage de la structure combine les algorithmes basés sur un score avec les algorithmes de recherche sous contraintes. Ces méthodes hybrides permettent de profiter des avantages des deux types d'approche (Naïm *et al.*, 2011) et sont particulièrement adaptées aux jeux de données contenant un grand nombre de variables (Trabelsi, 2013). Proposé par Friedman *et al.* (1999), l'algorithme « Sparse Candidate » (SC) consiste à restreindre les parents de chaque variable à un sous-ensemble limité de candidats sélectionnés selon des mesures de dépendance, puis à apprendre la structure du réseau bayésien en tenant compte de ces contraintes. Lors de l'itération suivante, la structure apprise est utilisée pour sélectionner de nouveaux sous-ensembles de parents candidats, et ainsi de suite jusqu'à la convergence du score. Cette méthode est notamment appliquée dans plusieurs travaux sur les réseaux bayésiens à mélanges gaussiens (Cansado et Soto, 2008 ; Ko *et al.*, 2009). L'algorithme « Max-Min Hill-Climbing » (MMHC) de Tsamardinos *et al.* (2006) adopte un principe similaire en sélectionnant une structure non orientée à l'aide d'un algorithme de recherche locale, puis en orientant les arcs à l'aide d'une procédure de recherche gloutonne.

2.4.2 Sélection experte et recherche gloutonne

Les méthodes d'apprentissage de la structure que nous avons décrites jusqu'à présent sont exclusivement basées sur les données. Or comme le détaille le chapitre 3, la connaissance experte joue un rôle prépondérant dans notre approche de modélisation. Bien que cette connaissance soit insuffisante pour déterminer précisément la structure du réseau bayésien, elle permet néanmoins de réduire l'espace de recherche en sélectionnant un sous-ensemble restreint de parents candidats pour chaque variable. Comme souvent lorsque l'expert intervient dans la construction du modèle, ces candidats sont déterminés en fonction des relations de causalité entre les variables.

La sélection experte des arcs candidats peut être complétée par une méthode classique d'optimisation de score. Dans notre étude, nous appliquons un algorithme de recherche gloutonne, dont chaque itération consiste à explorer le voisinage de la structure courante pour ajouter (si cette opération ne créé pas de circuit) ou retirer un arc candidat. L'opération retenue est celle qui, après réestimation des paramètres, maximise la fonction de score (la procédure étant réitérée jusqu'à ce que le score ne puisse plus être amélioré). Par analogie avec les critères de sélection présentés dans la sous-section 2.3.5, nous choisissons de maximiser le BIC, dont le terme de pénalité permet de privilégier les structures moins complexes et dont la propriété de décomposition (voir l'équation (2.35)) permet d'optimiser chaque structure locale de manière indépendante. Cet algorithme d'apprentissage de la structure est décrit dans l'algorithme 2.2, en notant E^{c} l'ensemble des arcs candidats.

Il est important de souligner que la qualité du maximum local atteint par un algorithme de recherche gloutonne dépend de la structure de départ du réseau bayésien. Naïm *et al.* (2011) présentent un certain nombre de méthodes permettant d'améliorer les chances d'atteindre le maximum global, mais aussi d'augmenter la vitesse de convergence. Citons par exemple les travaux de Leray et François (2004), qui proposent d'initialiser la structure par l'arbre de recouvrement maximal. Dans l'expérimentation du chapitre 4, nous testons deux initialisations différentes : celle par le graphe vide et celle par le graphe contenant tous les arcs candidats.

2.5 Apprentissage en cas de données incomplètes

Les méthodes d'apprentissage détaillées dans les sections précédentes impliquent l'utilisation d'un jeu de données complet. Or dans de nombreuses situations, cer-

ALGORITHME 2.2 – Apprentissage de la structure (et des paramètres) d'un réseau bayésien par recherche gloutonne après la sélection experte des arcs candidats

Entrée : $\mathcal{X}$
$G^0 = (V, E^0)$, où $E^0 = \bigcup_{i=1}^{n} E^0_{X_i|\mathrm{Pa}_{X_i}}$
$\Theta^0 = \left\{\theta^0_{X_1|\mathrm{Pa}_{X_1}}, \ldots, \theta^0_{X_n|\mathrm{Pa}_{X_n}}\right\}$
$E^{\mathrm{c}} = \bigcup_{i=1}^{n} E^{\mathrm{c}}_{X_i|\mathrm{Pa}_{X_i}}$

1. **pour** $i = 1, \ldots, n$ **faire**
2. $\quad E^{**}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow E^0_{X_i|\mathrm{Pa}_{X_i}}$
3. $\quad \theta^{**}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow \theta^0_{X_i|\mathrm{Pa}_{X_i}}$
4. $\quad$ **répéter**
5. $\quad\quad E^{*}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow E^{**}_{X_i|\mathrm{Pa}_{X_i}}$
6. $\quad\quad \theta^{*}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow \theta^{**}_{X_i|\mathrm{Pa}_{X_i}}$
7. $\quad\quad$ **pour chaque** $u \in E^{\mathrm{c}}_{X_i|\mathrm{Pa}_{X_i}} \backslash E^{*}_{X_i|\mathrm{Pa}_{X_i}}$ **faire**
8. $\quad\quad\quad$ **si** l'ajout de u ne créé pas de circuit **alors**
9. $\quad\quad\quad\quad E^{***}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow E^{*}_{X_i|\mathrm{Pa}_{X_i}} \cup \{u\}$
10. $\quad\quad\quad\quad \theta^{***}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow$ estimation des paramètres de $p(X_i|\mathrm{Pa}_{X_i})$ selon $E^{***}_{X_i|\mathrm{Pa}_{X_i}}$
11. $\quad\quad\quad\quad$ **si** $\mathrm{BIC}\left(\theta^{***}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\right) > \mathrm{BIC}\left(\theta^{**}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\right)$ **alors**
12. $\quad\quad\quad\quad\quad E^{**}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow E^{***}_{X_i|\mathrm{Pa}_{X_i}}$
13. $\quad\quad\quad\quad\quad \theta^{**}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow \theta^{***}_{X_i|\mathrm{Pa}_{X_i}}$
14. $\quad\quad$ **pour chaque** $u \in E^{*}_{X_i|\mathrm{Pa}_{X_i}}$ **faire**
15. $\quad\quad\quad E^{***}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow E^{*}_{X_i|\mathrm{Pa}_{X_i}} \backslash \{u\}$
16. $\quad\quad\quad \theta^{***}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow$ estimation des paramètres de $p(X_i|\mathrm{Pa}_{X_i})$ selon $E^{***}_{X_i|\mathrm{Pa}_{X_i}}$
17. $\quad\quad\quad$ **si** $\mathrm{BIC}\left(\theta^{***}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\right) > \mathrm{BIC}\left(\theta^{**}_{X_i|\mathrm{Pa}_{X_i}}|\mathcal{X}\right)$ **alors**
18. $\quad\quad\quad\quad E^{**}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow E^{***}_{X_i|\mathrm{Pa}_{X_i}}$
19. $\quad\quad\quad\quad \theta^{**}_{X_i|\mathrm{Pa}_{X_i}} \leftarrow \theta^{***}_{X_i|\mathrm{Pa}_{X_i}}$
20. $\quad$ **jusqu'à** $\theta^{**}_{X_i|\mathrm{Pa}_{X_i}} = \theta^{*}_{X_i|\mathrm{Pa}_{X_i}}$
21. $E^* \leftarrow \bigcup_{i=1}^{n} E^{*}_{X_i|\mathrm{Pa}_{X_i}}$
22. $G^* \leftarrow (V, E^*)$
23. $\Theta^* \leftarrow \left\{\theta^*_{X_1|\mathrm{Pa}_{X_1}}, \ldots, \theta^*_{X_n|\mathrm{Pa}_{X_n}}\right\}$
24. **retourner** (G^*, Θ^*)

taines variables ne sont que partiellement observées. Dans la typologie présentée dans la section 1.4, les données manquantes sont catégorisées en données MCAR, MAR et NMAR. Comme le soulignent Naïm *et al.* (2011), les cas MCAR et MAR sont les plus faciles à traiter car les données observées contiennent l'ensemble des informations nécessaires à l'estimation de la distribution des données manquantes. En revanche, le cas NMAR est plus difficile à appréhender et nécessite l'intégration

d'informations externes pour revenir au cas MCAR ou MAR.

2.5.1 Algorithme EM paramétrique

En présence de données MCAR, l'estimation des paramètres d'une distribution conditionnelle locale $p(X_i|\mathrm{Pa}_{X_i})$ peut être réalisée en conservant uniquement les observations pour lesquelles X_i et toutes les variables de Pa_{X_i} sont mesurées. Malgré la perte d'information engendrée par cette méthode, le caractère complètement aléatoire des données manquantes permet d'obtenir des estimations non biaisées. Cependant, quand Pa_{X_i} contient un grand nombre de variables, il devient difficile de trouver des observations complètement mesurées. Des techniques d'imputation simples peuvent alors être employées, telles que celles décrites dans l'ouvrage de Little et Rubin (1987).

Lorsque les données sont MAR, diverses méthodes permettent d'estimer les paramètres d'un réseau bayésien, parmi lesquelles l'algorithme EM (Dempster *et al.*, 1977), l'échantillonnage de Gibbs (Geman et Geman, 1984) ou encore la mise à jour séquentielle des probabilités conditionnelles (Spiegelhalter et Lauritzen, 1990). L'adaptation de l'algorithme EM aux réseaux bayésiens remonte aux travaux de Lauritzen (1995). Comme pour les modèles de mélanges gaussiens (voir la sous-section 2.3.2), cette méthode repose sur l'existence d'un ensemble $\mathcal{Y}$ représentant les données non observées. En partant de paramètres initiaux Θ^0 du réseau bayésien, chaque itération k consiste à calculer (étape E) :

$$Q(\Theta|\Theta^{k-1}) = \mathbb{E}[\ell(\Theta|\mathcal{X},\mathcal{Y})|\mathcal{X},\Theta^{k-1}]\,, \tag{2.36}$$

puis à mettre à jour les paramètres (étape M) :

$$\Theta^k = \arg\max_{\Theta} Q(\Theta|\Theta^{k-1})\,. \tag{2.37}$$

Ces deux étapes sont réitérées jusqu'à ce que la log-vraisemblance converge vers un maximum local.

En pratique, l'étape E consiste à compléter le jeu de données par inférence à partir des données observées et de Θ^{k-1}. L'étape M revient alors à estimer les nouveaux paramètres à l'aide du jeu de données complété en appliquant les algorithmes d'apprentissage décrits dans les sections précédentes. À noter que dans le cas des réseaux bayésiens à mélanges gaussiens, l'apprentissage des paramètres ne permet pas forcément de maximiser globalement $Q(\Theta|\Theta^{k-1})$. Néanmoins, le simple fait d'améliorer cette valeur permet de garantir la convergence monotone de la log-vraisemblance.

D'après l'algorithme EM généralisé de Dempster *et al.* (1977), l'étape M peut en effet être assouplie en substituant la recherche du maximum global par celle de paramètres Θ^k satisfaisant $Q(\Theta^k|\Theta^{k-1}) \geq Q(\Theta^{k-1}|\Theta^{k-1})$.

2.5.2 Algorithme EM structurel

L'algorithme EM paramétrique que nous venons de décrire peut être étendu à l'apprentissage de la structure graphique du réseau bayésien. Proposé par Friedman (1997, 1998), l'algorithme EM structurel s'articule de manière analogue à celui-ci. En partant d'une structure initiale G^0 et de paramètres initiaux Θ^0 du réseau bayésien, chaque itération k consiste à :

— calculer l'espérance du BIC des données complètes à partir des données observées, de la structure G^{k-1} et des paramètres Θ^{k-1} estimés lors de l'itération précédente (étape E) :

$$Q_{\text{BIC}}(G, \Theta|G^{k-1}, \Theta^{k-1}) = \mathbb{E}[\ell(G, \Theta|\mathcal{X}, \mathcal{Y})|\mathcal{X}, G^{k-1}, \Theta^{k-1}] - \frac{\log N}{2} \dim(G, \Theta)\,; \tag{2.38}$$

— déterminer la structure et les paramètres qui maximisent cette espérance (étape M) :

$$(G^k, \Theta^k) = \arg\max_{G,\Theta} Q_{\text{BIC}}(G, \Theta|G^{k-1}, \Theta^{k-1})\,. \tag{2.39}$$

Le BIC converge alors de façon monotone vers un maximum local. À l'instar de l'algorithme EM généralisé, il n'est pas forcément nécessaire de trouver le maximum global de $Q_{\text{BIC}}(G, \Theta|G^{k-1}, \Theta^{k-1})$ pour garantir cette convergence monotone, mais seulement de déterminer une structure G^k et des paramètres Θ^k satisfaisant $Q_{\text{BIC}}(G^k, \Theta^k|G^{k-1}, \Theta^{k-1}) \geq Q_{\text{BIC}}(G^{k-1}, \Theta^{k-1}|G^{k-1}, \Theta^{k-1})$ (Friedman, 1997).

En pratique, l'étape E de l'algorithme EM structurel consiste à appliquer l'algorithme EM paramétrique à partir des données observées, de G^{k-1} et de Θ^{k-1}, de manière à compléter le jeu de données tout en optimisant les paramètres associés à la structure courante. Les données complétées sont ensuite utilisées lors de l'étape M pour mettre à jour la structure et les paramètres du modèle. Cette étape peut être réalisée à l'aide de n'importe quelle méthode d'optimisation de score, en restreignant par exemple l'espace de recherche au voisinage du graphe (Friedman, 1997) ou à l'espace des arbres de recouvrement (Leray et François, 2005). Dans le cadre de notre étude, l'optimisation de la structure est opérée par recherche gloutonne parmi les arcs candidats sélectionnés de manière experte (voir l'algorithme 2.2).

2.5.3 Cas des réseaux bayésiens à mélanges gaussiens

L'algorithme EM structurel consiste à alterner entre l'algorithme EM paramétrique et la mise à jour de la structure par recherche gloutonne. Or au cours de ces deux étapes, les paramètres de chaque distribution conditionnelle locale $p\big(X_i\big|\mathrm{Pa}_{X_i}\big)$ sont estimés $K^{\mathrm{E}} + K^{\mathrm{M}}_{X_i|\mathrm{Pa}_{X_i}} n_{E^{\mathrm{c}}_{X_i|\mathrm{Pa}_{X_i}}}$ fois, où K^{E} est le nombre d'itérations de l'algorithme EM paramétrique, $K^{\mathrm{M}}_{X_i|\mathrm{Pa}_{X_i}}$ le nombre d'itérations de l'algorithme de recherche gloutonne et $n_{E^{\mathrm{c}}_{X_i|\mathrm{Pa}_{X_i}}}$ le nombre d'arcs candidats de la structure locale sélectionnés par l'expert. Dans le cas d'une distribution gaussienne, l'estimation des paramètres est réalisée analytiquement et impacte donc peu la complexité de l'algorithme. En revanche, la charge de calcul est beaucoup plus importante pour un modèle de mélange gaussien dont nous souhaitons à la fois estimer les paramètres et déterminer le nombre optimal de composantes. En effet, l'algorithme EM de division-fusion se comporte lui-même comme un algorithme de recherche gloutonne, dont chaque itération comprend jusqu'à $2C_{\max}$ applications de l'algorithme EM classique et $C_{\max}$ applications de l'algorithme EM partiel (voir l'algorithme 2.1). Toutefois, si nous ne modifions pas le nombre de composantes du modèle de mélange gaussien, l'estimation de ses paramètres se réduit à une unique application de l'algorithme EM classique.

Ces remarques nous conduisent à proposer une nouvelle version de l'algorithme EM structurel, qui s'avère beaucoup moins coûteuse en temps de calcul. Plutôt que d'optimiser le nombre de composantes à chaque estimation des paramètres, cette procédure est réalisée uniquement dans le cadre d'une étape distincte introduite après la mise à jour de la structure. L'ajout de cette nouvelle étape peut être vu comme un prolongement de l'étape M, où la mise à jour de la structure ne concerne plus seulement la structure graphique, mais également celle des modèles de mélanges gaussiens. Ainsi, pour chaque distribution conditionnelle locale $p\big(X_i\big|\mathrm{Pa}_{X_i}\big)$, l'algorithme EM de division-fusion n'est appliqué qu'une seule fois par itération, contre $K^{\mathrm{E}} + K^{\mathrm{M}}_{X_i|\mathrm{Pa}_{X_i}} n_{E^{\mathrm{c}}_{X_i|\mathrm{Pa}_{X_i}}}$ fois dans la version initiale de l'algorithme EM structurel. Décrite dans l'algorithme 2.3, cette version alternative est testée dans l'expérimentation du chapitre 4.

2.6 Réseaux bayésiens dynamiques

Dans de nombreux domaines d'application, l'utilisation d'un modèle statique se révèle insuffisante, dans la mesure où le système étudié évolue au cours du temps.

ALGORITHME 2.3 – Version alternative de l'algorithme EM structurel pour l'apprentissage d'un réseau bayésien à mélanges gaussiens avec des données incomplètes

Entrée : $\mathcal{X}$
G^0
Θ^0
E^{c}
$C_{\max}$

1. **pour** $k = 1, 2, \ldots$ jusqu'à convergence **faire**
2. $\quad\Theta^{k-1,0} \leftarrow \Theta^{k-1}$
3. $\quad$**pour** $k' = 1, 2, \ldots$ jusqu'à convergence **faire**
4. $\quad\quad\hat{\mathcal{Y}} \leftarrow \mathbb{E}[\mathcal{Y}|\mathcal{X}, G^{k-1}, \Theta^{k-1,k'-1}]$
5. $\quad\quad\Theta^{k-1,k'} \leftarrow$ estimation des paramètres par l'algorithme EM classique à partir de $(\mathcal{X}, \hat{\mathcal{Y}})$ et $\Theta^{k-1,k'-1}$
6. $\quad\left(G^k, \Theta^k\right) \leftarrow$ mise à jour de la structure par recherche gloutonne à partir de $(\mathcal{X}, \hat{\mathcal{Y}})$, G^{k-1}, $\Theta^{k-1,k'}$ et E^{c} (voir l'algorithme 2.2), avec estimation des paramètres par l'algorithme EM classique
7. $\quad\Theta^k \leftarrow$ mise à jour du nombre de composantes et estimation des paramètres par l'algorithme EM de division-fusion à partir de $(\mathcal{X}, \hat{\mathcal{Y}})$, Θ^k et $C_{\max}$ (voir l'algorithme 2.1)
8. **retourner** (G^k, Θ^k)

Introduits par Dean et Kanazawa (1989), les réseaux bayésiens dynamiques étendent le formalisme des réseaux bayésiens à la modélisation des relations temporelles entre les variables. En supposant que le système est observé à travers une séquence de pas de temps discrets, chaque nœud est associé à une variable X_i^t représentant l'instanciation de X_i au pas de temps t. Dans la suite de ce chapitre, l'ensemble des variables décrivant l'état du système à t est noté $X^t = \{X_1^t, \ldots, X_n^t\}$.

2.6.1 Réseaux bayésiens dynamiques d'ordre 1

D'après la définition donnée par Murphy (2002), un réseau bayésien dynamique d'ordre 1 est un couple $(\mathcal{B}_1, \mathcal{B}_{\rightarrow})$ tel que :

— $\mathcal{B}_1$ est un réseau bayésien représentant la distribution initiale :

$$p(X^1) = \prod_{i=1}^{n} p\left(X_i^1 \middle| \mathrm{Pa}_{X_i^1}\right), \tag{2.40}$$

où $\mathrm{Pa}_{X_i^1} \subset X^1$;

— $\mathcal{B}_{\rightarrow}$ est un réseau bayésien à deux pas de temps (2TBN) représentant, pour tout $t \geq 2$, la transition du pas de temps $t-1$ au pas de temps t, c'est-à-dire

la distribution :

$$p(X^t|X^{t-1}) = \prod_{i=1}^{n} p\big(X_i^t \big| \mathrm{Pa}_{X_i^t}\big)\,, \tag{2.41}$$

où $\mathrm{Pa}_{X_i^t} \subset X^{t-1} \cup X^t$.

La distribution jointe sur une séquence de T pas de temps s'obtient en « déroulant » $\mathcal{B}_{\rightarrow}$ autant de fois que nécessaire :

$$\begin{aligned} p(X^1, \ldots, X^T) &= p(X^1) \prod_{t=2}^{T} p(X^t|X^{t-1}) \\ &= \prod_{t=1}^{T} \prod_{i=1}^{n} p\big(X_i^t \big| \mathrm{Pa}_{X_i^t}\big)\,. \end{aligned} \tag{2.42}$$

2.6.2 Réseaux bayésiens dynamiques d'ordre r

La définition donnée par Murphy (2002) est souvent présentée comme la définition de base des réseaux bayésiens dynamiques. Cependant, elle repose sur l'hypothèse que le processus représenté est markovien d'ordre 1. Or à l'image des travaux de Hulst (2006), il arrive d'être confronté à des processus d'ordres plus élevés, notamment quand l'échelle temporelle est plus fine. Afin de modéliser convenablement ces processus, il est nécessaire d'étendre la définition de Murphy au cas plus général des réseaux bayésiens dynamiques d'ordre r, pour tout $r \geq 1$.

Un réseau bayésien dynamique d'ordre r peut être défini comme un $(r+1)$-uplet $(\mathcal{B}_1, \ldots, \mathcal{B}_r, \mathcal{B}_{\rightarrow})$ tel que :

— $\mathcal{B}_1$ est un réseau bayésien représentant la distribution initiale $p(X^1)$;

— si $r \geq 2$, alors pour tout $t \in \{2, \ldots, r\}$, $\mathcal{B}_t$ est un tTBN représentant la distribution de transition initiale :

$$p(X^t|X^{t-1}, \ldots, X^1) = \prod_{i=1}^{n} p\big(X_i^t \big| \mathrm{Pa}_{X_i^t}\big)\,, \tag{2.43}$$

où $\mathrm{Pa}_{X_i^t} \subset \bigcup_{k=1}^{t} X^k$;

— $\mathcal{B}_{\rightarrow}$ est un $(r+1)$TBN représentant, pour tout $t > r$, la distribution de transition :

$$p(X^t|X^{t-1}, \ldots, X^{t-r}) = \prod_{i=1}^{n} p\big(X_i^t \big| \mathrm{Pa}_{X_i^t}\big)\,, \tag{2.44}$$

où $\mathrm{Pa}_{X_i^t} \subset \bigcup_{k=t-r}^{t} X^k$.

La distribution jointe sur une séquence de T pas de temps s'obtient de manière analogue à l'équation (2.42). À noter qu'une définition équivalente peut être trouvée

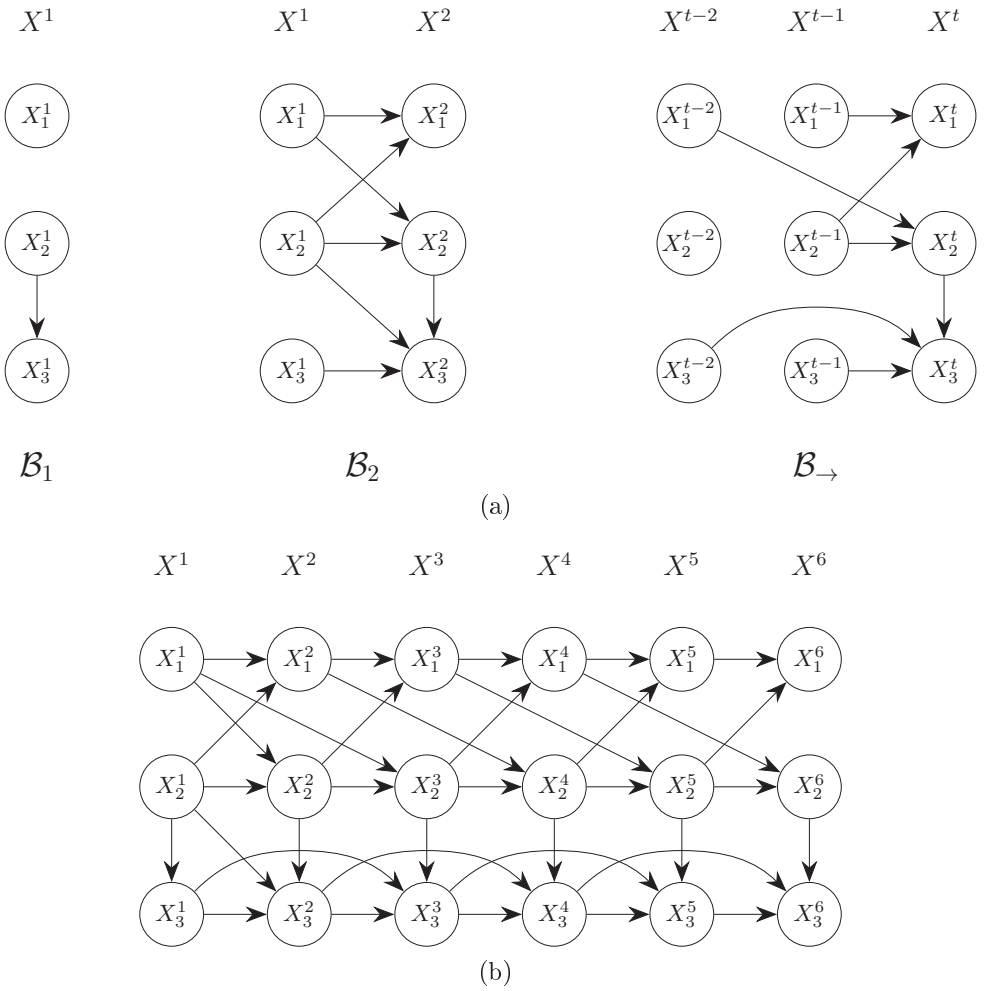


FIGURE 2.4 – (a) Exemple de réseau bayésien dynamique d'ordre 2. (b) Déroulement du réseau bayésien dynamique sur six pas de temps.

dans la thèse de Trabelsi (2013), où les réseaux bayésiens $\mathcal{B}_1, \ldots, \mathcal{B}_r$ sont présentés sous la forme d'un unique réseau bayésien dynamique initial représentant la distribution $p(X^1, \ldots, X^r)$.

Un exemple de réseau bayésien dynamique d'ordre 2 est donné dans la figure 2.4. Ce dernier est caractérisé par les réseaux bayésiens initiaux $\mathcal{B}_1$ et $\mathcal{B}_2$ ainsi que par le réseau bayésien de transition $\mathcal{B}_\rightarrow$ (voir la figure 2.4a). La version « déroulée » du modèle illustre bien le rôle central de $\mathcal{B}_\rightarrow$, dont la structure est répliquée invariablement à partir du troisième pas de temps (voir la figure 2.4b).

Notre définition des réseaux bayésiens dynamiques repose sur l'hypothèse que le processus modélisé est stationnaire. Autrement dit, ni la structure graphique ni les paramètres des distributions de probabilité ne varient au cours du temps. Cette hypothèse est retenue dans la plupart des travaux et permet de décrire le modèle à

partir d'un nombre restreint de paramètres. Toutefois, certains auteurs s'intéressent à des formalismes plus souples autorisant les structures évolutives (Song *et al.*, 2009 ; Robinson et Hartemink, 2010).

2.6.3 Apprentissage des paramètres

Soit $\mathcal{B} = (\mathcal{B}_1, \ldots, \mathcal{B}_r, \mathcal{B}_{\rightarrow})$ un réseau bayésien dynamique d'ordre r dont nous souhaitons apprendre les paramètres $\Theta = (\Theta_1, \ldots, \Theta_r, \Theta_{\rightarrow})$. Nous disposons pour cela d'un ensemble de données $\mathcal{X}$ composé de N_{seq} séquences d'observations complètes, telles que la s-ème séquence $x^{s,1}, \ldots, x^{s,T_s}$ $(T_s > r)$ spécifie respectivement des valeurs de $X^1, \ldots, X^{T_s}$. La log-vraisemblance de Θ se décompose ainsi :

$$\begin{aligned}
\ell(\Theta|\mathcal{X}) &= \sum_{s=1}^{N_{\text{seq}}} \log p(x^{s,1}, \ldots, x^{s,T_s}|\Theta) \\
&= \sum_{s=1}^{N_{\text{seq}}} \log p(x^{s,1}|\Theta) + \sum_{s=1}^{N_{\text{seq}}} \sum_{t=2}^{r} \log p(x^{s,t}|x^{s,t-1}, \ldots, x^{s,1}, \Theta) \\
&\quad + \sum_{s=1}^{N_{\text{seq}}} \sum_{t=r+1}^{T_s} \log p(x^{s,t}|x^{s,t-1}, \ldots, x^{s,t-r}, \Theta) \\
&= \sum_{t=1}^{r} \ell(\Theta_t|\mathcal{X}) + \ell(\Theta_{\rightarrow}|\mathcal{X})\,.
\end{aligned} \tag{2.45}$$

Mise en évidence par Friedman *et al.* (1998) pour les réseaux bayésiens dynamiques d'ordre 1, cette propriété de décomposition se généralise donc aux ordres plus élevés. Maximiser la log-vraisemblance de Θ équivaut donc à maximiser les log-vraisemblances de $\Theta_1, \ldots, \Theta_r$ et $\Theta_{\rightarrow}$. En d'autres termes, l'apprentissage des paramètres de $\mathcal{B}$ revient à apprendre indépendamment les paramètres de $\mathcal{B}_1, \ldots, \mathcal{B}_r$ et $\mathcal{B}_{\rightarrow}$. L'ensemble de données $\mathcal{X}$ est alors divisé de sorte que :

- les N_{seq} observations du pas de temps 1 sont utilisées pour estimer Θ_1 ;
- si $r \geq 2$, alors pour tout $t \in \{2, \ldots, r\}$, les N_{seq} instances de transition des pas de temps $1, \ldots, t-1$ au pas de temps t sont utilisées pour estimer Θ_t ;
- les $N_{\rightarrow} = \sum_{s=1}^{N_{\text{seq}}} (T_s - r)$ instances de transition restantes sont utilisées pour estimer $\Theta_{\rightarrow}$.

En présence de données incomplètes, l'algorithme EM paramétrique présenté dans la sous-section 2.5.1 peut être aisément étendu aux réseaux bayésiens dynamiques. Une fois les données complétées par inférence (étape E), l'étape M est réalisée en appliquant la propriété de décomposition de la log-vraisemblance que nous venons de décrire.

2.6.4 Apprentissage de la structure

Nous souhaitons apprendre la structure graphique $G = (G_1, \ldots, G_r, G_{\rightarrow})$ du réseau bayésien dynamique $\mathcal{B}$ défini dans la sous-section précédente. La méthode utilisée résulte d'une simple extension de la méthode d'apprentissage des paramètres. En effet, la propriété de décomposition de la log-vraisemblance permet de décomposer le BIC de manière analogue (Friedman *et al.*, 1998) :

$$\mathrm{BIC}(G, \Theta|\mathcal{X}) = \sum_{t=1}^{r} \mathrm{BIC}(G_t, \Theta_t|\mathcal{X}) + \mathrm{BIC}(G_{\rightarrow}, \Theta_{\rightarrow}|\mathcal{X}), \tag{2.46}$$

où, pour tout $t \in \{1, \ldots, r\}$:

$$\mathrm{BIC}(G_t, \Theta_t|\mathcal{X}) = \ell(G_t, \Theta_t|\mathcal{X}) - \frac{\log N_{\mathrm{seq}}}{2} \dim(G_t, \Theta_t) \tag{2.47}$$

et :

$$\mathrm{BIC}(G_{\rightarrow}, \Theta_{\rightarrow}|\mathcal{X}) = \ell(G_{\rightarrow}, \Theta_{\rightarrow}|\mathcal{X}) - \frac{\log N_{\rightarrow}}{2} \dim(G_{\rightarrow}, \Theta_{\rightarrow}). \tag{2.48}$$

Ainsi, la structure et les paramètres de $\mathcal{B}$ sont appris en maximisant indépendamment le BIC de $\mathcal{B}_1, \ldots, \mathcal{B}_r$ et $\mathcal{B}_{\rightarrow}$. Les méthodes d'optimisation de score présentées dans la section 2.4 peuvent alors être appliquées à chacun de ces réseaux bayésiens.

À noter que certaines méthodes d'apprentissage de la structure sont spécifiquement développées dans le cadre des réseaux bayésiens dynamiques. Pour une revue récente et un benchmark de ces méthodes, nous référons le lecteur à la thèse de Trabelsi (2013).

En présence de données incomplètes, l'algorithme EM structurel présenté dans la sous-section 2.5.2 peut être aisément étendu aux réseaux bayésiens dynamiques, tout comme sa version alternative détaillée dans l'algorithme 2.3. Une fois les données complétées à l'aide de l'algorithme EM paramétrique (étape E), l'étape M est réalisée en appliquant la propriété de décomposition du BIC que nous venons de décrire (Friedman *et al.*, 1998).

2.6.5 Inférence

En présence de données incomplètes, nous avons vu que l'apprentissage d'un réseau bayésien dynamique pouvait être réalisé à l'aide des algorithmes EM paramétrique et structurel. Or en pratique, l'étape E de ces algorithmes consiste à compléter les données manquantes à partir des données observées et du modèle estimé lors de l'itération précédente (l'algorithme EM structurel faisant alors appel à l'algorithme EM paramétrique). Étant donnée une séquence d'observations incomplètes $o^1, \ldots, o^T$

telle que o^t est l'ensemble des valeurs des variables observées O^t au pas de temps t, l'objectif est de calculer pour tout $t \in \{1, \ldots, T\}$:

$$\hat{m}^t = \mathbb{E}[M^t | o^1, \ldots, o^T], \tag{2.49}$$

où M^t est l'ensemble des variables non observées à t.

Une fois le modèle appris, celui-ci peut être exploité à des fins de prédiction pour déterminer les états futurs du système étudié compte tenu de ses états passés et actuel. Supposons qu'à chaque pas de temps t, nous disposons d'une séquence d'observations $o^1, \ldots, o^t$ (possiblement incomplètes) et nous souhaitons prédire en temps réel les valeurs d'un sous-ensemble de variables Y^{t+h} au pas de temps $t+h$ ($h \geq 1$). Supposons également que pour tout pas de temps $k \in \{t+1, \ldots, t+h\}$, les valeurs z^k d'un sous-ensemble de variables Z^k sont connues a priori ($Z^{t+h} \cap Y^{t+h} = \emptyset$). La démarche de prédiction revient alors à calculer :

$$\hat{y}^{t+h} = \mathbb{E}[Y^{t+h} | o^1, \ldots, o^t, z^{t+1}, \ldots, z^{t+h}]. \tag{2.50}$$

Bien que ces procédures soient distinctes, elles constituent toutes les deux des problèmes d'inférence dans le réseau bayésien dynamique. Une grande diversité de méthodes exactes et approchées sont conçues pour réaliser ce type de tâche (Murphy, 2002 ; Koller et Friedman, 2009). En pratique, le coût des algorithmes d'inférence exacte se révèle souvent prohibitif lorsque le nombre de variables est élevé. Par ailleurs, l'applicabilité de ces algorithmes aux réseaux bayésiens continus se limite généralement au cas gaussien. Afin de traiter de larges réseaux bayésiens dynamiques et des distributions plus complexes, il est préférable d'utiliser des algorithmes d'inférence approchée ne requérant pas d'hypothèse paramétrique sur les distributions considérées (Koller et Friedman, 2009).

Le filtre bootstrap est une méthode d'inférence approchée couramment utilisée en présence de systèmes dynamiques non linéaires. Initialement proposée pour les processus markoviens d'ordre 1 (Gordon *et al.*, 1993), cette méthode peut être étendue aux processus d'ordres plus élevés (Pan et Schonfeld, 2011). Son principe est de propager dans le temps plusieurs séquences d'échantillons pondérées, ou « particules », construites en échantillonnant les variables non observées. Afin d'illustrer ce processus de propagation, nous considérons un ensemble de N_{par} particules $\bar{m}^{t-1} = \{\bar{m}^{t-1,1}, \ldots, \bar{m}^{t-1,N_{\text{par}}}\}$ propagées jusqu'au pas de temps $t-1$ et associées à un ensemble de poids $\dot{w}^{t-1} = \{\dot{w}^{t-1,1}, \ldots, \dot{w}^{t-1,N_{\text{par}}}\}$, tel que $\sum_{q=1}^{N_{\text{par}}} \dot{w}^{t-1,q} = 1$. Chaque particule $\bar{m}^{t-1,q}$ se caractérise par un ensemble $\{m^{t-1,q,1}, \ldots, m^{t-1,q,t-1}\}$, où $m^{t-1,q,k}$ contient les échantillons des variables de M^k. À partir de l'ensemble des

valeurs observées jusqu'au pas de temps t $\bar{o}^t = \{o^1, \ldots, o^t\}$, notre objectif est de poursuivre la propagation des particules vers t, c'est-à-dire de construire un nouvel ensemble $\bar{m}^t = \{\bar{m}^{t,1}, \ldots, \bar{m}^{t,N_{\text{par}}}\}$.

Dans un premier temps, un ensemble $\{\bar{\bar{m}}^{t-1,1}, \ldots, \bar{\bar{m}}^{t-1,N_{\text{par}}}\}$ est construit en tirant aléatoirement (avec remise) N_{par} particules dans $\bar{m}^{t-1}$. Chaque particule $\bar{m}^{t-1,q}$ est sélectionnée avec une probabilité égale à son poids $\dot{w}^{t-1,q}$. Plus celui-ci est élevé, plus elle a de chances d'être répliquée. À l'inverse, les particules dont le poids est plus faible tendent naturellement à disparaître. Dans un second temps, le nouvel ensemble $\bar{m}^t$ est construit en ajoutant à chaque particule $\bar{\bar{m}}^{t-1,q}$ un ensemble d'échantillons pour les variables de M^t. En supposant que $M^t \neq \emptyset$, nous notons $M_1^t, \ldots, M_{n_{M^t}}^t$ ces variables selon un ordre topologique du réseau bayésien dynamique. Pour chaque variable M_i^t, un échantillon $m_i^{t,q,t}$ est généré aléatoirement selon la distribution $p\big(M_i^t \big| \text{pa}_{M_i^t}^q\big)$, où $\text{pa}_{M_i^t}^q$ désigne les valeurs des parents de M_i^t prises parmi les valeurs observées de $\bar{o}^t$, les échantillons de $\bar{\bar{m}}^{t-1,q}$ et les échantillons précédemment générés $m_1^{t,q,t}, \ldots, m_{i-1}^{t,q,t}$ ($M_1^t, \ldots, M_{n_{M^t}}^t$ doivent donc être échantillonnées dans cet ordre). La particule est finalement complétée :

$$\bar{m}^{t,q} = \bar{\bar{m}}^{t-1,q} \cup \left\{m^{t,q,t}\right\}, \tag{2.51}$$

où $m^{t,q,t} = \left\{m_1^{t,q,t}, \ldots, m_{n_{M^t}}^{t,q,t}\right\}$.

La dernière étape consiste à mettre à jour le poids de chaque particule selon la vraisemblance des valeurs observées à t compte tenu des échantillons générés. En notant $O_1^t, \ldots, O_{n_{O^t}}^t$ les variables de O^t (si $O^t \neq \emptyset$) et $o_1^t, \ldots, o_{n_{O^t}}^t$ leurs valeurs respectives, le poids de $\bar{m}^{t,q}$ est calculé de la manière suivante :

$$w^{t,q} = \prod_{i=1}^{n_{O^t}} p\big(o_i^t \big| \text{pa}_{O_i^t}^q\big), \tag{2.52}$$

où $\text{pa}_{O_i^t}^q$ désigne les valeurs des parents de O_i^t prises parmi les valeurs observées de $\bar{o}^t$ et les échantillons de $\bar{m}^{t,q}$. Ce poids est ensuite normalisé :

$$\dot{w}^{t,q} = \frac{w^{t,q}}{\sum_{l=1}^{N_{\text{par}}} w^{t,l}}. \tag{2.53}$$

Nous disposons finalement d'un nouvel ensemble de particules $\bar{m}^t$ auquel est associé un ensemble de poids $\dot{w}^t = \{\dot{w}^{t,1}, \ldots, \dot{w}^{t,N_{\text{par}}}\}$ (voir l'algorithme 2.4). La sélection aléatoire opérée au début de ce processus permet de propager en priorité les particules qui expliquent le mieux les valeurs mesurées. Cette étape a pour but de pallier les problèmes de dégénérescence liés au caractère aléatoire de l'échantillonnage (Koller et Friedman, 2009).

ALGORITHME 2.4 – Propagation d'un ensemble de particules pondérées du pas de temps $t-1$ au pas de temps t

Entrée : $\bar{m}^{t-1} = \{\bar{m}^{t-1,1}, \ldots, \bar{m}^{t-1,N_{\text{par}}}\}$
$\dot{w}^{t-1} = \{\dot{w}^{t-1,1}, \ldots, \dot{w}^{t-1,N_{\text{par}}}\}$
$\bar{o}^t = \{o^1, \ldots, o^t\}$, où $o^t = \{o_1^t, \ldots, o_{n_{O^t}}^t\}$ (si $O^t \neq \emptyset$)
$\mathcal{B}$

1. $\{\bar{\bar{m}}^{t-1,1}, \ldots, \bar{\bar{m}}^{t-1,N_{\text{par}}}\} \leftarrow$ tirage aléatoire dans $\bar{m}^{t-1}$ pondéré selon $\dot{w}^{t-1}$
2. $M^t \leftarrow$ variables non observées à t
3. **si** $M^t \neq \emptyset$ **alors**
4. $\quad (M_1^t, \ldots, M_{n_{M^t}}^t) \leftarrow$ classement des variables de M^t dans un ordre topologique de $\mathcal{B}$
5. **pour** $q = 1, \ldots, N_{\text{par}}$ **faire**
6. $\quad$ **si** $M^t \neq \emptyset$ **alors**
7. $\quad\quad$ **pour** $i = 1, \ldots, n_{M^t}$ **faire**
8. $\quad\quad\quad m_i^{t,q,t} \leftarrow$ génération aléatoire d'un échantillon selon $p\big(M_i^t \big| \text{pa}_{M_i^t}^q\big)$
9. $\quad\quad m^{t,q,t} \leftarrow \{\bar{m}_1^{t,q,t}, \ldots, \bar{m}_{n_{M^t}}^{t,q,t}\}$
10. $\quad$ **sinon**
11. $\quad\quad m^{t,q,t} \leftarrow \emptyset$
12. $\quad \bar{m}^{t,q} \leftarrow \bar{\bar{m}}^{t-1,q} \cup \{m^{t,q,t}\}$
13. $\quad$ **si** $O^t \neq \emptyset$ **alors**
14. $\quad\quad w^{t,q} \leftarrow \prod_{i=1}^{n_{O^t}} p\big(o_i^t \big| \text{pa}_{O_i^t}^q\big)$
15. $\quad$ **sinon**
16. $\quad\quad w^{t,q} \leftarrow \frac{1}{N_{\text{par}}}$
17. **pour** $q = 1, \ldots, N_{\text{par}}$ **faire**
18. $\quad \dot{w}^{t,q} \leftarrow \frac{w^{t,q}}{\sum_{l=1}^{N_{\text{par}}} w^{t,l}}$
19. $\bar{m}^t \leftarrow \{\bar{m}^{t,1}, \ldots, \bar{m}^{t,N_{\text{par}}}\}$
20. $\dot{w}^t \leftarrow \{\dot{w}^{t,1}, \ldots, \dot{w}^{t,N_{\text{par}}}\}$
21. **retourner** $(\bar{m}^t, \dot{w}^t)$

Le filtre bootstrap s'applique aisément aux problèmes d'inférence exposés au début de cette sous-section. Dans le cas des algorithmes EM paramétrique et structurel, l'estimation des valeurs manquantes est réalisée une fois que les particules ont été propagées jusqu'au pas de temps T. Pour tout $t \in \{1, \ldots, T\}$, les valeurs des variables de M^t sont estimées en calculant :

$$\hat{m}^t \approx \sum_{q=1}^{N_{\text{par}}} \dot{w}^{T,q} m^{T,q,t}\,, \tag{2.54}$$

Décrite dans l'algorithme 2.5, cette méthode est analogue à l'algorithme de lissage proposé par Isard et Blake (1998).

ALGORITHME 2.5 – Filtre bootstrap pour l'estimation des valeurs manquantes

Entrée : $o^1, \ldots, o^T$
$\mathcal{B}$
N_{par}

1. $\bar{o}^0 \leftarrow \emptyset$
2. $\bar{m}^0 \leftarrow \{\emptyset, \ldots, \emptyset\}$ de taille N_{par}
3. $\dot{w}^0 \leftarrow \left\{\frac{1}{N_{\text{par}}}, \ldots, \frac{1}{N_{\text{par}}}\right\}$ de taille N_{par}
4. **pour** $t = 1, \ldots, T$ **faire**
5. $\quad \bar{o}^t \leftarrow \bar{o}^{t-1} \cup \{o^t\}$
6. $\quad (\bar{m}^t, \dot{w}^t) \leftarrow$ propagation des particules de $t-1$ à t à partir de $\bar{m}^{t-1}$, $\dot{w}^{t-1}$, $\bar{o}^t$ et $\mathcal{B}$ (voir l'algorithme 2.4)
7. **pour** $t = 1, \ldots, T$ **faire**
8. $\quad \hat{m}^t \leftarrow \sum_{q=1}^{N_{\text{par}}} \dot{w}^{T,q} m^{T,q,t}$
9. **retourner** $\{\hat{m}^1, \ldots, \hat{m}^T\}$

La démarche de prédiction consiste à propager un ensemble de particules au fur et à mesure de la réception des données. Ainsi, les observations $o^1, \ldots, o^t$ permettent de propager ces particules jusqu'au pas de temps t. Pour estimer les valeurs des variables de Y^{t+h}, il suffit alors de poursuivre « virtuellement » cette propagation jusqu'au pas de temps $t+h$ en considérant $z^{t+1}, \ldots, z^{t+h}$ comme des ensembles de valeurs observées (au même titre que $o^1, \ldots, o^t$). Les valeurs prédites s'obtiennent en calculant :

$$\hat{y}^{t+h} \approx \sum_{q=1}^{N_{\text{par}}} \dot{w}^{t+h,q} y^{t+h,q,t+h}, \tag{2.55}$$

où $y^{t+h,q,t+h}$ est le sous-ensemble des échantillons de $m^{t+h,q,t+h}$ générés pour les variables de Y^{t+h} (voir l'algorithme 2.6). Il est important de souligner que $\hat{y}^{t+h}$ est calculée dès le pas de temps t. Par conséquent, cette méthode de prédiction peut être mise en œuvre en temps réel, ce qui constitue un critère fondamental de notre étude. Pour un réseau bayésien dynamique d'ordre r, une fois que les particules ont été propagées jusqu'à t, les échantillons (de même que les données observées) antérieurs au pas de temps $t-r+1$ ne sont plus d'aucune utilité dans le processus. Ces derniers n'ont donc plus besoin d'être stockés, ce qui permet de ne pas augmenter la mémoire requise au fil du temps.

ALGORITHME 2.6 – Filtre bootstrap pour la prédiction en temps réel

À l'initialisation de l'algorithme :

Entrée : Y^h
$z^1, \ldots, z^h$
$\mathcal{B}$
N_{par}

1. $\bar{o}^0 \leftarrow \emptyset$
2. $\bar{m}^0 \leftarrow \{\emptyset, \ldots, \emptyset\}$ de taille N_{par}
3. $\dot{w}^0 \leftarrow \left\{\frac{1}{N_{\text{par}}}, \ldots, \frac{1}{N_{\text{par}}}\right\}$ de taille N_{par}
4. $\bar{z}^0 \leftarrow \emptyset$
5. **pour** $k = 1, \ldots, h$ **faire**
6. $\quad \bar{z}^k \leftarrow \bar{z}^{k-1} \cup \{z^k\}$
7. $\quad (\bar{m}^k, \dot{w}^k) \leftarrow$ propagation des particules de $k-1$ à k à partir de $\bar{m}^{k-1}$, $\dot{w}^{k-1}$, $\bar{z}^k$ et $\mathcal{B}$ (voir l'algorithme 2.4)
8. $\hat{y}^h \leftarrow \sum_{q=1}^{N_{\text{par}}} \dot{w}^{h,q} y^{h,q,h}$
9. **retourner** $\hat{y}^h$

À chaque pas de temps $t \geq 1$:

Entrée : Y^{t+h}
o^t
$z^{t+1}, \ldots, z^{t+h}$

1. $\bar{o}^t \leftarrow \bar{o}^{t-1} \cup \{o^t\}$
2. $(\bar{m}^t, \dot{w}^t) \leftarrow$ propagation des particules de $t-1$ à t à partir de $\bar{m}^{t-1}$, $\dot{w}^{t-1}$, $\bar{o}^t$ et $\mathcal{B}$ (voir l'algorithme 2.4)
3. $\bar{z}^t \leftarrow \bar{o}^t$
4. **pour** $k = t+1, \ldots, t+h$ **faire**
5. $\quad \bar{z}^k \leftarrow \bar{z}^{k-1} \cup \{z^k\}$
6. $\quad (\bar{m}^k, \dot{w}^k) \leftarrow$ propagation des particules de $k-1$ à k à partir de $\bar{m}^{k-1}$, $\dot{w}^{k-1}$, $\bar{z}^k$ et $\mathcal{B}$ (voir l'algorithme 2.4)
7. $\hat{y}^{t+h} \leftarrow \sum_{q=1}^{N_{\text{par}}} \dot{w}^{t+h,q} y^{t+h,q,t+h}$
8. **retourner** $\hat{y}^{t+h}$

Chapitre 3

Données et construction du modèle

Dans le chapitre 2, nous avons présenté les algorithmes d'apprentissage et d'inférence associés aux réseaux bayésiens et à leur extension dynamique. À présent, nous souhaitons exploiter ce formalisme pour prédire à court terme les flux de voyageurs. Les données utilisées sont collectées sur le réseau ferré (métro et RER) de la RATP et issues de trois sources distinctes : les validations des titres de transport, les comptages par pesée des voyageurs à bord des trains et l'offre de transport. Dans ce chapitre, nous procédons à une description détaillée de ces données, après avoir défini un référentiel spatial permettant de les associer. Basés sur les relations de causalité spatio-temporelles entre les flux, les mécanismes de construction du modèle sont exposés dans la section 3.4. L'offre de transport est intégrée par l'intermédiaire des intervalles de départ des trains, dont l'impact sur les flux est mis en évidence dans une courte expérimentation réalisée sur la gare de Nanterre-Préfecture.

3.1 Définition du référentiel spatial

3.1.1 Description spatiale du réseau

La RATP collecte une grande diversité de données sur la mobilité des voyageurs au sein de son réseau. La plupart de ces données sont issues des validations des titres de transport, de comptages manuels ou automatiques, ou encore d'enquêtes origine-destination. En tant qu'opérateur de transport public, la RATP possède également des données sur l'offre de transport qu'elle réalise, c'est-à-dire sur les horaires d'arrivée et de départ des véhicules à chaque point d'arrêt. Elle dispose enfin de nombreuses sources d'information internes et externes susceptibles de mieux expliquer la fréquentation de son réseau (calendrier, perturbations d'exploitation, données issues des autres réseaux, conditions météorologiques, événements sportifs

et culturels, études socio-économiques, etc.).

Cette diversité des données s'accompagne d'une hétérogénéité des référentiels qui leur sont associés. Bien souvent, ces référentiels possèdent des architectures et des nomenclatures qui leur sont propres. Ils ne font pas toujours intervenir les mêmes concepts spatiaux (stations, espaces contrôlés, points d'arrêt, etc.), ce qui rend leur unification difficile. Or si nous souhaitons combiner plusieurs sources de données au sein d'un même modèle, nous devons disposer d'un référentiel spatial suffisamment flexible pour que chaque source puisse trouver sa place. Pour définir un tel référentiel, il convient d'introduire un certain nombre de notions permettant de décrire les éléments qui composent le réseau de transport public. La description spatiale que nous proposons s'articule autour de deux concepts topologiques simples et génériques : les zones et les accès.

Zone : Une zone est un espace délimité accessible aux voyageurs. Il s'agit soit d'une zone piétonne, soit d'une zone embarquée.

Zone piétonne : Une zone piétonne est une zone où les voyageurs circulent à pied. Elle est caractérisée par un emplacement géographique unique.

Zone embarquée : Une zone embarquée est une zone où les voyageurs sont à bord d'un véhicule. Il s'agit soit d'un point d'arrêt, soit d'un tronçon. Exploitée par un opérateur, une zone embarquée appartient à une ligne et à un sens de circulation. Son emplacement géographique dépend de la position des véhicules circulant sur cette ligne et dans ce sens. Par conséquent, une même zone embarquée peut être caractérisée par plusieurs emplacements géographiques différents (par exemple, les voies 2 et 4 de la station Porte de Versailles peuvent être attribuées toutes les deux au point d'arrêt de la ligne 12 du métro en direction de Mairie d'Issy). À l'inverse, un même emplacement peut être attribué à plusieurs zone embarquées différentes (par exemple, la voie 44 de Gare du Nord peut être attribuée à un point d'arrêt du RER B ou D).

Point d'arrêt : Un point d'arrêt est une zone embarquée où les voyageurs peuvent monter à bord ou descendre d'un véhicule.

Tronçon : Un tronçon est une zone embarquée située entre deux points d'arrêt successifs d'une même ligne et d'un même sens de circulation, où les voyageurs ne peuvent pas monter à bord ni descendre d'un véhicule.

Accès : Un accès représente le passage d'une zone à une autre qui lui est contiguë. Il s'agit d'un accès piéton, d'un accès d'embarquement, d'un accès de

débarquement ou d'un accès embarqué.

Accès piéton : Un accès piéton est un accès d'une zone piétonne à une autre.

Accès d'embarquement : Un accès d'embarquement est un accès d'une zone piétonne à un point d'arrêt.

Accès de débarquement : Un accès de débarquement est un accès d'un point d'arrêt à une zone piétonne.

Accès embarqué : Un accès embarqué est un accès d'une zone embarquée à une autre. Il s'agit soit d'un accès de départ, soit d'un accès d'arrivée.

Accès de départ : Un accès de départ est un accès embarqué d'un point d'arrêt à un tronçon.

Accès d'arrivée : Un accès d'arrivée est un accès embarqué d'un tronçon à un point d'arrêt.

Station : Une station est un ensemble de zones regroupées sous une appellation commune, permettant aux voyageurs d'emprunter, de quitter les transports publics ou d'effectuer une correspondance.

Mode de transport : Un mode de transport est un moyen de locomotion caractérisé par un type particulier de véhicule et d'infrastructure.

Ligne : Une ligne est un ensemble de zones embarquées regroupées sous une appellation commune, décrivant un ou plusieurs chemins parcourus par des véhicules d'un même mode de transport.

Opérateur : Un opérateur est une entreprise qui fournit une offre de transport public.

La figure 3.1 permet de visualiser un exemple de zones et d'accès du réseau de transport public. Par son caractère générique, cette méthode de description des espaces est applicable quel que soit le mode de transport et la configuration spatiale des lieux considérés. Elle permet ainsi de représenter un vaste réseau de transport multimodal en un seul et même système cohérent. Naturellement, le découpage des zones doit être opéré de manière intelligente en tenant compte par exemple de la position des dispositifs de collecte des données. Si nous souhaitons faire coïncider un de ces dispositifs avec un accès, il est toujours possible de subdiviser la zone dans laquelle il se trouve pour créer ce nouvel accès.

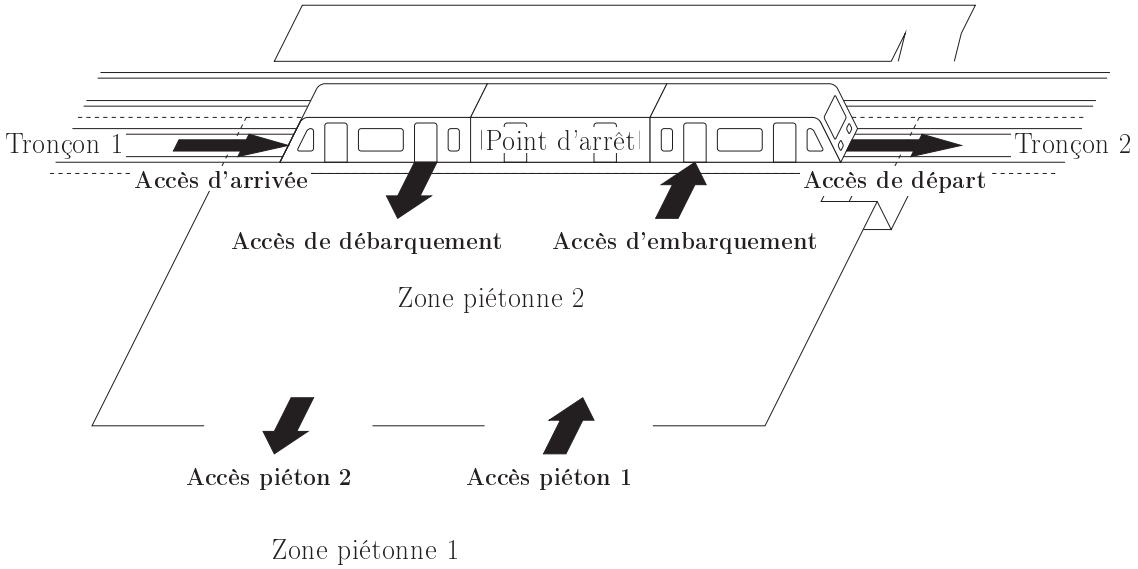


FIGURE 3.1 – Exemple de zones et d'accès du réseau de transport public.

3.1.2 Définition des flux de voyageurs

La notion de flux de voyageurs joue un rôle central dans notre étude. Grâce aux concepts spatiaux introduits précédemment, cette notion peut être définie de manière formelle :

Flux de voyageurs : Un flux de voyageurs F est une variable associée à un accès qui, pour chaque pas de temps t, représente le nombre de voyageurs F^t transitant par cet accès pendant t. Un accès est associé à un flux seulement si des mesures fiables de ce flux sont recueillies. Par analogie avec le type d'accès auquel il est associé, un flux peut être piéton, d'embarquement, de débarquement, de départ ou d'arrivée.

Les relations spatiales entre les flux de voyageurs constituent le cœur de notre approche de modélisation. À partir de la définition que nous venons de donner, il est possible de décrire précisément ces relations :

Chemin (spatial) : Au sens spatial du terme, un chemin est une suite de zones distinctes $(Z_1, \ldots, Z_n)$ $(n \geq 2)$ telle que, pour tout $i \in \{1, \ldots, n-1\}$, il existe un accès de Z_i à Z_{i+1}, noté (Z_i, Z_{i+1}).

Flux de voyageurs amont-aval : Un flux de voyageurs F_1 est en amont d'un autre flux F_2 (et F_2 est en aval de F_1) s'il existe un chemin $(Z_1, \ldots, Z_n)$ $(n \geq 3)$ tel que F_1 est associé à (Z_1, Z_2) et F_2 est associé à (Z_{n-1}, Z_n). À noter qu'une relation amont-aval peut exister selon plusieurs chemins différents.

Flux de voyageurs adjacents : Un flux de voyageurs F_1 est amont-adjacent à un autre flux F_2 (et F_2 est aval-adjacent à F_1) si F_1 est en amont de F_2 selon un chemin $(Z_1, \ldots, Z_n)$ satisfaisant l'une des conditions suivantes :

— $n = 3$, autrement dit la zone de destination de F_1 est la même que la zone d'origine de F_2 ;

— si $n > 3$, alors pour tout $i \in \{2, \ldots, n-2\}$, (Z_i, Z_{i+1}) n'est pas associé à un flux.

Afin de mieux comprendre ces relations spatiales, nous nous intéressons à nouveau à l'exemple de la figure 3.1. Nous supposons tout d'abord que chacun des accès présentés dans cette figure est associé à un flux de voyageurs. En prenant comme exemple le flux de départ, ce dernier est situé en aval des flux d'arrivée, d'embarquement et du flux piéton 1. Il n'est toutefois adjacent qu'aux flux d'arrivée et d'embarquement. Si nous supposons à présent que l'accès d'embarquement n'est pas associé à un flux, alors le flux de départ est adjacent au flux d'arrivée et au flux piéton 1.

3.2 Description des données

Notre étude se focalise sur le réseau ferré de la RATP, à travers l'exploitation de trois sources de données différentes : les validations des titres de transport, les comptages par pesée des voyageurs à bord des trains et l'offre de transport [1]. Ces données ont l'avantage d'être collectées à grande échelle et de manière continue. Grâce aux validations et aux comptages par pesée, nous disposons de mesures de flux recueillies à la fois sur des accès piétons et sur des accès embarqués. Malheureusement, les systèmes de collecte actuels ne permettent pas d'exploiter ces données en temps réel, ce qui représente un inconvénient majeur dans la démarche de prédiction. Notre approche de modélisation ne peut donc être testée que de manière prospective (en simulant le temps réel sur un jeu de données expérimental), sans possibilité d'industrialisation immédiate.

3.2.1 Validations

Les voyageurs qui utilisent les transports publics doivent valider leur titre de transport sur des bornes prévues à cette effet. Sur le réseau ferré, ces bornes sont

1. Utilisées uniquement dans la courte expérimentation de la sous-section 3.4.2, les données issues des comptages manuels (par ailleurs faciles à comprendre) ne sont pas explicitées.

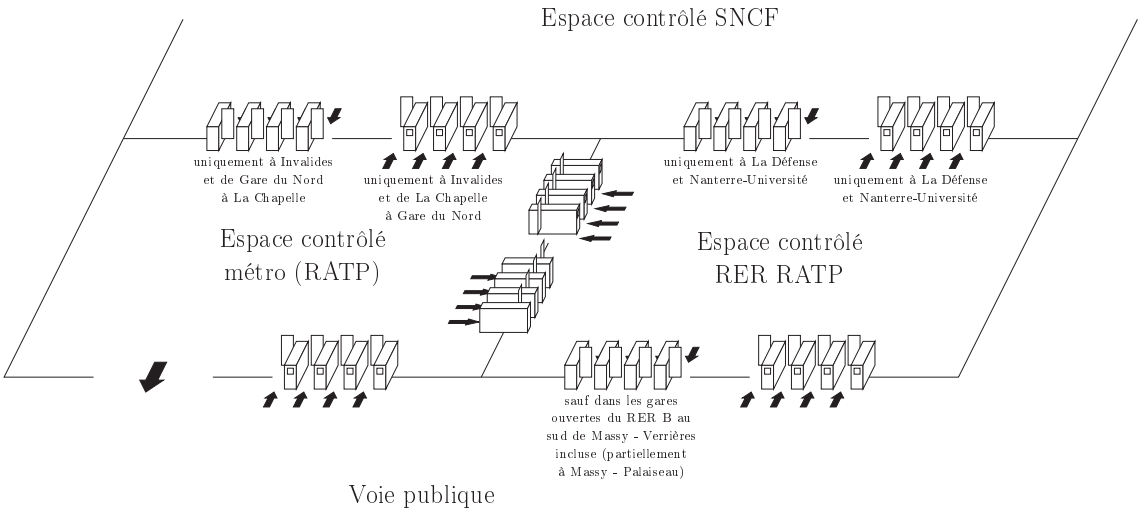


FIGURE 3.2 – Accès piétons du réseau ferré où la RATP recueille des données de validation.

positionnées sur des accès piétons et regroupées en lignes de contrôles délimitant des espaces contrôlés :

Espace contrôlé : Un espace contrôlé est un ensemble d'une ou plusieurs zones piétonnes contiguës auxquelles seuls les voyageurs munis d'un titre de transport valide ont le droit d'accéder. Il est contrôlé par un opérateur et associé à un mode de transport (dans le cas du métro, l'opérateur est forcément la RATP). Les zones piétonnes qui n'appartiennent pas à un espace contrôlé constituent la voie publique.

Chaque ligne de contrôle est rattachée à un espace contrôlé, à une recette (c'est-à-dire un lieu de vente ou d'information) et est caractérisée par un numéro. Les bornes qui la composent sont également numérotées. Dans la majorité des cas, les lignes de contrôle ferment totalement les accès de manière à contraindre les voyageurs à la validation. Les gares dites « ouvertes » du RER B, situées au sud de Massy - Verrières incluse, constituent toutefois une exception à cette règle (bien que certains accès de la gare de Massy - Palaiseau soient fermés).

Sur le réseau ferré, la RATP recueille des validations aux accès piétons suivants (voir la figure 3.2) :

— les accès de la voie publique aux espaces contrôlés par la RATP ;
— les accès des espaces contrôlés du RER RATP à la voie publique, sauf dans les gares ouvertes du RER B (partiellement à Massy - Palaiseau) ;
— les accès des espaces contrôlés du métro aux espaces contrôlés du RER RATP, et inversement ;

— les accès des espaces contrôlés par la Société Nationale des Chemins de Fer Français (SNCF) aux espaces contrôlés par la RATP, et inversement, uniquement dans les gares et stations Invalides, La Défense et Nanterre-Université, ainsi qu'entre Gare du Nord et La Chapelle.

Les validations recueillies sont stockées dans les bases de données de la RATP. Si une validation est opérée avec un titre de transport sur support télébillettique, elle remonte dans le Système d'Information Central (SIC) ferré. Si le titre est sur support magnétique, elle remonte dans le Centre de Traitement des Contrôles (CTK), ou dans le Système d'Acquisition des Contrôles (SAK) si la borne concernée est rattachée à un espace contrôlé desservi par la ligne 14 du métro (à l'exception de la station Châtelet), à l'espace contrôlé de la gare de Val d'Europe ou à la recette A de la gare de Noisy - Champs. Les validations du SIC ferré et du SAK (75 à 80 % des validations) sont horodatées à la seconde près. En revanche, celles du CTK sont agrégées par tranche de 10 minutes.

Il arrive fréquemment que des voyageurs franchissent des lignes de contrôle sans effectuer de validation, soit parce qu'ils ne possèdent pas de titre de transport valide (fraude), soit parce qu'ils n'y sont pas contraints physiquement (par exemple, si les portiques de la ligne de contrôle sont ouverts). Une enquête réalisée tous les deux ans pendant des jours ouvrés hors vacances scolaires permet d'obtenir une estimation du taux de non-validation par espace contrôlé. Au global, ce taux se situe généralement entre 2 et 3 %.

3.2.2 Comptages par pesée

Sur le réseau ferré, la RATP dispose de plusieurs parcs de matériels roulants (MF01, MP05, MI09, etc.), dont les éléments (ou rames) sont utilisés pour constituer des trains. Au sein d'un même parc de matériels, chaque train est identifié par un numéro unique. Il ne peut être constitué que d'un seul élément sur le réseau métro, contre un ou plusieurs éléments sur le réseau RER.

Les matériels les plus récents sont équipés de capteurs de pression de suspension pneumatique qui peuvent être utilisés pour mesurer la masse embarquée et en déduire le nombre de voyageurs à bord des trains. À l'heure actuelle, ce type de système est déjà exploité sur les matériels MF01, qui équipent les lignes 2, 5 et 9 du métro. Son déploiement est en cours sur les matériels MI09, qui équipent le RER A. Il est également à l'étude sur les matériels MP05, qui équipent les lignes 1 et 14, ainsi que sur les futures matériels MP14 et MF19, qui équiperont une dizaine de

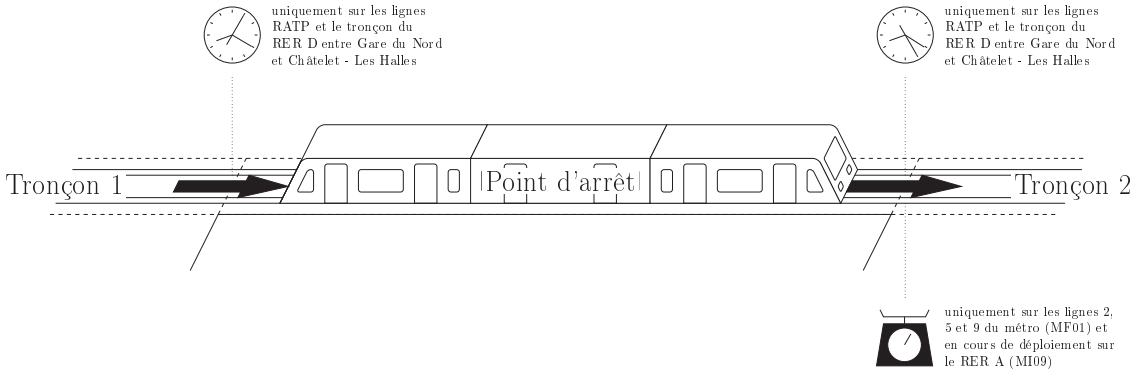


FIGURE 3.3 – Accès embarqués du réseau ferré où la RATP recueille des données de comptage par pesée et d'offre de transport réelle.

lignes de métro à partir de 2019 et 2023 (respectivement).

Les comptages par pesée des MF01 se déroulent de la manière suivante. Lorsqu'un train est mis en circulation pour la première fois de la journée, une tare de chaque voiture est réalisée afin de déterminer sa masse à vide. Lors du départ d'un point d'arrêt, une nouvelle série de pesées est initialisée. À chaque fois que le train franchit un seuil de vitesse (en accélérant ou décélérant), une pesée pneumatique de chaque voiture est effectuée. Lors de l'arrivée au point d'arrêt suivant, la masse embarquée M^v_{emb} de la voiture v est déduite de la moyenne des mesures réalisées tout au long du tronçon en tenant compte de la masse à vide M^v_{vide} de v :

$$M^v_{\text{emb}} = \frac{\sum_{i=1}^{N_{\text{pes}}} M^{v,i}_{\text{tot}}}{N_{\text{pes}}} - M^v_{\text{vide}}\,, \tag{3.1}$$

où $M^{v,i}_{\text{tot}}$ est la masse totale de v mesurée lors de la i-ème pesée et N_{pes} le nombre de pesées effectuées sur le tronçon. À partir de la masse moyenne d'un voyageur M_{voy}, il est alors possible d'estimer le nombre de voyageurs à bord de v :

$$N^v_{\text{voy}} = \frac{M^v_{\text{emb}}}{M_{\text{voy}}}\,. \tag{3.2}$$

Les masses embarquées enregistrées par voiture sont transmises au Centre de Réception des Informations Train (CRIT), où elles sont mises à disposition du Système en Ligne d'Enregistrement du Trafic Voyageurs (SYLEVE). Elles sont alors converties en comptages en fixant le poids moyen d'un voyageur par défaut à 70 kg. Sur le plan spatial, chaque comptage est attribué à l'accès de départ du tronçon sur lequel il est réalisé (voir la figure 3.3). En général, plus la densité de voyageurs à bord est élevée, plus ce comptage est précis.

3.2.3 Offre de transport

Pendant ses horaires de service, un train effectue un certain nombre de courses, c'est-à-dire de trajets suivant un itinéraire précis d'une ligne dans un sens donné. Sur le réseau métro, ce train se voit attribuer un numéro de mission à deux chiffres qu'il conserve durant plusieurs courses successives. Sur le réseau RER, ce numéro est composé de quatre lettres associées à l'itinéraire emprunté et aux points d'arrêt desservis, ainsi que de deux chiffres permettant d'identifier chaque course de manière unique.

Les données d'offre de transport permettent de connaître les points d'arrêt desservis lors des différentes courses, ainsi que les horaires de ces dessertes. Trois niveaux d'information se distinguent :

— l'offre de transport théorique, planifiée en dehors des horaires de service, en principe plusieurs mois à l'avance ;
— l'offre de transport estimée, prévue plusieurs minutes à l'avance en fonction des conditions du réseau et utilisée pour informer les voyageurs en temps réel ;
— l'offre de transport réelle, enregistrée a posteriori et correspondant à l'offre effectivement réalisée.

L'offre de transport réelle étant naturellement plus fiable que l'offre théorique et l'offre estimée, nous nous focalisons exclusivement sur ce type d'information dans la suite de notre étude.

Le suivi de la circulation des trains s'effectue par l'analyse de l'occupation et de la libération des circuits de voie. Ce système permet de détecter, par le biais d'un circuit électrique, si un train est présent ou non sur une section donnée de la voie ferrée. Lorsque le train entre sur la section, on dit qu'il procède à une occupation du circuit de voie. S'il quitte totalement la section, on dit qu'il procède à une libération du circuit de voie. L'enregistrement de ces deux types d'événements partout sur le réseau permet de déterminer les horaires d'arrivée et de départ à chaque point d'arrêt. En effet, l'arrivée correspond à l'instant où le train libère le circuit de voie qui précède le point d'arrêt, tandis que le départ correspond à l'instant où le train ouvre le circuit de voie qui suit le point d'arrêt [2].

Les données d'occupation et de libération des circuits de voie sont collectées sur l'ensemble des lignes de métro et de RER exploitées par la RATP, ainsi que

2. En réalité, il existe un décalage de quelques secondes entre l'instant de libération (respectivement d'occupation) du circuit de voie et le moment exact où le train s'arrête (respectivement redémarre). Ce léger décalage varie en fonction du point d'arrêt et de la vitesse du train. Par souci de simplification, nous n'en tenons pas compte dans la suite de ces travaux.

sur le tronçon du RER D entre Gare du Nord et Châtelet - Les Halles (qui utilise les infrastructures de la RATP). Ces données sont transférées et stockées dans le Concentrateur Diffuseur (CONDIF), qui alimente ensuite diverses applications de suivi de l'exploitation. Il est alors possible de connaître l'offre de transport réelle à chaque accès embarqué, tel qu'illustré dans la figure 3.3.

3.3 Traitement des données

À travers la description précise des éléments qui composent le réseau de transport, nous avons construit un socle commun permettant de réunir diverses sources de données au sein du même modèle. Si les données présentées diffèrent par leur nature et leur méthode de collecte, toutes s'intègrent dans cette représentation normalisée des espaces. Chaque donnée peut en effet être associée à un accès et, par extension, à un flux de voyageurs. Cet ancrage spatial lui permet d'être mise en relation avec les autres données collectées partout ailleurs sur le réseau.

Inspiré des diagrammes de classes UML (Unified Modeling Language), le modèle conceptuel des données proposé dans la figure 3.4 rassemble les différents concepts spatiaux et relatifs aux données présentés depuis le début de ce chapitre. Afin de faciliter sa compréhension, les noms des entités et des attributs sont exprimés de manière intelligible, sans tenir compte de la nomenclature des référentiels de la RATP. En prenant comme exemple les entités « Borne » et « Validation », les associations se lisent de la manière suivante : une validation est effectuée sur une et une seule borne (cardinalité 1), tandis qu'une borne peut recevoir plusieurs validations (cardinalité 0..*). Ce modèle conceptuel témoigne de notre capacité à réunir les données de validation, de comptage par pesée et d'offre de transport réelle au sein d'un formalisme commun. Bien entendu, son utilisation ne se limite pas à notre seule étude et il est tout à fait possible de l'enrichir en intégrant de nouvelles sources de données.

Avant d'utiliser les données à des fins expérimentales, un certain nombre de traitements doivent être réalisés. Tout d'abord, les comptages par pesée sont réajustés en modifiant la masse moyenne d'un voyageur en fonction de la ligne et de la période de la journée. Ces nouvelles masses (qui demeurent proches du poids par défaut de 70 kg) sont déterminées en comparant les données issues de SYLEVE avec des comptages manuels opérés plusieurs fois par mois au départ de certains points d'arrêt. Concernant les données d'offre de transport, les courses qui ne prennent pas de voyageurs (ou haut-le-pied) sont écartées. Ces dernières sont identifiables sur le réseau métro par le numéro de mission 00, et sur le réseau RER par un numéro de

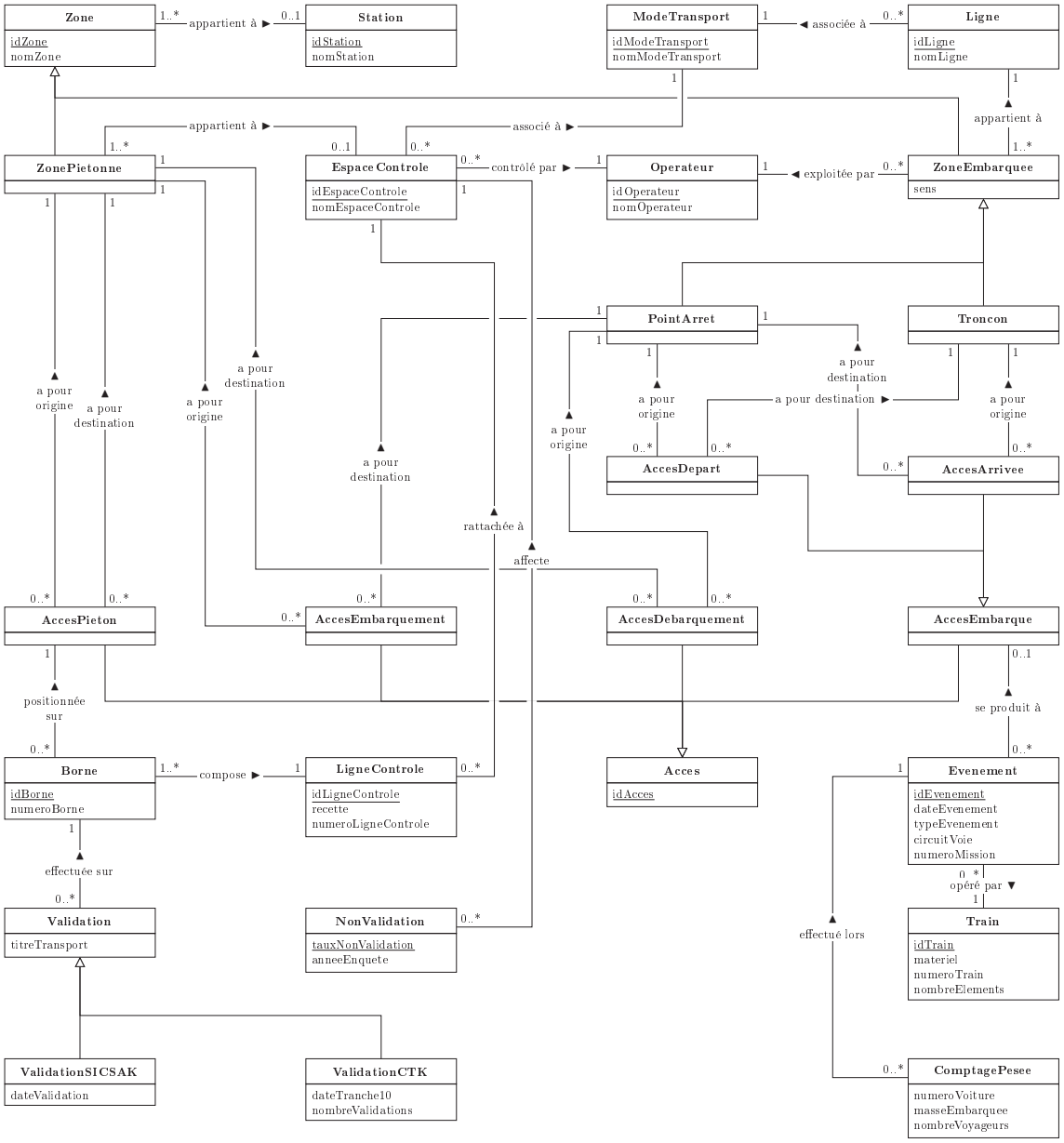


FIGURE 3.4 – Modèle conceptuel des données de validation, de comptage par pesée et d'offre de transport réelle.

mission commençant par W.

La seconde partie des traitements consiste à transformer les données de validation et de comptage par pesée en mesures de flux. Pour un flux piéton mesuré par le biais des validations, nous devons prendre en compte le fait que les validations issues du CTK sont agrégées par tranche de 10 minutes. Nous pouvons alors supposer qu'à l'intérieur d'une même tranche de 10 minutes t_{10}, la distribution de ces validations est la même que celle des validations du SIC et du SAK. Autrement dit, pour un

pas de temps t de durée D (en minutes) inclus dans t_{10}, le nombre de validations du CTK effectuées pendant t est estimé de la manière suivante :

$$N^t_{\text{CTK}} = \begin{cases} \frac{N^t_{\text{SIC}}}{N^{t_{10}}_{\text{SIC}}} N^{t_{10}}_{\text{CTK}} & \text{si } N^{t_{10}}_{\text{SIC}} > 0 \\ \frac{D}{10} N^{t_{10}}_{\text{CTK}} & \text{si } N^{t_{10}}_{\text{SIC}} = 0 \end{cases}, \tag{3.3}$$

où N^t_{SIC} et $N^{t_{10}}_{\text{SIC}}$ sont les nombres de validations du SIC et du SAK enregistrées respectivement pendant t et t_{10}, et $N^{t_{10}}_{\text{CTK}}$ est le nombre de validations du CTK enregistrées pendant t_{10}.

Afin de disposer d'un comptage exhaustif des voyageurs, il convient enfin d'appliquer le taux de non-validation T_{nval} relatif à l'espace contrôlé considéré. La mesure du flux pour le pas de temps t s'obtient donc ainsi :

$$F^t_{\text{val}} = \frac{N^t_{\text{SIC}} + N^t_{\text{CTK}}}{1 - T_{\text{nval}}}. \tag{3.4}$$

En ce qui concerne les flux de départ mesurés par le biais de comptages par pesée, la mesure pour le pas de temps t est calculée par une simple somme des comptages réalisés à bord des trains dont le départ a lieu pendant t :

$$F^t_{\text{pes}} = \begin{cases} \sum_{j=1}^{N^t_{\text{dép}}} N^{t,j}_{\text{voy}} & \text{si } N^t_{\text{dép}} > 0 \\ 0 & \text{si } N^t_{\text{dép}} = 0 \end{cases}, \tag{3.5}$$

où $N^t_{\text{dép}}$ est le nombre de départs de trains pendant t et $N^{t,j}_{\text{voy}}$ le nombre de voyageurs à bord du j-ème train.

En raison de défaillances techniques ou d'absences de systèmes de collecte, il arrive que des flux de voyageurs ne soient que partiellement observés. Dans le cas des validations, les lignes de contrôle et les systèmes intervenant dans la remontée des données sont parfois sujets à des dysfonctionnements. Si les pertes de données qui en résultent représentent, en temps normal, moins de 5 % des validations totales du réseau ferré, elles sont souvent concentrées sur un petit nombre de flux piétons et peuvent donc avoir un impact significatif à l'échelle locale. Dans le cas des flux de départ, les données manquantes proviennent de dysfonctionnements des systèmes de comptage par pesée, mais aussi de l'absence d'équipement de certains trains. Dans l'expérimentation du chapitre 4, ce dernier point concerne uniquement la ligne 9 qui, au moment du recueil des données, n'est que partiellement équipée de matériels MF01.

3.4 Construction du modèle

3.4.1 Prise en compte de l'information spatiale

Dans un réseau de transport public, il existe une relation de causalité intuitive entre les flux de voyageurs amont et aval. Lorsqu'un phénomène impacte un flux à un endroit du réseau, ses effets se propagent naturellement dans le sens de circulation des voyageurs. Par exemple, si une forte affluence est observée en entrée d'un espace contrôlé, il est probable que celle-ci se répercute sur le nombre de voyageurs montant à bord des trains à l'intérieur de cet espace. À l'inverse, si la circulation des trains est interrompue à partir d'un point d'arrêt, le nombre de voyageurs sortant des espaces contrôlés situés en aval est mécaniquement plus faible.

La structure graphique des réseaux bayésiens constitue un outil naturel pour la représentation de ces relations de causalité. En supposant que chaque flux de voyageurs dépend de ses flux amont-adjacents, tous les arcs qui composent cette structure sont orientés d'amont en aval. Ce principe de construction permet à l'information de circuler de proche en proche, dans le sens de circulation des voyageurs. La structure dérive alors directement de la topologie du réseau de transport, ce qui facilite sa construction.

3.4.2 Application sur une gare de RER

Afin de tester le principe de construction énoncé précédemment, nous réalisons une première expérimentation sur la gare de Nanterre-Préfecture, desservie par le RER A, dont un plan simplifié est fourni dans la figure 3.5a. Dans le cadre de cette expérimentation, six flux de voyageurs sont mesurés :

— le flux piéton passant de l'entrée Boulevard de Pesaro au quai direction Cergy-le-Haut / Poissy (F_1), mesuré par le biais des validations ;
— le flux piéton passant de la partie publique à la partie contrôlée de la salle des billets (F_2), mesuré par le biais des validations ;
— le flux piéton passant de la salle des billets au quai direction Saint-Germain-en-Laye (F_3), mesuré par le biais de comptages manuels ;
— le flux piéton passant de l'entrée Préfecture au quai direction Saint-Germain-en-Laye (F_4), mesuré par le biais des validations ;
— le flux piéton passant de l'entrée Préfecture au quai direction Paris (F_5), mesuré par le biais des validations ;
— le flux d'embarquement à bord des trains direction Paris en provenance de

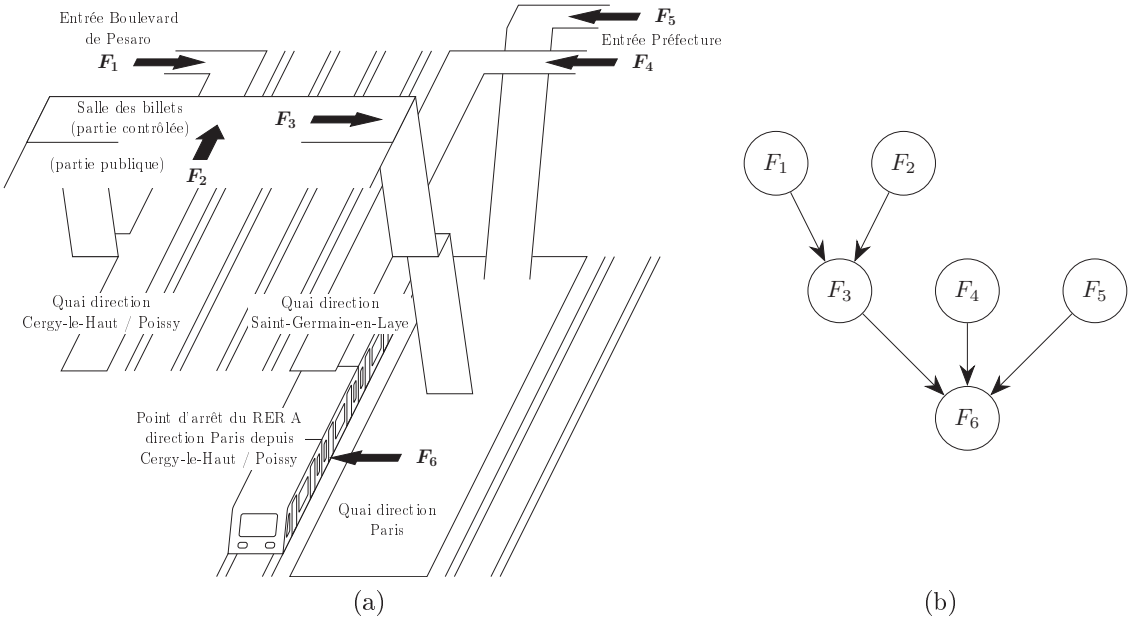


FIGURE 3.5 – (a) Plan simplifié de la gare de Nanterre-Préfecture avec les flux de voyageurs mesurés. (b) Réseau bayésien représentant les relations de causalité entre les flux adjacents.

Cergy-le-Haut / Poissy (F_6), mesuré par le biais de comptages manuels.

D'après la définition des relations d'adjacence donnée dans la sous-section 3.1.2, F_3 est aval-adjacent à F_1 et F_2, tandis que F_6 est aval-adjacent à F_3, F_4 et F_5. Le réseau bayésien qui en résulte est présenté dans la figure 3.5b.

Les distributions conditionnelles locales du réseau bayésien sont décrites par des modèles linéaires gaussiens. Dans le cadre de cette expérimentation, nous nous intéressons exclusivement aux distributions $p(F_3|F_1, F_2)$ et $p(F_6|F_3, F_4, F_5)$. Plus précisément, nous cherchons à prédire les valeurs de F_3 compte tenu de celles de F_1 et F_2, ainsi que les valeurs de F_6 compte tenu de celles de F_3, F_4 et F_5. Ces deux cas sont traités de manière indépendante à l'aide de deux ensembles de données recueillies durant des jours ouvrés hors vacances scolaires. Pour la prédiction de F_3, 264 observations complètes de 10 minutes sont collectées :

— le jeudi 5 décembre 2013, de 5 h à 1 h le jour suivant ;

— les mardi 11 et jeudi 27 mars 2014, de 7 h à 13 h et de 14 h à 20 h.

Concernant la prédiction de F_6, 143 observations complètes de 10 minutes sont collectées les 11 et 27 mars durant les mêmes horaires [3].

Les distances entre les flux sont telles qu'au cours d'une même tranche de 10

3. L'observation du 27 mars entre 18 h 10 et 18 h 20 est écartée en raison d'un défaut de fiabilité des comptages.

minutes, les voyageurs ont généralement le temps de transiter par plusieurs flux différents. Par conséquent, nous pouvons considérer que les relations de causalité entre les flux adjacents interviennent dans la même tranche, ce qui justifie l'utilisation d'un réseau bayésien statique.

Étant donné le nombre limité d'observations de chaque ensemble de données, les performances de prédiction sont évaluées par validation croisée à 10 plis (Kohavi, 1995). Cette méthode consiste à diviser aléatoirement l'ensemble de données $\mathcal{X}$ en 10 sous-ensembles $\mathcal{X}_1, \ldots, \mathcal{X}_{10}$ (les plis) de taille à peu près équivalente. Le nombre d'observations de $\mathcal{X}$ n'étant pas forcément un multiple de 10, certains sous-ensembles peuvent contenir une observation de plus que les autres. Pour chaque $k \in \{1, \ldots, 10\}$, les paramètres du modèle linéaire gaussien sont estimés à partir de $\mathcal{X} \backslash \mathcal{X}_k$, selon la méthode du maximum de vraisemblance décrite dans la sous-section 2.2.2. L'erreur de prédiction e_k est ensuite calculée en testant le modèle sur $\mathcal{X}_k$. Les 10 erreurs obtenues sont alors moyennées pour estimer l'erreur de prédiction finale :

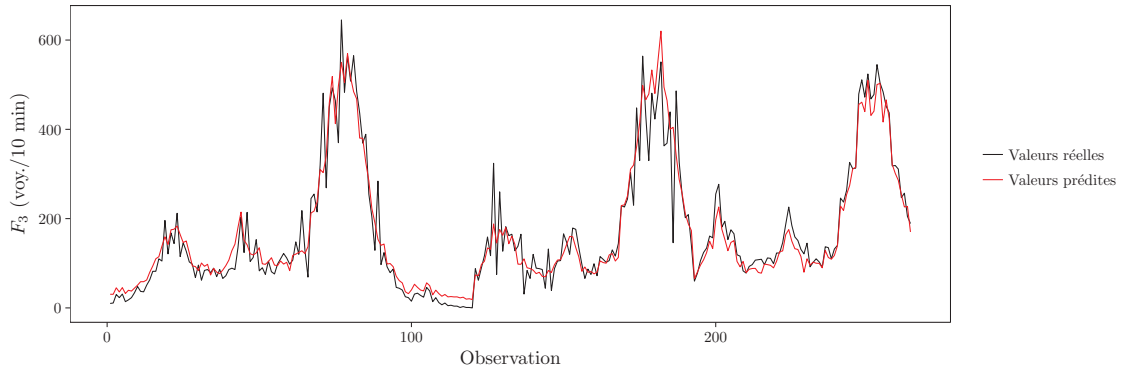
$$e_{\text{vc}} = \frac{\sum_{k=1}^{10} e_k}{10} \,. \tag{3.6}$$

L'erreur de prédiction que nous utilisons est l'erreur absolue moyenne en pourcentage pondérée (WMAPE) :

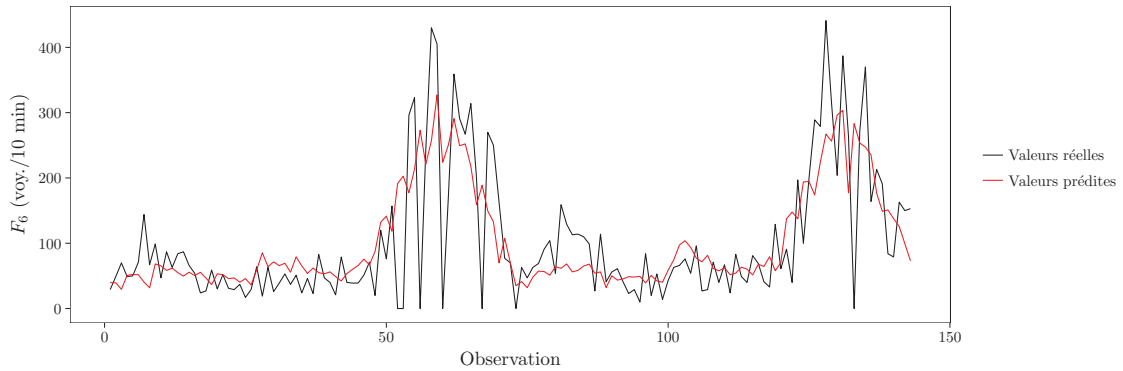
$$\text{WMAPE}(x, \hat{x}) = \frac{\sum_{m=1}^{N} |x^m - \hat{x}^m|}{\sum_{m=1}^{N} x^m} \,, \tag{3.7}$$

où $x = \{x^1, \ldots, x^N\}$ est l'ensemble des valeurs réelles et $\hat{x} = \{\hat{x}^1, \ldots, \hat{x}^N\}$ l'ensemble des valeurs prédites. La WMAPE est une variante de l'erreur absolue moyenne en pourcentage (MAPE), qui possède la même facilité d'interprétation que celle-ci. La différence est qu'elle pondère les écarts relatifs absolus par les valeurs réelles, ce qui la rend moins sensible aux écarts qui concernent les valeurs les plus faibles. Dans notre étude, elle permet donc de favoriser les modèles qui prédisent le mieux les valeurs élevées des flux de voyageurs.

Comme le montre la figure 3.6a, les résultats de prédiction de F_3 se révèlent particulièrement encourageants, avec une WMAPE estimée à 16.3 %. Nous pouvons voir que la courbe des valeurs prédites parvient bien à épouser celle des valeurs réelles. En revanche, les performances de prédiction diminuent significativement lorsque nous considérons F_6, avec une WMAPE estimée à 48.2 %. En effet, comme le montre la figure 3.6b, ce flux de voyageurs est sujet à de larges fluctuations liées à la circulation des trains. Si aucun train en provenance de Cergy-le-Haut / Poissy ne dessert la gare, le flux est logiquement interrompu, ce qui explique les creux observés par



(a)



(b)

FIGURE 3.6 – Valeurs réelles et prédites (lors de la validation croisée) des flux de voyageurs de la gare de Nanterre-Préfecture : (a) F_3 ; (b) F_6.

intermittence. En utilisant uniquement l'information issue des flux amont-adjacents, le modèle ne parvient pas à épouser ces fluctuations. Ce résultat illustre le caractère spécifique des réseaux de transport public, dont les flux de voyageurs sont étroitement liés à l'offre de transport.

3.4.3 Extension au facteur temporel

Dans l'expérimentation menée sur la gare de Nanterre-Préfecture, les flux de voyageurs sont prédits à partir de leur voisinage spatial observé au même pas de temps. Or notre objectif est de réaliser des prédictions dans le futur. Pour ce faire, nous décidons d'étendre la modélisation au formalisme des réseaux bayésiens dynamiques. Désormais, un nœud ne représente plus un flux de voyageurs, mais une instanciation de ce flux à un pas de temps donné.

En nous référant aux travaux de Sun *et al.* (2006), le voisinage spatial peut être étendu au voisinage spatio-temporel. Nous supposons qu'il existe une relation de

dépendance causale entre chaque flux de voyageurs au pas de temps t et ses flux amont-adjacents à $t-1, \ldots, t-r_1$, où r_1 est un paramètre défini expérimentalement. Les arcs concernés sont donc orientés en avant dans l'espace et dans le temps. Toujours selon Sun *et al.* (2006), les valeurs historiques d'un flux nous renseignent sur sa tendance actuelle. Nous définissons donc un autre paramètre r_2 tel que chaque flux à t dépend de ses propres valeurs à $t-1, \ldots, t-r_2$. Au final, l'ordre du réseau bayésien dynamique est défini selon le paramètre r_1 ou r_2 maximal :

$$r = \max(r_1, r_2)\,. \tag{3.8}$$

Contrairement à l'expérimentation précédente où les données sont agrégées par tranche de 10 minutes, il convient de choisir ici une résolution temporelle suffisamment fine pour que les relations de causalité spatio-temporelles entre les flux soient cohérentes (de l'ordre de 1 à 2 minutes dans le cadre de notre étude).

3.4.4 Intégration de l'offre de transport

L'expérimentation précédente a mis en évidence le lien étroit entre les flux de voyageurs et la circulation des trains. L'intégration de l'offre de transport peut donc contribuer grandement à l'amélioration des performances de prédiction. Lorsque des voyageurs souhaitent emprunter une ligne au départ d'un point d'arrêt, ils rejoignent le quai correspondant et attendent l'arrivée d'un train. Quand le train arrive à quai, l'embarquement est possible jusqu'au départ de celui-ci. Passé cet instant, le quai se remplit à nouveau de voyageurs jusqu'à l'arrivée du train suivant, et ainsi de suite. Plus l'intervalle de temps entre deux départs successifs est important, plus les voyageurs s'accumulent sur le quai avant d'embarquer à bord du second train. Le nombre de voyageurs au départ de ce train dépend donc directement de cet intervalle.

Ces observations nous conduisent à introduire la notion d'« intervalle de départ », ainsi que les relations spatiales qui lui sont associées :

Intervalle de départ : Soient $h_1, \ldots, h_{n_H}$ les horaires de départ des trains enregistrés dans l'ordre chronologique sur un accès de départ, et H^t l'ensemble des horaires enregistrés au pas de temps t. L'intervalle de départ I associé à cet accès est une variable qui, pour chaque pas de temps t, représente la somme des intervalles de temps [4], exprimée en secondes, entre les couples

4. La notion d'intervalle de temps est assimilée ici à une durée.

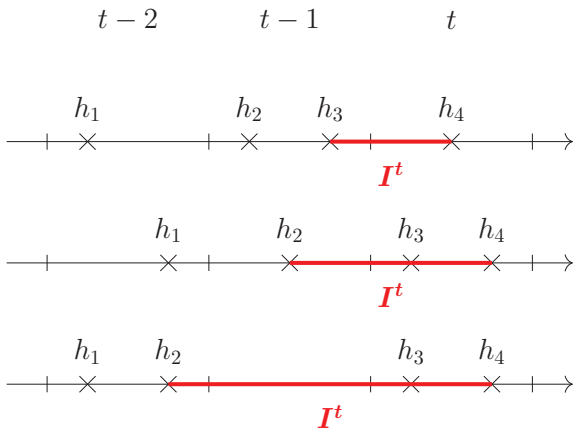


FIGURE 3.7 – Exemples d'horaires de départ et d'intervalles de départ au pas de temps t.

d'horaires successifs (h_i, h_{i+1}) tels que $h_{i+1} \in H^t$:

$$I^t = \sum_{h_{i+1} \in H^t} h_{i+1} - h_i = \begin{cases} \max H^t - \max \bigcup_{k<t} H^k & \text{si } H^t \neq \emptyset \\ 0 & \text{si } H^t = \emptyset \end{cases} . \tag{3.9}$$

Si $H^t \neq \emptyset$, I^t représente l'intervalle de temps entre le dernier horaire enregistré avant t et le dernier horaire enregistré pendant t, autrement dit l'intervalle durant lequel les voyageurs peuvent partir au plus tôt à t. La figure 3.7 permet de mieux visualiser le lien entre les horaires de départ et I^t. À noter qu'un accès de départ est associé à un intervalle de départ seulement si des mesures fiables de cet intervalle sont recueillies.

Intervalle de départ et flux de départ associés : Un intervalle de départ I est associé à un flux de départ F si I et F sont associés au même accès de départ.

Intervalle de départ et flux de voyageurs adjacents : Un intervalle de départ I est amont-adjacent à un flux de voyageurs F (ou F est aval-adjacent à I) s'il existe un chemin $(Z_1, \ldots, Z_n)$ $(n \geq 3)$ tel que :

— I est associé à (Z_1, Z_2) et F est associé à (Z_{n-1}, Z_n) ;

— pour tout $i \in \{1, \ldots, n-2\}$, (Z_i, Z_{i+1}) n'est pas associé à un flux de voyageurs.

Intervalles de départ adjacents : Un intervalle de départ I_1 est amont-adjacent à un autre intervalle de départ I_2 (ou I_2 est aval-adjacent à I_1) s'il existe un chemin $(Z_1, \ldots, Z_n)$ $(n \geq 3)$ tel que :

— $Z_1, \ldots, Z_n$ sont des zones embarquées appartenant à la même ligne et au même sens de circulation ;
— I_1 est associé à (Z_1, Z_2) et I_2 est associé à (Z_{n-1}, Z_n) ;
— pour tout $i \in \{2, \ldots, n-2\}$ (si $n > 3$), (Z_i, Z_{i+1}) n'est pas associé à un intervalle de départ.

Les intervalles de départ constituent un nouvel ensemble de variables du réseau bayésien dynamique. Les observations émises précédemment nous conduisent à supposer que ces intervalles possèdent une relation de causalité avec leur flux de départ associé. Plus précisément, chaque flux de départ au pas de temps t dépend de son intervalle associé au même pas de temps. Nous supposons également qu'un flux de voyageurs à t dépend de ses intervalles de départ amont-adjacents à $t-1, \ldots, t-r_1$. Enfin, de manière analogue aux flux, chaque intervalle de départ à t dépend de ses intervalles amont-adjacents à $t-1, \ldots, t-r_1$ ainsi que de ses propres valeurs à $t-1, \ldots, t-r_2$.

La figure 3.8a fournit un exemple de flux de voyageurs et d'intervalles de départ mesurés sur le réseau ferré pour deux stations desservies successivement par la même ligne. Ces derniers sont représentatifs des types de flux et d'intervalles considérés dans l'expérimentation du chapitre 4. Dans cet exemple, le flux F_2 est aval-adjacent à F_1, F_3 est aval-adjacent à F_2 et F_5 est aval-adjacent à F_2 et F_4. L'intervalle I_1 est associé au même accès que F_2 et I_2 au même accès que F_5 (I_2 est aval-adjacent à I_1). La figure 3.8b permet de visualiser le réseau bayésien de transition $\mathcal{B}_{\rightarrow}$ du réseau bayésien dynamique associé à ces flux et à ces intervalles, en choisissant comme paramètres $r_1 = r_2 = 2$.

3.4.5 Recherche de la sous-structure optimale

En fonction des valeurs de r_1 et r_2, le nombre d'arcs du réseau bayésien dynamique peut être très élevé. Considérer l'ensemble de ces arcs engendre des temps de calcul importants et favorise les risques de surapprentissage. Afin de prévenir ce genre de situation, la structure peut être construite de manière indirecte en sélectionnant d'abord un ensemble d'arcs candidats à partir de r_1 et r_2, puis en cherchant parmi ces arcs une sous-structure optimale. La recherche de cette sous-structure est réalisée conjointement à l'apprentissage des paramètres en appliquant les algorithmes décrits dans le chapitre 2 à un jeu de données historiques.

En résumé, la construction du réseau bayésien dynamique s'articule en deux étapes successives. Dans un premier temps, les arcs candidats sont sélectionnés de

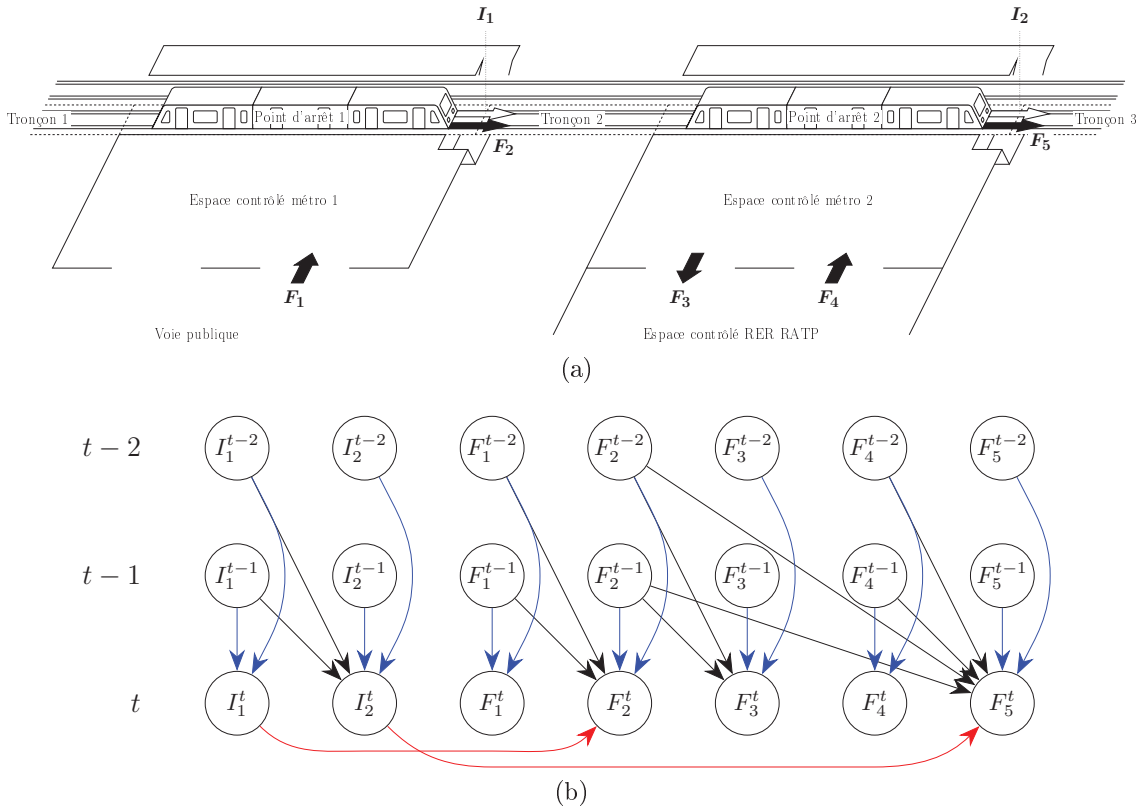


FIGURE 3.8 – (a) Exemple de flux de voyageurs (flèches noires) et d'intervalles de départ (flèches blanches) mesurés sur le réseau ferré. (b) Réseau bayésien de transition $\mathcal{B}_{\rightarrow}$ du réseau bayésien dynamique associé ($r_1 = r_2 = 2$). Les relations d'adjacence sont représentées par des arcs noirs, les relations entre la valeur courante et les valeurs historiques d'une même variable par des arcs bleus et les relations entre les flux de départ et leur intervalle de départ associé par des arcs rouges.

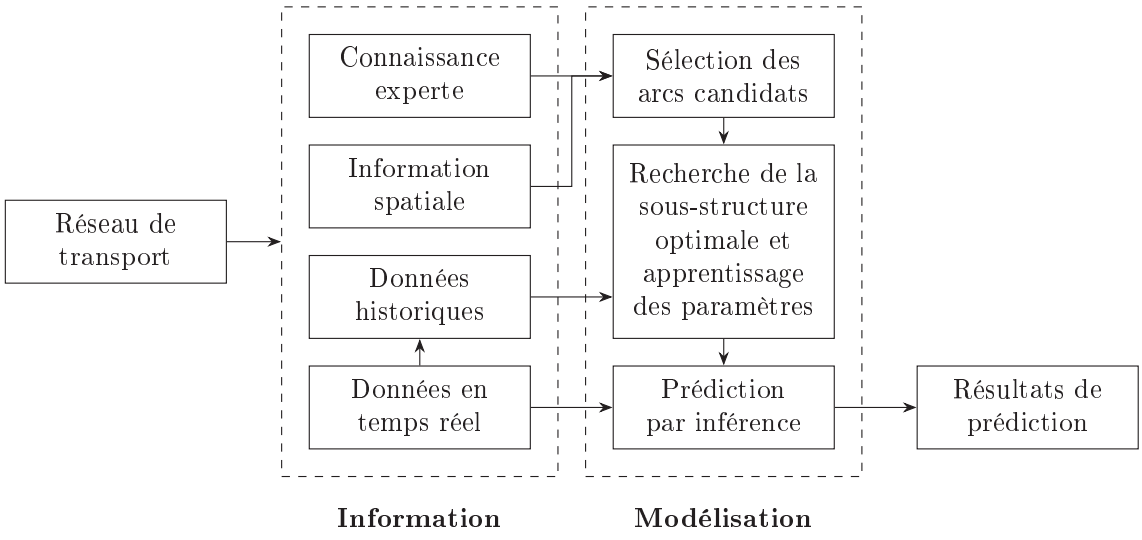


FIGURE 3.9 – Démarche de prévision à court terme des flux de voyageurs par les réseaux bayésiens dynamiques.

façon experte selon les relations spatiales entre les flux de voyageurs et les intervalles de départ. Dans un second temps, la sous-structure optimale et les paramètres sont déterminés par apprentissage à partir des données collectées sur le réseau de transport. Décrite dans la figure 3.9 (de manière analogue à la figure 1.1), cette démarche de modélisation repose ainsi sur des principes topologiques simples et intuitifs. Son applicabilité demeure indépendante de la configuration spatiale du réseau et des types de systèmes de collecte utilisés. Par conséquent et bien que nos travaux se concentrent exclusivement sur le réseau ferré, elle peut être facilement élargie à l'ensemble du réseau de transport multimodal.

Chapitre 4

Expérimentation à grande échelle

Dans le chapitre 3, nous avons développé une méthode de prévision à court terme des flux de voyageurs basée sur les réseaux bayésiens dynamiques. Afin de tester l'efficacité de cette méthode, nous réalisons une expérimentation sur l'ensemble de la ligne 2 du métro de Paris. Dans un premier temps, les distributions jointes locales du réseau bayésien dynamique sont décrites par de simples gaussiennes. Par la suite, ces distributions sont étendues aux modèles de mélanges gaussiens, en travaillant d'abord avec un nombre fixe de composantes (défini a priori), puis en cherchant à optimiser ce nombre à l'aide de l'algorithme EM de division-fusion. Les performances de notre approche sont finalement comparées à celles fournies par d'autres méthodes de prédiction, ceci afin de mieux évaluer la pertinence des principes de modélisation que nous avons adoptés.

4.1 Données d'entrée

4.1.1 Périmètre de l'expérimentation

Nous appliquons notre méthode de prévision à court terme des flux de voyageurs à l'ensemble de la ligne 2 du métro. Longue de 12.3 kilomètres, cette ligne suit un parcours semi-circulaire au nord de Paris entre les stations Nation et Porte Dauphine. Elle dessert 25 stations, dont les trois grands pôles multimodaux de Nation, Charles de Gaulle - Étoile et Gare du Nord (par La Chapelle). Sa fréquentation est estimée à environ 560000 voyageurs par jour ouvré. Au moment où nous réalisons cette expérimentation, il s'agit de l'une des deux seules lignes (avec la ligne 5) à être entièrement équipée de systèmes de comptage par pesée [1].

1. En 2015, lors du recueil des données utilisées dans cette expérimentation, la ligne 9 n'est encore que partiellement équipée de matériels MF01.

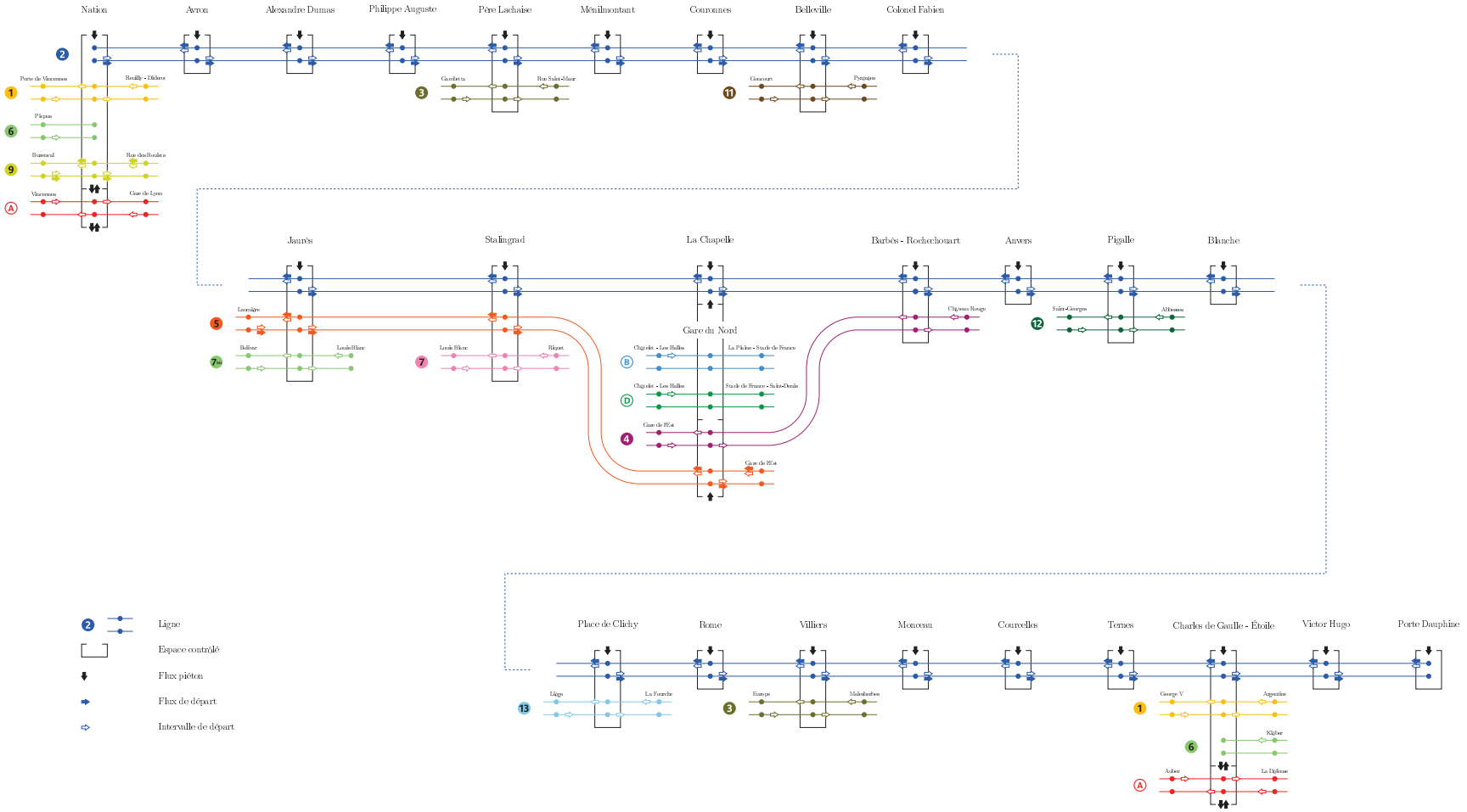


FIGURE 4.1 – Plan schématique de la ligne 2 et des lignes en correspondance, avec les flux de voyageurs et les intervalles de départ mesurés.

Dans le cadre de cette expérimentation, nous considérons 95 flux de voyageurs relatifs à la ligne 2 ou spatialement proches de celle-ci. Ces flux se répartissent en :

— 35 flux piétons mesurés par le biais des validations : 30 flux entrant ou sortant des espaces contrôlés desservis par la ligne 2 et 5 flux entrant ou sortant des espaces contrôlés en correspondance ;

— 60 flux de départ mesurés par le biais de comptages par pesée : 48 flux de la ligne 2 et 12 flux des lignes 5 et 9 en correspondance.

À ces 95 flux s'ajoutent 114 intervalles de départ mesurés à partir de l'offre de transport : 48 intervalles de la ligne 2 et 66 intervalles des lignes de métro et de RER en correspondance. La figure 4.1 fournit un plan schématique de la ligne 2 et des lignes en correspondance, où sont représentés l'ensemble des flux de voyageurs et des intervalles de départ. Ce plan permet de mieux visualiser les relations spatiales entre ces flux et ces intervalles.

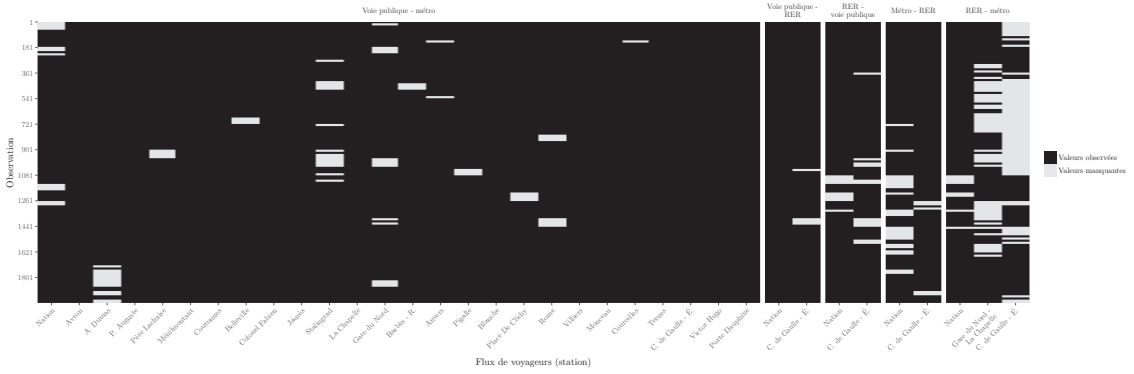
Afin de réaliser cette expérimentation, nous disposons d'un ensemble de données recueillies durant 33 jours ouvrés hors vacances scolaires, du 2 mars au 17 avril 2015 [2], entre 7 h 30 et 9 h 30 (durant la période de pointe du matin). Ces données sont agrégées par pas de temps de 2 minutes, totalisant ainsi 1980 observations.

4.1.2 Données manquantes

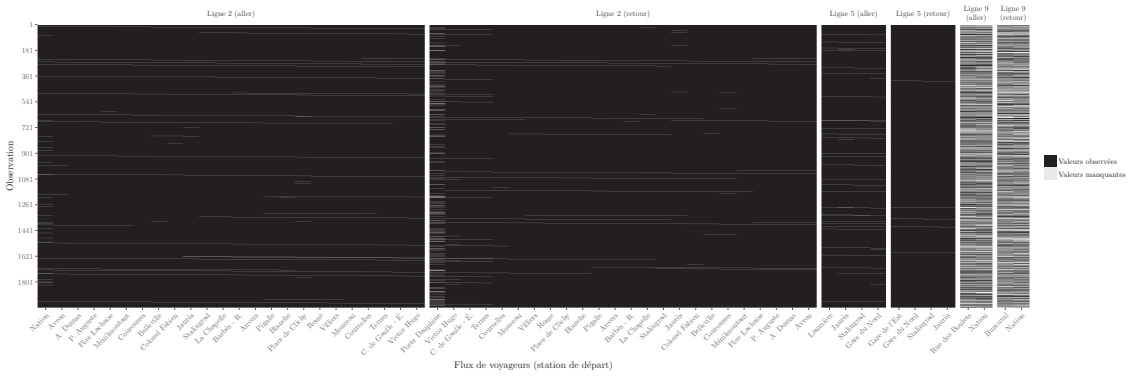
Comme nous l'avons évoqué dans le chapitre 3, les défaillances techniques et les absences de systèmes de collecte sont susceptibles d'engendrer des données manquantes. Dans cette expérimentation, le taux d'incomplétude des données est de 2.3 %. Il est inégalement réparti entre les flux de voyageurs (4.8 %) et les intervalles de départ (0.2 %).

Dans le cas des flux piétons, les données manquantes sont majoritairement concentrées sur les flux sortant des espaces contrôlés (vers la voie publique ou vers d'autres espaces contrôlés), avec un taux de 18.5 %. Par comparaison, ce taux est de seulement 2.3 % pour les flux entrant dans les espaces contrôlés depuis la voie publique. Comme le montre la figure 4.2a, les valeurs manquantes apparaissent par groupes de points séquentiels plus ou moins éloignés dans le temps. Elles ne peuvent donc pas être considérées comme MCAR, dans la mesure où leur probabilité d'absence dépend de leur voisinage temporel. Toutefois, la répartition de ces groupes

2. Le lundi 23 mars est sujet à un pic de pollution entraînant la gratuité des transports publics d'Île-de-France. En raison de son caractère atypique et de l'absence totale de données de validation, cette journée est écartée. Le lundi 6 avril étant férié, il ne fait pas non plus partie de notre jeu de données.



(a)



(b)

FIGURE 4.2 – Distribution des valeurs manquantes des flux de voyageurs : (a) flux piétons ; (b) flux de départ.

ne semble pas obéir à des patterns précis. Par conséquent, nous pouvons supposer que l'information connue est suffisante pour estimer la distribution des données manquantes, autrement dit que ces dernières sont de type MAR.

Dans le cas des flux de départ, le taux de données manquantes est de 1.1 % pour la ligne 2 et de 0.9 % pour la ligne 5. Seuls les flux partant des terminus de la ligne 2 affichent des taux supérieurs à 2 % (2.1 % au départ de Nation et 9.6 % au départ de Porte Dauphine). En ce qui concerne la ligne 9, certains trains ne sont pas équipés de systèmes de comptage, ce qui induit un taux d'incomplétude de 51.1 %. Comme le montre la figure 4.2b, les valeurs manquantes se présentent sous forme de séquences spatiales correspondant au cheminement des trains. Ainsi, il apparaît clairement qu'un train qui ne remonte pas de comptage pour un flux donné a moins de chances de remonter des comptages pour les flux voisins. La probabilité que la valeur d'un flux soit manquante dépend donc des matériels roulants impliqués dans la mesure de ce flux. La circulation des matériels dont le système de collecte est

défaillant (ou absent) étant aléatoire dans le temps, nous pouvons considérer que les données sont de type MAR.

4.2 Méthode expérimentale

Dans le cadre de cette expérimentation, nous souhaitons évaluer les performances de prédiction de l'approche développée dans le chapitre 3. Pour ce faire, nous divisons le jeu de données en deux échantillons de tailles différentes. Composé des cinq premières semaines de données, soit 24 journées (73 % des observations), le premier échantillon est utilisé pour l'apprentissage de la structure et des paramètres du réseau bayésien dynamique. Composé des deux semaines restantes, soit 9 journées (27 % des observations), le second échantillon (ou échantillon de test) est destiné à l'évaluation des performances de prédiction.

Dans un premier temps, les arcs candidats du réseau bayésien dynamique sont sélectionnés selon la démarche détaillée dans la section 3.4, après avoir préalablement défini les paramètres r_1 et r_2. Dans un second temps, la sous-structure optimale et les paramètres des distributions sont estimés à partir de l'échantillon d'apprentissage. Compte tenu de la présence de données MAR, il nous semble judicieux d'appliquer l'algorithme EM structurel décrit dans les sous-sections 2.5.2 et 2.6.4. Afin de réduire les temps de calcul, la réalisation de cet algorithme est limitée à 10 itérations, de même que l'algorithme EM paramétrique mis en œuvre lors de l'étape E. Nous montrons par la suite que ce nombre maximal d'itérations est amplement suffisant pour garantir la convergence.

Une fois que le réseau bayésien dynamique est construit, ses performances sont évaluées sur l'échantillon de test en simulant le temps réel. À chaque pas de temps t, notre objectif est de prédire les flux de voyageurs au pas de temps $t+1$ (c'est-à-dire observés au cours des deux prochaines minutes) en tenant compte des valeurs mesurées jusqu'à t. Étant donné que l'offre de transport est planifiée par l'opérateur de transport public, nous partons du principe que celle-ci est connue a priori et que nous disposons donc déjà des valeurs des intervalles de départ à $t+1$. Cette hypothèse peut paraître un peu trop optimiste, dans la mesure où des changements imprévus peuvent survenir au dernier moment du fait d'un incident d'exploitation ou d'une réadaptation de l'offre aux conditions du réseau. Il serait en effet plus réaliste de considérer les intervalles calculés à partir de l'offre de transport estimée (voir la sous-section 3.2.3), c'est-à-dire de l'offre prévue (et non réalisée) à $t+1$ au moment de la prédiction. Malheureusement, nous ne disposons pas d'un historique

de données pour ce type d'information et ne pouvons donc pas l'exploiter dans cette expérimentation.

Les problèmes d'inférence relatifs à l'algorithme EM structurel et à la prédiction des flux en temps réel sont traités par la méthode du filtre bootstrap en appliquant respectivement les algorithmes 2.5 et 2.6. Dans le cas de l'algorithme de prédiction en temps réel, les intervalles de départ à $t+1$ représentent le sous-ensemble de variables connues a priori Z^{t+1}. De manière empirique, un nombre de 1000 particules offre un bon compromis entre la vitesse d'exécution du filtre bootstrap et la précision de l'inférence.

Comme dans l'expérimentation de la sous-section 3.4.2, l'erreur de prédiction est mesurée à l'aide de la WMAPE (voir l'équation (3.7)). Les différentes méthodes sont évaluées au regard des 30 flux piétons entrant et sortant des espaces contrôlés desservis par la ligne 2 et des 48 flux de départ de cette ligne. Une plus grande attention est toutefois portée aux flux de départ, ceci pour les raisons suivantes :

— ces flux sont directement soumis à l'offre de transport, ce qui constitue l'un des défis majeurs de la modélisation ;

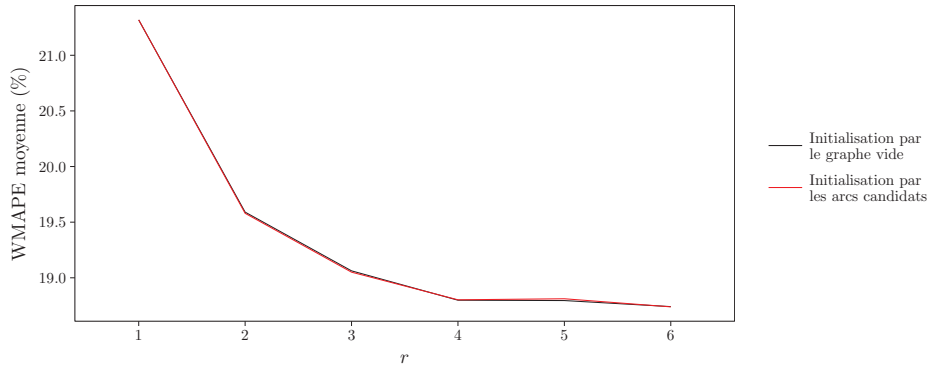
— contrairement aux flux piétons, situés en périphérie de notre terrain d'étude, la plupart des flux de départ possèdent des flux amont-adjacents et exploitent donc pleinement le potentiel de notre approche.

Les algorithmes d'apprentissage et d'inférence sont implémentés à l'aide du langage R (version 3.3), en exploitant les bibliothèques du tidyverse (Wickham et Grolemund, 2016). Il sont exécutés sur une machine Windows 7 64 bits équipée d'un processeur Intel Xeon E3-1240 v2 (3.4 GHz) et de 16 Go de mémoire vive.

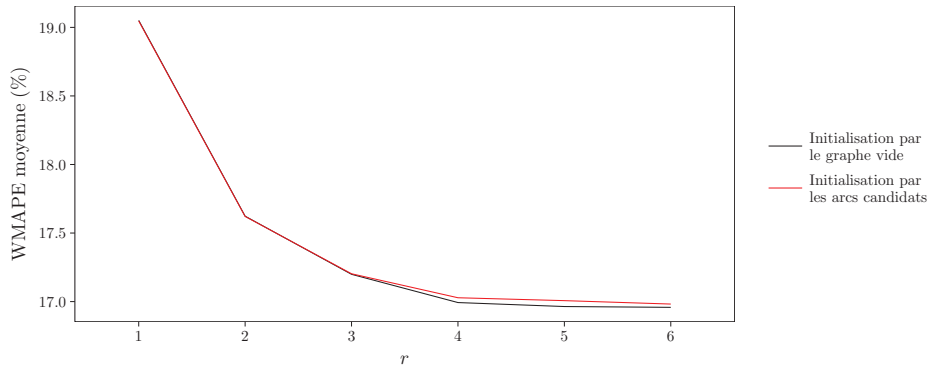
4.3 Utilisation de distributions gaussiennes

Dans la première partie de l'expérimentation, nous supposons que les flux de voyageurs et les intervalles de départ suivent une distribution gaussienne et que leurs relations sont linéaires. Partant de cette hypothèse, nous décrivons les distributions conditionnelles locales du réseau bayésien dynamique par des modèles linéaires gaussiens, dont les paramètres sont estimés par la méthode classique du maximum de vraisemblance (voir la sous-section 2.2.2).

Afin de mesurer l'influence de r_1 et r_2 sur les performances de prédiction, nous faisons varier ces paramètres entre 1 et 6. Pour limiter le nombre de tests, nous décidons de les regrouper en un unique paramètre $r = r_1 = r_2$, de manière à définir directement l'ordre du réseau bayésien dynamique. L'algorithme EM structurel étant



(a)



(b)

FIGURE 4.3 – Moyenne des erreurs de prédiction à $t+1$ selon l'ordre r du réseau bayésien gaussien dynamique, pour chaque initialisation de la structure : (a) flux piétons relatifs à la ligne 2 ; (b) flux de départ de la ligne 2.

sensible à la structure de départ, nous testons deux initialisations différentes : l'initialisation par le graphe vide et celle par le graphe contenant tous les arcs candidats sélectionnés.

Les résultats de prédiction à $t+1$ sur l'échantillon de test sont détaillés dans la figure 4.3. Dans la figure 4.3a, la moyenne des erreurs de prédiction des 30 flux piétons relatifs à la ligne 2 diminue à mesure que r augmente. Partant de 21.3 % pour $r = 1$, la WMAPE moyenne se stabilise entre 18.7 % et 18.8 % à partir de $r = 4$, ceci quelle que soit l'initialisation de la structure. Dans la figure 4.3b, une tendance similaire est observée pour les 48 flux de départ de la ligne 2, dont la WMAPE moyenne se stabilise également à partir de $r = 4$, autour de 17.0 % (contre 19.1 % pour $r = 1$). Ces résultats confirment notre intuition selon laquelle les processus qui régissent le comportement des flux ne sont pas nécessairement markoviens d'ordre 1. L'utilisation d'un réseau bayésien dynamique d'ordre plus élevé présente donc ici

r	Nombre d'arcs candidats	Nombre d'arcs retenus		Nombre d'arcs en commun
		Initialisation par le graphe vide	Initialisation par les arcs candidats	
1	677	516	519	516
2	1294	874	888	873
3	1911	1125	1162	1113
4	2528	1344	1401	1334
5	3145	1501	1593	1476
6	3762	1650	1777	1623

TABLE 4.1 – Nombre d'arcs candidats de $\mathcal{B}_{\rightarrow}$ retenus selon l'ordre r du réseau bayésien gaussien dynamique, pour chaque initialisation de la structure et communément aux deux initialisations.

un intérêt certain.

Les deux courbes de chaque graphique de la figure 4.3 étant presque toujours confondues, la structure initiale du réseau bayésien dynamique ne semble pas impacter significativement les performances de prédiction. Comme le montre la table 4.1, le nombre d'arcs candidats du réseau bayésien de transition $\mathcal{B}_{\rightarrow}$ retenus par l'algorithme EM structurel varie peu d'une initialisation à l'autre. En outre, la quasi-totalité des arcs retenus lors de l'initialisation par le graphe vide le sont également lors de l'initialisation par les arcs candidats. En considérant par exemple le réseau bayésien dynamique d'ordre 4, ces deux structures initiales aboutissent respectivement à la conservation de 1344 et 1401 arcs (sur 2528). Elles partagent 1334 arcs en commun, soit 99.3 % des arcs retenus lors de l'initialisation par le graphe vide. Ces résultats nous montrent que les deux structures finalement obtenues sont très proches l'une de l'autre, ce qui explique la similarité des performances de prédiction.

La recherche de la sous-structure optimale est plus rapide si nous partons du graphe vide. En effet, lorsque la structure est initialisée par le graphe contenant tous les arcs candidats, les premières itérations de l'algorithme EM structurel sont exécutées plus lentement en raison de la complexité plus importante du réseau bayésien dynamique (voir la figure 4.4). Par conséquent, puisque la structure initiale impacte peu les performances de prédiction, nous privilégions l'initialisation par le graphe vide dans toute la suite de l'expérimentation.

Il peut être intéressant d'analyser en détail la structure conservée par l'algorithme EM structurel. Pour ce faire, nous considérons le réseau bayésien dynamique d'ordre 4 initialisé par le graphe vide et calculons le taux d'arcs candidats de $\mathcal{B}_{\rightarrow}$ retenus selon le type de variable enfant et le type de variable parent. D'après les taux affichés

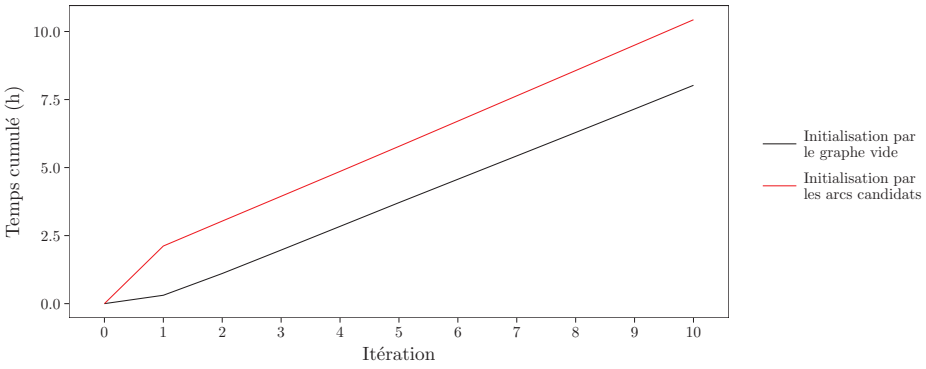


FIGURE 4.4 – Temps cumulé après chaque itération de l'algorithme EM structurel pour l'apprentissage du réseau bayésien gaussien dynamique d'ordre 4, pour chaque initialisation de la structure.

Type de variable enfant	Type de variable parent	Taux d'arcs retenus (%)
Flux piéton	Flux piéton amont-adjacent	46.9
Flux piéton	Flux de départ amont-adjacent	27.8
Flux piéton	Intervalle de départ amont-adjacent	30.0
Flux piéton	Valeur historique	83.6
Flux de départ	Flux piéton amont-adjacent	62.0
Flux de départ	Flux de départ amont-adjacent	
	(même ligne et même sens)	87.3
	(autre ligne ou autre sens)	8.9
Flux de départ	Intervalle de départ amont-adjacent	4.3
Flux de départ	Valeur historique	24.2
Flux de départ	Intervalle de départ associé	100.0
Intervalle de départ	Intervalle de départ amont-adjacent	63.8
Intervalle de départ	Valeur historique	97.1

TABLE 4.2 – Taux d'arcs candidats de $\mathcal{B}_{\rightarrow}$ retenus selon le type de variable enfant et le type de variable parent, dans le réseau bayésien gaussien dynamique d'ordre 4 initialisé par le graphe vide.

dans la table 4.2, cette structure optimale contient l'intégralité des arcs entre les flux de départ et leur intervalle de départ associé. Comme nous l'avions pressenti, l'offre de transport joue donc un rôle central dans la prédiction de ces flux.

Les relations d'adjacence occupent également une place importante dans la structure. En particulier, 87.3 % des arcs entre les flux de départ adjacents de même ligne et de même sens de circulation sont retenus, ainsi que 62.0 % des arcs entre les flux de départ et leurs flux piétons amont-adjacents. En revanche, le taux d'arcs retenus

est beaucoup plus faible entre les flux de départ adjacents de lignes ou de sens de circulation différents (8.9 %), ainsi qu'entre les flux de départ et leurs intervalles de départ amont-adjacents (4.3 %). En d'autres termes, les interactions entre les lignes en correspondance sont peu représentées dans la structure finale du réseau bayésien dynamique.

Concernant les relations de dépendance entre les valeurs courantes et historiques des flux de voyageurs, les arcs correspondants sont retenus à hauteur de 83.6 % pour les flux piétons, contre seulement 24.2 % pour les flux de départ. Cette différence s'explique principalement par la situation périphérique des flux piétons. La plupart de ces flux (28 sur 35) ne possédant pas de flux ni d'intervalle de départ amont-adjacent, leurs valeurs historiques constituent leurs seuls parents candidats, ce qui explique le taux élevé d'arcs retenus.

4.4 Utilisation de modèles de mélanges gaussiens

L'utilisation d'un réseau bayésien gaussien repose sur l'hypothèse de normalité des données et de linéarité des relations entre les variables. Or si la manipulation de distributions gaussiennes facilite les calculs, cette hypothèse s'avère très restrictive et peut donc légitimement être remise en question. En effet, nous pouvons nous demander si les flux de voyageurs obéissent à des processus non linéaires et si l'utilisation de distributions plus complexes, mais aussi plus flexibles, est susceptible d'améliorer les performances de prédiction.

Dans cette seconde partie de l'expérimentation, nous décrivons les distributions jointes locales du réseau bayésien dynamique par des modèles de mélanges gaussiens. Avec un nombre suffisant de composantes gaussiennes, ces derniers sont capables d'approximer une grande variété de distributions et de surmonter ainsi les limites des modèles linéaires gaussiens. Dans la suite de ces travaux, nous travaillons exclusivement avec le réseau bayésien dynamique d'ordre 4 ($r = r_1 = r_2 = 4$) dont la structure est initialisée par le graphe vide. Afin de limiter la charge de calcul, seules les distributions locales du réseau bayésien de transition $\mathcal{B}_{\rightarrow}$ sont décrites par des modèles de mélanges gaussiens. Les réseaux bayésiens initiaux $\mathcal{B}_1, \ldots, \mathcal{B}_4$ ayant une importance moindre dans la modélisation (leur unique rôle étant de représenter les distributions des pas de temps 1 à 4), leurs distributions sont décrites par de simples gaussiennes. Dans cette section, nous nous concentrons exclusivement sur la prédiction des 48 flux de départ de la ligne 2, qui présente davantage d'intérêt que celle des flux piétons.

4.4.1 Nombre de composantes fixé a priori

Dans un premier temps, nous partons du principe que le nombre de composantes M des modèles de mélanges gaussiens est fixé a priori. La structure du réseau bayésien dynamique étant initialisée par le graphe vide, les modèles de mélanges initiaux sont univariés. Leurs paramètres de départ sont définis de la manière suivante (pour la j-ème composante) :

$$\begin{aligned} \alpha_j &= \frac{1}{M} \\ \mu_j &= \mu + \epsilon_j \\ \Sigma_j &= \sigma^2 \,, \end{aligned} \tag{4.1}$$

où μ et σ^2 sont respectivement la moyenne et la variance calculées sur les valeurs observées de la variable, et ϵ_j est une petite perturbation aléatoire permettant de dissocier les composantes.

Dans le cadre de l'algorithme EM structurel, la mise à jour des paramètres des modèles de mélanges gaussiens est réalisée par le biais de l'algorithme EM, selon la démarche détaillée dans la sous-section 2.3.2. Afin de pallier les problèmes de singularité, une régularisation bayésienne est effectuée en choisissant $\lambda = 0.01$ (empiriquement, nous observons que la valeur de λ a un impact négligeable sur les résultats tant qu'elle demeure proche de 0). Les paramètres estimés maximisent donc localement la log-vraisemblance pénalisée explicitée dans l'équation (2.19). Pour finir, un test d'acceptation ou de rejet des paramètres est mis en place afin d'éviter une éventuelle dégradation de la log-vraisemblance conditionnelle (voir la sous-section 2.3.3).

Afin de tester différents niveaux de flexibilité, nous faisons varier le nombre de composantes M des modèles de mélanges gaussiens entre 2 et 6. Comme le montre la figure 4.5, la moyenne des erreurs de prédiction à $t+1$ des flux de départ de la ligne 2 diminue à mesure que M augmente. Partant d'une valeur de 15.9 % pour $M = 2$, la WMAPE moyenne se stabilise autour de 14.1 % à partir de $M = 5$. En comparant ces résultats avec ceux obtenus dans la première partie de l'expérimentation, nous constatons que l'utilisation de modèles de mélanges gaussiens aboutit à de meilleurs performances de prédiction que l'utilisation de simples distributions gaussiennes, ceci quelle que soit la valeur de M. Cette amélioration des performances valide l'hypothèse selon laquelle les flux suivent un comportement non linéaire, que la flexibilité des modèles de mélanges permet de mieux représenter.

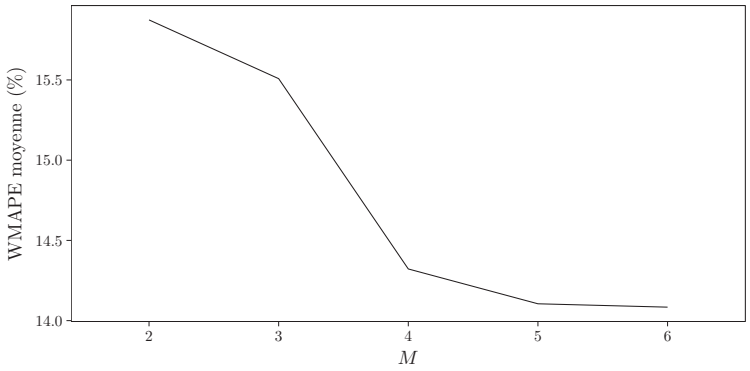


FIGURE 4.5 – Moyenne des erreurs de prédiction à $t+1$ des flux de départ de la ligne 2 selon le nombre M de composantes des modèles de mélanges gaussiens du réseau bayésien dynamique d'ordre 4 initialisé par le graphe vide.

4.4.2 Optimisation du nombre de composantes

Dans l'approche précédente, le nombre optimal de composantes des modèles de mélanges gaussiens est déterminé de manière empirique en réitérant les phases d'apprentissage et de test pour chaque valeur de M. Outre son caractère fastidieux, l'inconvénient de cette méthode est que le nombre optimal de composantes n'est pas nécessairement le même d'une distribution à l'autre. Définir un nombre M unique pour l'ensemble du réseau bayésien dynamique peut donc avoir des répercussions négatives sur les performances de prédiction. Afin de pallier ces problèmes, il est préférable d'utiliser une méthode d'apprentissage permettant d'optimiser automatiquement le nombre de composantes de chaque modèle de mélange gaussien. Cette méthode consiste à remplacer l'algorithme EM classique utilisé pour l'estimation des paramètres par l'algorithme EM de division-fusion détaillé dans l'algorithme 2.1.

Comme nous l'avons proposé dans la sous-section 2.5.3, il est possible d'implémenter deux versions différentes de l'algorithme EM structurel :

— la version initiale, qui consiste à appliquer l'algorithme EM de division-fusion à chaque estimation des paramètres ;
— la version alternative détaillée dans l'algorithme 2.3, qui consiste à appliquer l'algorithme EM de division-fusion une seule fois par itération, à l'issue de l'étape M (le reste du temps, les paramètres sont estimés par l'algorithme EM classique).

L'algorithme EM de division-fusion permet de déterminer le nombre optimal de composantes d'un modèle de mélange gaussien en procédant de manière itérative à des divisions et des fusions de composantes. Ces opérations sont choisies selon un

critère de sélection à maximiser. Dans la sous-section 2.3.5, nous proposons deux critères différents :

— le BIC de la distribution jointe locale, couramment utilisé pour l'optimisation de modèles de mélanges gaussiens ;

— le BIC de la distribution conditionnelle locale, dont la maximisation contribue directement à celle du BIC du réseau bayésien dynamique.

Afin de tester ces différentes configurations, nous proposons trois scénarios d'apprentissage du réseau bayésien dynamique :

Scénario 1 : L'algorithme EM structurel est implémenté dans sa version initiale. Le critère de sélection utilisé dans l'algorithme EM de division-fusion est le BIC de la distribution jointe locale (avec la pénalité de régularisation de la log-vraisemblance).

Scénario 2 : L'algorithme EM structurel est implémenté dans sa version alternative. Comme dans le scénario 1, le critère de sélection utilisé dans l'algorithme EM de division-fusion est le BIC de la distribution jointe locale (avec la pénalité de régularisation de la log-vraisemblance).

Scénario 3 : Comme dans le scénario 2, l'algorithme EM structurel est implémenté dans sa version alternative. En revanche, le critère de sélection utilisé dans l'algorithme EM de division-fusion est le BIC de la distribution conditionnelle locale.

Dans ces trois scénarios, les modèles de mélanges gaussiens du réseau bayésien dynamique sont initialisés avec une seule composante (c'est-à-dire comme de simples distributions gaussiennes), ceci afin de simplifier l'initialisation de leurs paramètres. Comme nous l'avons expliqué dans la sous-section 2.3.5, la mise en œuvre de l'algorithme EM de division-fusion nécessite de spécifier un paramètre $C_{\max}$ de classement des divisions et des fusions candidates. Pour chaque scénario d'apprentissage, nous faisons varier ce paramètre entre 1 et 4.

Les résultats de prédiction à $t+1$ des flux de départ de la ligne 2 sont présentés dans la figure 4.6. De manière assez surprenante, la valeur de $C_{\max}$ semble avoir peu d'impact sur les performances de prédiction des scénarios 1 et 2, la WMAPE moyenne oscillant dans tous les cas autour de 13.9 et 14.0 %. Cet impact est légèrement plus important dans le cas du scénario 3, avec une WMAPE moyenne variant de 14.3 % pour $C_{\max} = 1$ à 13.8 % pour $C_{\max} = 3$. Quoi qu'il en soit, les écarts constatés entre les trois scénarios demeurent relativement faibles et ne permettent pas de conclure à un scénario plus performant que les autres.

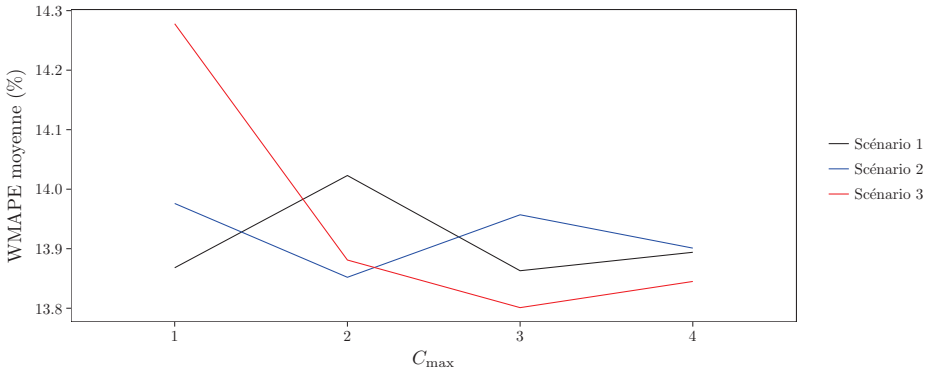


FIGURE 4.6 – Moyenne des erreurs de prédiction à $t+1$ des flux de départ de la ligne 2 selon le paramètre de classement $C_{\max}$, pour chaque scénario d'apprentissage du réseau bayésien à mélanges gaussiens dynamique.

$C_{\max}$	Temps d'exécution (h)		
	Scénario 1	Scénario 2	Scénario 3
1	63.1	13.1	8.4
2	120.6	16.2	9.5
3	176.5	19.9	10.8
4	231.1	22.7	12.0

TABLE 4.3 – Temps d'exécution des 10 premières itérations de l'algorithme EM structurel selon le paramètre de classement $C_{\max}$, pour chaque scénario d'apprentissage du réseau bayésien à mélanges gaussiens dynamique.

La principale différence entre les scénarios réside dans le temps nécessaire à la mise en œuvre de l'algorithme EM structurel. La table 4.3 contient les temps d'exécution des 10 premières itérations de cet algorithme (contenant chacune 10 itérations de l'algorithme EM paramétrique) pour chaque scénario et chaque valeur de $C_{\max}$. Comme attendu, le scénario 1 se révèle beaucoup plus coûteux que les scénarios 2 et 3, avec des temps de calcul allant de 63.1 heures pour $C_{\max} = 1$ à 231.1 heures (soit près de 10 jours !) pour $C_{\max} = 4$. Optimiser le nombre de composantes à chaque estimation des paramètres n'est donc clairement pas une option efficace, d'où l'intérêt d'utiliser la version alternative de l'algorithme EM structurel.

Un autre fait notable est que le scénario 3 implique des temps de calcul moins élevés que le scénario 2, ces derniers allant de 8.4 heures pour $C_{\max} = 1$ à 12.0 heures pour $C_{\max} = 4$ (contre respectivement 13.1 à 22.7 heures). Nous remarquons en effet que l'algorithme EM de division-fusion converge plus rapidement lorsque le critère de sélection à maximiser est le BIC de la distribution conditionnelle locale. Il réalise

$C_{\max}$	Nombre moyen de composantes		
	Scénario 1	Scénario 2	Scénario 3
1	5.8	5.0	2.9
2	6.4	5.8	3.4
3	6.6	6.3	3.7
4	6.7	6.5	3.8

(a)

$C_{\max}$	Nombre moyen d'arcs retenus		
	Scénario 1	Scénario 2	Scénario 3
1	2.1	2.3	2.9
2	2.0	2.2	2.8
3	2.0	2.1	2.7
4	2.0	2.2	2.7

(b)

TABLE 4.4 – Caractéristiques de $\mathcal{B}_{\rightarrow}$ selon le paramètre de classement $C_{\max}$, pour chaque scénario d'apprentissage du réseau bayésien à mélanges gaussiens dynamique : (a) nombre moyen de composantes par distribution locale ; (b) nombre moyen d'arcs retenus par structure locale.

moins d'itérations et opère donc moins de divisions ou de fusions de composantes que lorsque le critère de sélection est le BIC de la distribution jointe locale. Ce phénomène a des conséquences directes sur le nombre de composantes des modèles de mélanges gaussiens. Comme le montre la table 4.4a, chaque modèle de mélange de $\mathcal{B}_{\rightarrow}$ possède en moyenne 5.0 à 6.5 composantes à l'issue du scénario 2, contre seulement 2.9 à 3.8 composantes à l'issue du scénario 3. Ainsi, le scénario 3 possède le double avantage d'impliquer une charge de calcul plus faible et de générer des distributions avec moins de composantes.

L'utilisation de modèles de mélanges gaussiens a des conséquences intéressantes sur la structure graphique du réseau bayésien dynamique. En effet, le nombre d'arcs candidats retenus par l'algorithme EM structurel est beaucoup moins élevé que lorsque les distributions conditionnelles locales sont décrites par des modèles linéaires gaussiens. D'après la table 4.4b, le nombre d'arcs moyen par structure locale de $\mathcal{B}_{\rightarrow}$ n'excède pas 2.9, tandis que nous avons vu qu'il était de 6.4 dans le cas gaussien (1344 arcs pour 209 structures locales). Cette diminution de la taille de la structure vient contrebalancer l'augmentation de la complexité due aux modèles de mélanges gaussiens. Elle provient du fait que le BIC pénalise chaque distribution locale en

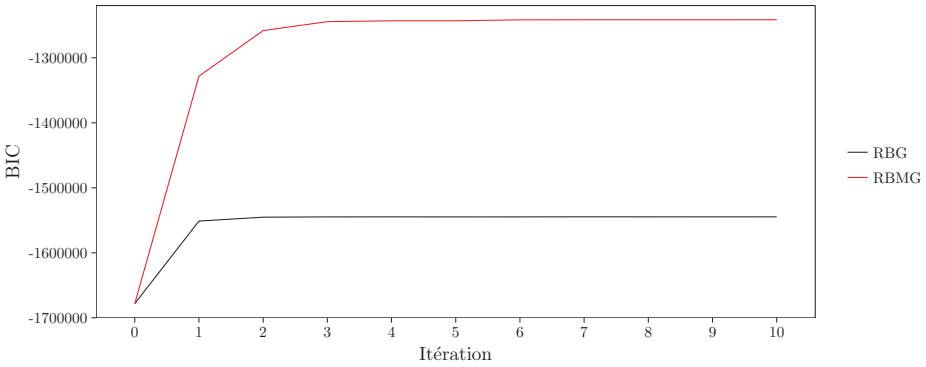


FIGURE 4.7 – BIC du réseau bayésien dynamique (sur l'échantillon d'apprentissage) après chaque itération de l'algorithme EM structurel, pour les méthodes de prédiction RBG et RBMG.

fonction de son nombre de paramètres libres, lequel dépend à la fois du nombre de composantes et du nombre de variables qui interviennent dans cette distribution. Ainsi, un modèle de mélange avec moins de composantes a tendance à impliquer davantage de variables, et inversement. C'est pour cette même raison que le scénario 3, qui génère des distributions avec moins de composantes que les scénarios 1 et 2, aboutit à des structures locales contenant en moyenne 2.7 à 2.9 arcs, contre seulement 2.0 à 2.3 pour les deux autres scénarios.

4.5 Étude comparative des méthodes de prédiction

Dans la section précédente, nous avons mis en évidence la capacité du réseau bayésien dynamique à mieux prédire les flux de départ de la ligne 2 lorsque ses distributions jointes locales sont décrites par des modèles de mélanges gaussiens. Afin d'approfondir cette comparaison, nous considérons les deux méthodes de prédiction suivantes :

RBG : La méthode RBG correspond au réseau bayésien gaussien dynamique d'ordre 4 dont la structure est initialisée par le graphe vide.

RBMG : La méthode RBMG correspond au réseau bayésien à mélanges gaussiens dynamique d'ordre 4 dont la structure est initialisée par le graphe vide et dont l'apprentissage est réalisé selon le scénario 3, en choisissant le paramètre de classement $C_{\max} = 3$.

Pour la méthode RBG comme pour la méthode RBMG, la convergence de l'algorithme EM structurel est relativement rapide. Comme le montre la figure 4.7, les

premières itérations contribuent très majoritairement à l'amélioration du BIC du réseau bayésien dynamique, ce qui justifie notre décision d'interrompre l'algorithme au bout de 10 itérations. Nous remarquons au passage que la progression du BIC induite par la méthode RBMG est plus de trois fois supérieure à celle induite par la méthode RBG.

4.5.1 Comparaison à l'échelle d'un flux

Nous souhaitons analyser plus précisément le comportement individuel des flux de voyageurs. Pour ce faire, nous prenons comme exemple le flux de départ de la ligne 2 partant de la station Villiers vers la station Monceau, que nous nommons de manière plus concise flux de départ « Villiers - Monceau ». D'après le plan de la figure 4.1, ce flux est aval-adjacent :

— au flux piéton entrant dans l'espace contrôlé de Villiers ;
— aux flux de départ de la ligne 2 Rome - Villiers et Monceau - Villiers ;
— aux intervalles de départ de la ligne 3 Europe - Villiers et Malesherbes - Villiers.

Dans le réseau bayésien dynamique d'ordre 4, le flux Villiers - Monceau au pas de temps t admet 25 parents candidats, composés de ses flux et intervalles de départ amont-adjacents à $t-1,\dots,t-4$, de ses propres valeurs à $t-1,\dots,t-4$ et de son intervalle de départ associé à t. Dans le cas de la méthode RBG, ses 8 parents retenus dans le réseau bayésien de transition $\mathcal{B}_{\rightarrow}$ sont :

— le flux piéton entrant dans l'espace contrôlé de Villiers à $t-1$, $t-2$ et $t-4$;
— le flux de départ de la ligne 2 Rome - Villiers à $t-1$, $t-2$ et $t-3$;
— sa propre valeur à $t-2$;
— son intervalle de départ associé à t.

Dans le cas de la méthode RBMG, le flux Villiers - Monceau à t n'admet plus que 3 parents dans $\mathcal{B}_{\rightarrow}$:

— le flux piéton entrant dans l'espace contrôlé de Villiers à $t-4$;
— le flux de départ de la ligne 2 Rome - Villiers à $t-1$;
— son intervalle de départ associé à t.

Le modèle de mélange gaussien décrivant la distribution jointe locale entre ces variables est représenté de manière bivariée dans la figure 4.8. Globalement, ses cinq composantes sont correctement ajustées aux données d'apprentissage et se concentrent sur les régions à forte densité de points. Elles témoignent de la capacité de l'algorithme EM de division-fusion à bien optimiser les paramètres du modèle de

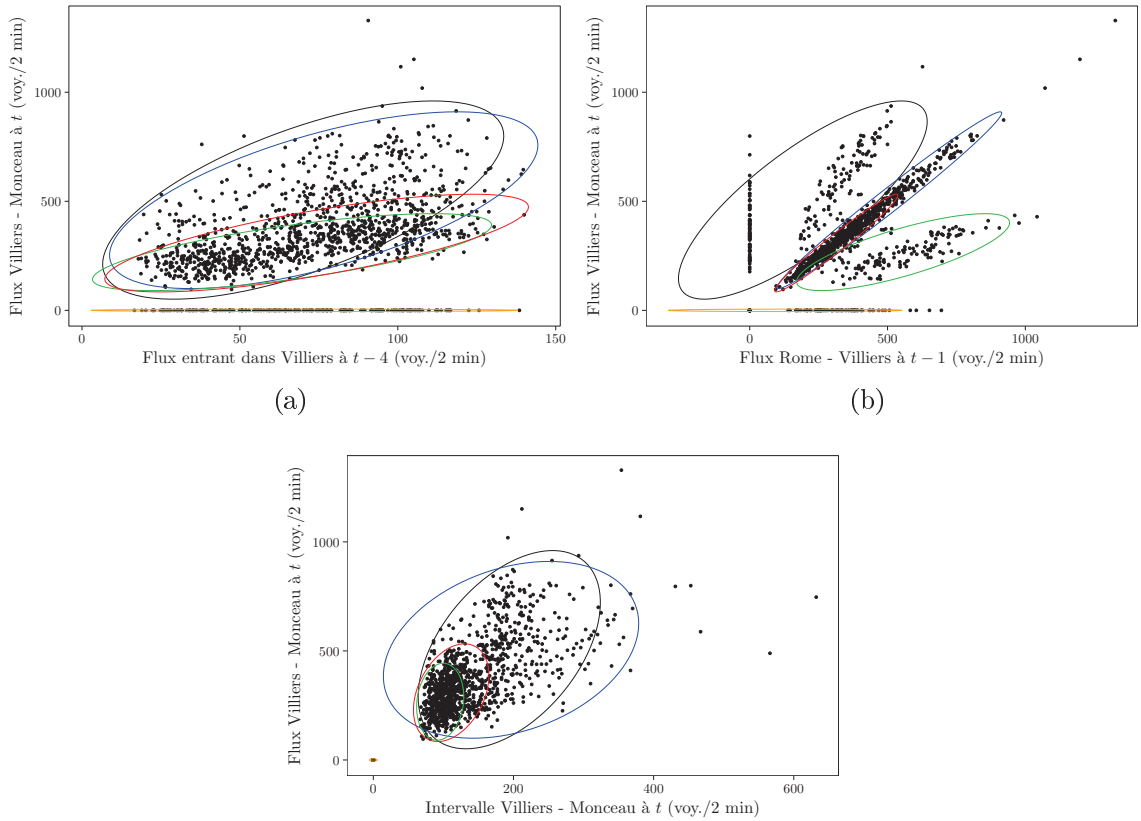


FIGURE 4.8 – Représentation bivariée des données d'apprentissage et du modèle de mélange gaussien décrivant la distribution jointe locale de $\mathcal{B}_{\rightarrow}$ entre le flux de départ Villiers - Monceau et ses parents, pour la méthode RBMG : flux Villiers - Monceau à t en fonction (a) du flux entrant dans l'espace contrôlé de Villiers à $t-4$, (b) du flux de départ Rome - Villiers à $t-1$, (c) de l'intervalle de départ Villiers - Monceau à t. Les données manquantes sont estimées par inférence à partir des données observées et des paramètres. Les ellipses de même couleur représentent la même composante.

mélange gaussien.

Dans la figure 4.8b, les données sont réparties sous forme de nuages de points allongés, auxquels viennent s'ajuster les composantes du modèle de mélange gaussien. La pente décrite par chacun de ces nuages est liée au nombre de départs de trains opérés pour le flux Villiers - Monceau à t et pour le flux Rome - Villiers à $t-1$. Comme nous pouvons le voir, ce nombre de départs constitue un important critère de différenciation des composantes. Ainsi, les points générés par les composantes rouge et bleue correspondent aux situations (les plus fréquentes) où le nombre de départs est le même pour les deux flux. Les points générés par les composantes noire et verte correspondent aux situations où le nombre de départs opérés pour le

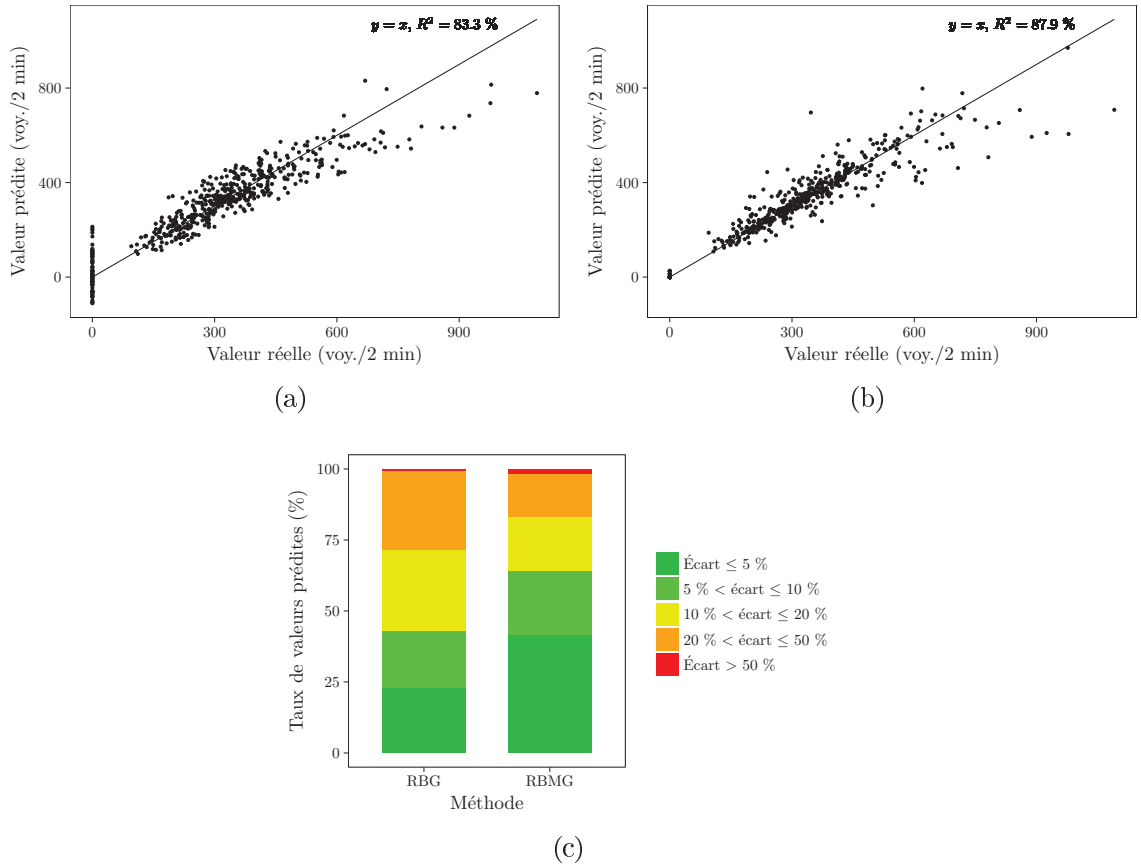


FIGURE 4.9 – Valeurs prédites à $t+1$ en fonction des valeurs réelles du flux de départ Villiers - Monceau sur l'échantillon de test : (a) méthode RBG ; (b) méthode RBMG. (c) Répartition des valeurs prédites selon leur écart relatif absolu avec les valeurs réelles, pour les deux méthodes de prédiction (valeurs non nulles uniquement).

flux Villiers - Monceau à t est respectivement supérieur et inférieur au nombre de départ opérés pour le flux Rome - Villiers à $t-1$. Enfin, les points générés par la composante orange caractérisent l'absence de départ pour le flux Villiers - Monceau à t. Ces derniers sont entièrement expliqués par les valeurs nulles de l'intervalle de départ associé à t (voir la figure 4.8c).

Lors de la prédiction à $t+1$ sur l'échantillon de test, la WMAPE obtenue pour le flux Villiers - Monceau est de 16.8 % pour la méthode RBG et 11.2 % pour la méthode RBMG. Dans les figures 4.9a et 4.9b, les valeurs prédites par chacune de ces méthodes sont exprimées en fonction des valeurs réelles. En comparant ces deux figures, nous remarquons que les écarts obtenus par la méthode RBMG sont plus faibles. Les points se concentrent davantage sur la droite d'équation $y = x$, avec un coefficient de détermination R^2 de 87.9 %, contre 83.3 % pour la méthode RBG. Ces

meilleurs résultats de prédiction sont corroborés par la figure 4.9c. Dans le cas de la méthode RBMG, 41.5 % des valeurs non nulles sont prédites avec un écart relatif absolu inférieur à 5 %, 64.2 % avec un écart inférieur à 10 % et 83.3 % avec un écart inférieur à 20 %. Dans le cas de la méthode RBG, ces taux sont respectivement de 23.1 %, 43.1 % et 71.6 %.

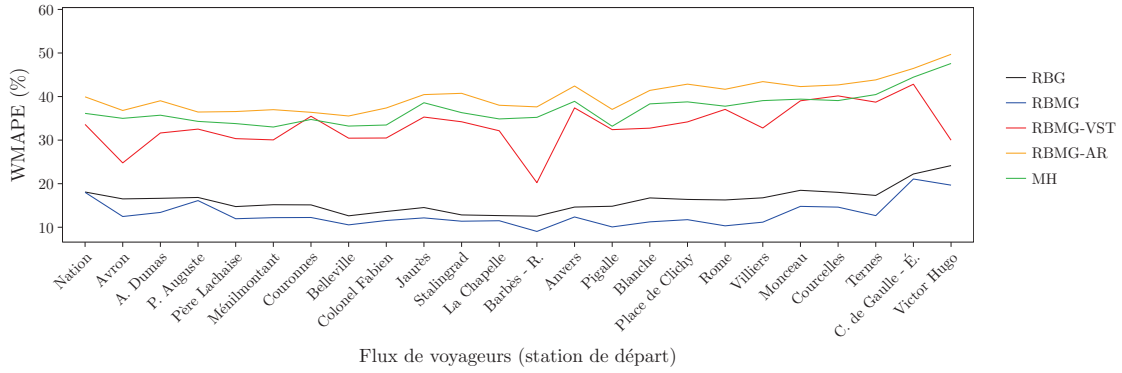
Concernant la méthode RBMG, nous avons vu qu'une composante du modèle de mélange gaussien (la composante orange dans la figure 4.8) était entièrement dédiée à l'explication des valeurs nulles du flux Villiers - Monceau. Comme le montre la figure 4.9b, la présence de cette composante favorise la prédiction de ces valeurs nulles, les écarts observés étant beaucoup plus faibles que dans la figure 4.9a. Dans les deux graphiques, nous constatons toutefois que la plupart des points correspondant aux valeurs les plus élevées se retrouvent en dessous de la droite d'équation $y = x$. Ce phénomène témoigne d'une certaine tendance de la méthode RBMG, comme de la méthode RBG, à sous-estimer ces valeurs extrêmes.

4.5.2 Comparaison à l'échelle de la ligne

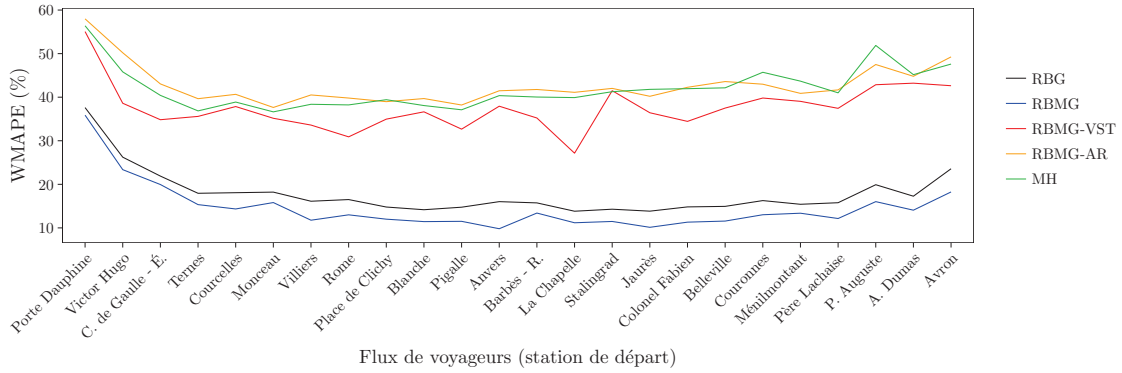
À présent, nous souhaitons élargir la comparaison à l'ensemble des flux de départ de la ligne 2. Afin d'évaluer plus précisément les bénéfices de notre approche, nous introduisons trois nouvelles méthodes de prédiction :

RBMG-VST : Comme la méthode RBMG, la méthode RBMG-VST consiste à utiliser le réseau bayésien à mélanges gaussiens dynamique d'ordre 4 dont la structure est initialisée par le graphe vide. La différence est que les flux sont prédits à partir de leurs flux amont-adjacents et de leurs valeurs historiques, sans prendre en compte les intervalles de départ. Ce principe de construction de la structure est analogue à celui proposé dans l'approche par les réseaux bayésiens de Sun *et al.* (2006). Afin de faciliter la comparaison avec la méthode RBMG, l'apprentissage est également réalisé selon le scénario 3, avec le paramètre de classement $C_{\max} = 3$.

RBMG-AR : La méthode RBMG-AR consiste aussi à utiliser le réseau bayésien à mélanges gaussiens dynamique d'ordre 4 dont la structure est initialisée par le graphe vide. La différence est que les flux sont prédits de manière « autorégressive », uniquement à partir de leurs valeurs historiques. Ce principe de construction de la structure est analogue à celui proposé dans l'approche par les chaînes de Markov de Yu *et al.* (2003). L'apprentissage est également réalisé selon le scénario 3, avec le paramètre de classement $C_{\max} = 3$.



(a)



(b)

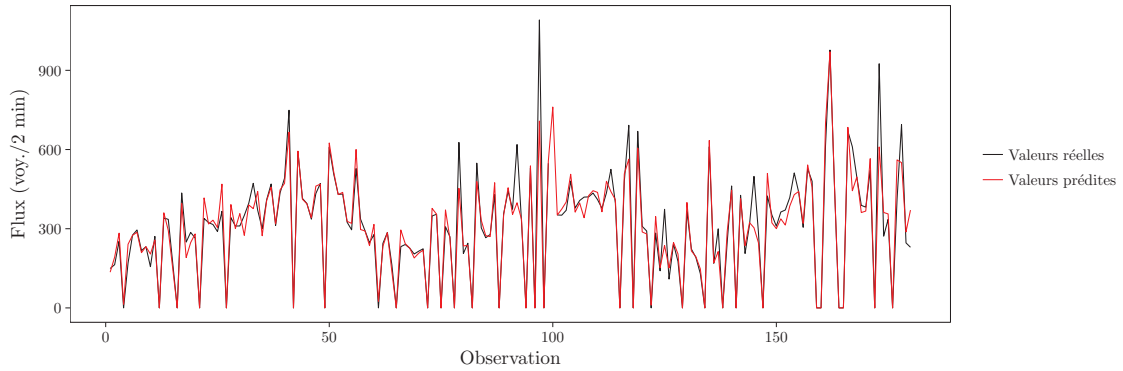
FIGURE 4.10 – Erreurs de prédiction à $t+1$ des flux de départ de la ligne 2, pour chaque méthode de prédiction : (a) sens aller (direction Porte Dauphine) ; (b) sens retour (direction Nation).

MH : La méthode MH désigne la méthode de la moyenne historique. Cette dernière consiste à prédire la valeur de chaque flux à un pas de temps donné en calculant la moyenne des valeurs de ce flux observées au même pas de temps dans l'échantillon d'apprentissage.

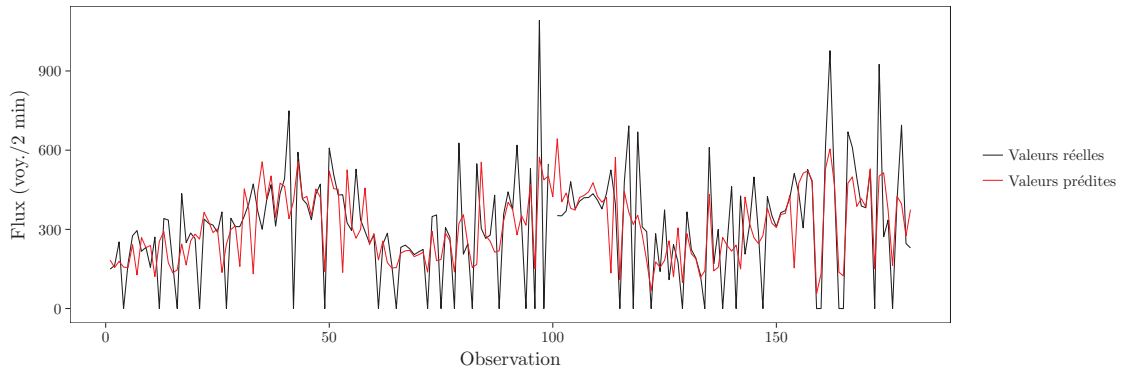
Les performances des méthodes RBG, RBMG, RBMG-VST, RBMG-AR et MH sont comparées sur l'échantillon de test. Les erreurs de prédiction à $t+1$ des 48 flux de départ de la ligne 2 sont listées dans la table 4.5 et visualisées dans la figure 4.10, dans l'ordre topologique de chaque sens de circulation. Le principal enseignement de ces résultats est que la méthode RBMG surpasse les quatre autres méthodes de prédiction, ceci quel que soit le flux considéré. En comparant ses performances avec celles de la méthode RBG, nous constatons que 42 flux sur 48 observent une baisse relative de leur WMAPE de plus de 10 %, cette baisse dépassant même 20 % dans le cas de 18 flux. Rappelons que les WMAPEs moyennes obtenues par ces deux

Sens	Flux de voyageurs (station de départ)	WMAPE (%)				
		RBG	RBMG	RBMG-VST	RBMG-AR	MH
Aller	Nation	18.1	18.0	33.6	39.9	36.1
Aller	Avron	16.5	12.5	24.8	36.8	35.0
Aller	Alexandre Dumas	16.6	13.4	31.6	39.0	35.7
Aller	Philippe Auguste	16.8	16.1	32.5	36.4	34.3
Aller	Père Lachaise	14.7	12.0	30.3	36.6	33.8
Aller	Ménilmontant	15.2	12.2	30.1	37.0	33.0
Aller	Couronnes	15.1	12.2	35.5	36.4	34.7
Aller	Belleville	12.6	10.5	30.4	35.5	33.2
Aller	Colonel Fabien	13.6	11.6	30.5	37.4	33.5
Aller	Jaurès	14.5	12.1	35.3	40.4	38.6
Aller	Stalingrad	12.8	11.4	34.2	40.7	36.3
Aller	La Chapelle	12.7	11.5	32.1	38.0	34.9
Aller	Barbès - Rochechouart	12.5	9.0	20.2	37.6	35.2
Aller	Anvers	14.6	12.4	37.4	42.4	38.9
Aller	Pigalle	14.8	10.1	32.4	37.1	33.2
Aller	Blanche	16.7	11.2	32.7	41.4	38.3
Aller	Place de Clichy	16.4	11.7	34.2	42.9	38.8
Aller	Rome	16.3	10.3	37.1	41.7	37.8
Aller	Villiers	16.8	11.2	32.8	43.4	39.1
Aller	Monceau	18.5	14.8	39.0	42.3	39.4
Aller	Courcelles	18.0	14.6	40.2	42.7	39.1
Aller	Ternes	17.3	12.7	38.7	43.8	40.5
Aller	Charles de Gaulle - Étoile	22.2	21.1	42.8	46.5	44.5
Aller	Victor Hugo	24.2	19.7	30.0	49.7	47.6
Retour	Porte Dauphine	37.6	35.9	55.1	58.0	56.4
Retour	Victor Hugo	26.2	23.4	38.6	50.2	45.8
Retour	Charles de Gaulle - Étoile	21.9	20.0	34.8	43.0	40.4
Retour	Ternes	17.9	15.3	35.6	39.7	36.9
Retour	Courcelles	18.1	14.3	37.9	40.7	38.9
Retour	Monceau	18.2	15.8	35.2	37.6	36.6
Retour	Villiers	16.1	11.8	33.6	40.5	38.4
Retour	Rome	16.5	13.0	30.9	39.8	38.2
Retour	Place de Clichy	14.8	12.0	35.0	39.0	39.4
Retour	Blanche	14.2	11.4	36.6	39.7	38.1
Retour	Pigalle	14.7	11.5	32.7	38.2	37.1
Retour	Anvers	16.0	9.8	37.9	41.5	40.4
Retour	Barbès - Rochechouart	15.7	13.4	35.2	41.8	40.0
Retour	La Chapelle	13.8	11.2	27.1	41.1	39.9
Retour	Stalingrad	14.3	11.5	41.5	42.0	41.3
Retour	Jaurès	13.8	10.1	36.4	40.2	41.8
Retour	Colonel Fabien	14.8	11.3	34.4	42.3	42.0
Retour	Belleville	14.9	11.6	37.5	43.6	42.2
Retour	Couronnes	16.3	13.0	39.8	43.0	45.7
Retour	Ménilmontant	15.4	13.4	39.0	40.9	43.7
Retour	Père Lachaise	15.8	12.2	37.4	41.7	41.0
Retour	Philippe Auguste	19.9	16.0	42.9	47.5	51.9
Retour	Alexandre Dumas	17.3	14.1	43.2	44.8	45.2
Retour	Avron	23.6	18.2	42.6	49.3	47.6
	WMAPE moyenne	17.0	13.8	35.4	41.5	39.6

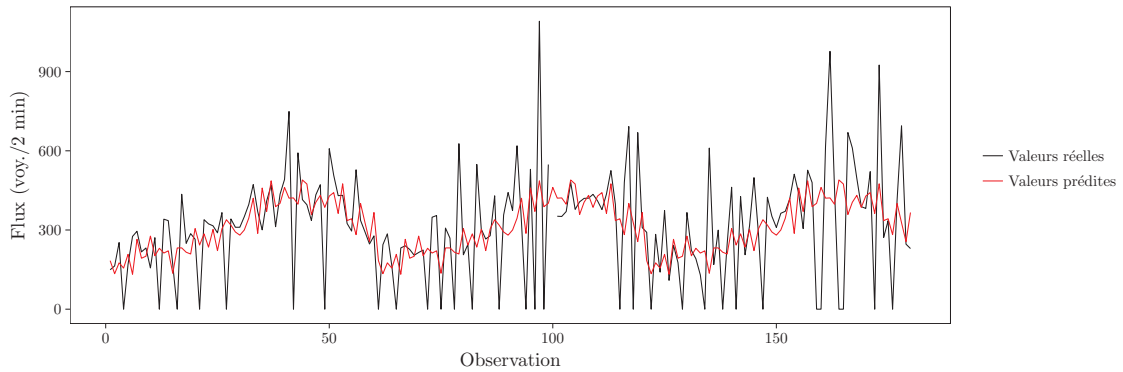
TABLE 4.5 – Erreurs de prédiction à $t + 1$ des flux de départ de la ligne 2, pour chaque méthode de prédiction.



(a)



(b)



(c)

FIGURE 4.11 – Valeurs réelles et prédites à $t+1$ du flux de départ Villiers - Monceau sur les trois premières journées de l'échantillon de test (du 7 au 9 avril 2015) : (a) méthode RBMG ; (b) méthode RBMG-VST ; (c) méthode MH.

méthodes sont respectivement de 13.8 % et 17.0 %.

Avec des WMAPEs moyennes respectives de 35.4 % et 41.5 %, la méthode RBMG-VST affiche des performances de prédiction supérieures à la méthode RBMG-AR, ceci pour l'ensemble des flux de départ de la ligne 2. Ces résultats illustrent

l'intérêt d'exploiter le voisinage spatio-temporel des flux, plutôt que de prédire ces derniers à partir de leurs seules valeurs historiques. Toutefois, la contribution la plus significative provient de l'intégration de l'offre de transport. En effet, les méthodes RBG et RBMG obtiennent des erreurs beaucoup plus faibles que la méthode RBMG-VST, ce qui confirme le rôle fondamental de ce type de données dans la modélisation.

Afin de mieux visualiser les bénéfices de l'offre de transport, les valeurs réelles et prédites du flux Villiers - Monceau sur les trois premières journées de l'échantillon de test (du 7 au 9 avril 2015) sont représentées dans les figures 4.11a pour la méthode RBMG et 4.11b pour la méthode RBMG-VST. Dans le cas de la méthode RBMG, les fluctuations inhérentes à la circulation des trains sont bien modélisées, malgré quelques imprécisions concernant la prédiction des valeurs les plus élevées (comme nous l'avons déjà constaté dans la figure 4.9b). Dans le cas de la méthode RBMG-VST, l'amplitude de ces fluctuations est en revanche largement sous-estimée.

Avec une WMAPE moyenne de 39.6 %, la méthode MH se révèle moins performante que les méthodes RBG, RBMG et RBMG-VST. Contrairement à ces dernières, la moyenne historique repose uniquement sur les situations observées par le passé. Elle ne tient pas compte des conditions courantes du réseau et s'adapte donc mal aux phénomènes non récurrents qui surviennent en temps réel. En outre, ses prédictions pâtissent de l'effet « lissant » de la moyenne, qui entrave sa capacité à modéliser les larges fluctuations. Les valeurs réelles et prédites du flux Villiers - Monceau pour la méthode MH sont représentées dans la figure 4.11c.

Dans la figure 4.10, nous remarquons que les performances des cinq méthodes de prédiction se dégradent aux extrémités de la ligne, notamment sur la section entre les stations Porte Dauphine et Charles de Gaulle - Étoile. Lorsque les trains circulent à proximité des terminus, le nombre de voyageurs à bord est plus faible que sur le reste de la ligne. Par conséquent, les flux concernés sont mesurés avec moins de précision par les systèmes de comptage par pesée. Les données utilisées lors de l'apprentissage et de la prédiction sont donc de moins bonne qualité, ce qui explique cette baisse de performances.

4.5.3 Cas des flux piétons

Jusqu'à présent, nous avons comparé les performances des cinq méthodes de prédiction au regard des flux de départ de la ligne 2. Bien que les flux piétons présentent un intérêt limité dans cette expérimentation, ceci pour les raisons évoquées dans la

Méthode	WMAPE moyenne (%)
RBG	18.8
RBMG	19.1
RBMG-VST	19.3
RBMG-AR	19.4
MH	17.2

TABLE 4.6 – Moyenne des erreurs de prédiction à $t+1$ des flux piétons relatifs à la ligne 2, pour chaque méthode de prédiction.

section 4.2, il peut être intéressant de se pencher brièvement sur leurs résultats de prédiction.

Les moyennes des erreurs à $t+1$ des 30 flux piétons relatifs à la ligne 2 sont détaillées dans la table 4.6. Dans le cas présent, les différences de performances sont beaucoup moins marquées que lors de la prédiction des flux de départ. En effet, les WMAPEs moyennes obtenues par les méthodes basées sur les réseaux bayésiens dynamiques varient de 18.8 % pour la méthode RBG à 19.4 % pour la méthode RBMG-AR. La prise en compte du voisinage spatio-temporel et de l'offre de transport apporte peu de bénéfices, ce qui s'explique par la situation géographique des flux piétons. Comme nous l'avons vu précédemment, la plupart de ces flux ne sont pas directement soumis à l'offre de transport et ne possèdent pas de flux amont-adjacents. Par conséquent, très peu d'entre eux exploitent réellement ces principes de modélisation. Notons enfin que l'utilisation de modèles de mélanges gaussiens ne semble pas non plus améliorer la qualité de la prédiction, la méthode RBMG obtenant même des résultats légèrement inférieurs à ceux de la méthode RBG (avec une WMAPE moyenne de 19.1 %).

De manière assez surprenante, les meilleures performances de prédiction sont obtenues par la méthode MH, avec une WMAPE moyenne de 17.2 %. Ces résultats s'expliquent par une certaine régularité des flux piétons, que la moyenne historique parvient correctement à retranscrire (voir la figure 4.12 pour l'exemple du flux entrant dans la station Villiers). Les flux piétons n'étant pas sujets à de larges fluctuations, ils pâtissent moins que les flux de départ de l'effet lissant de la moyenne. Néanmoins, il convient de relativiser la supériorité de la méthode MH, dans la mesure où les données utilisées ne contiennent pas d'événement majeur (non récurrent) ayant un impact significatif sur les flux piétons (perturbation d'exploitation, fermeture de station, manifestation culturelle ou sportive, etc.). En présence d'un tel événement, l'incapacité de la moyenne historique à prendre en compte les conditions

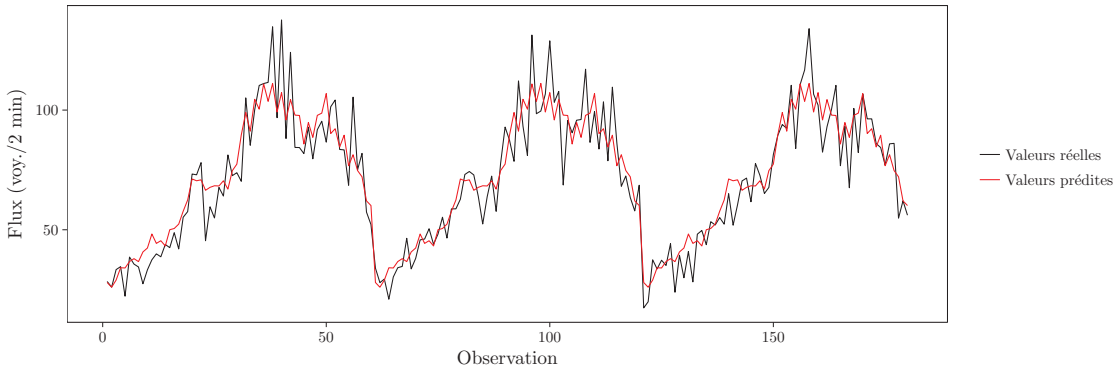


FIGURE 4.12 – Valeurs réelles et prédites à $t+1$ par la méthode MH du flux piéton entrant dans la station Villiers sur les trois premières journées de l'échantillon de test (du 7 au 9 avril 2015).

courantes du réseau entraînerait certainement une dégradation des performances de prédiction.

4.5.4 Extension de l'horizon de prédiction

Tout au long de cette expérimentation, nous nous sommes contentés de prédire les flux de voyageurs au pas de temps $t+1$, c'est-à-dire au cours des deux prochaines minutes. Cependant, il est parfois nécessaire de réaliser ces prédictions plus en amont dans le temps, afin de laisser une marge de manœuvre suffisante à l'utilisateur pour réagir, par exemple en cas de mise en place d'un dispositif de régulation. C'est pourquoi il peut être intéressant d'analyser les performances de notre approche à des horizons de prédiction plus lointains. Pour ce faire, nous faisons varier l'horizon h jusqu'à 10 pas de temps (soit 20 minutes) dans le futur. Comme pour la prédiction à $t+1$, la prédiction à $t+h$ est réalisée à l'aide du filtre bootstrap (voir l'algorithme 2.6). Nous admettons également que l'offre de transport est connue a priori, bien que le réalisme de cette hypothèse soit de plus en plus discutable à mesure que h augmente (en conséquence, les résultats de ce test doivent être considérés avec précaution).

Les résultats de prédiction des flux de départ de la ligne 2 sont présentés dans la figure 4.13, pour les méthodes RBG et RBMG. Sans surprise, les performances de ces deux méthodes diminuent quand l'horizon de prédiction h augmente. Cette dégradation est particulièrement marquée lorsque nous comparons les résultats de prédiction à $t+1$ et à $t+2$, la WMAPE moyenne passant respectivement de 13.8 % à 16.9 % dans le cas de la méthode RBMG. Par ailleurs, les différences observées entre les méthodes RBG et RBMG s'estompent au fur et à mesure que l'horizon s'éloigne,

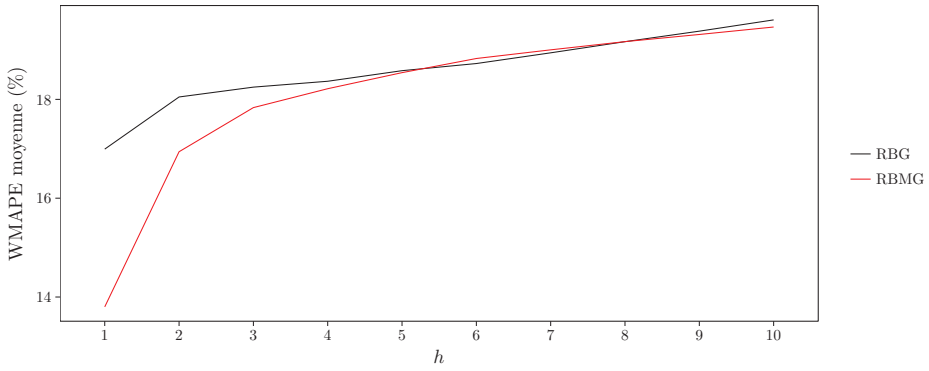


FIGURE 4.13 – Moyenne des erreurs de prédiction à $t + h$ des flux de départ de la ligne 2 selon l'horizon de prédiction h, pour les méthodes RBG et RBMG.

les deux courbes du graphique finissant par se confondre à partir de $h = 5$. En d'autres termes, l'utilisation de modèles de mélanges gaussiens contribue de moins en moins à l'amélioration des performances, jusqu'à ne plus du tout impacter les résultats si les flux sont prédits à $t + 5$ ou au-delà.

Conclusion et perspectives

Conclusion

Au cours de ces travaux, nous avons tout d'abord présenté la démarche de prévision des flux à court terme, avant de dresser un état de l'art des différentes méthodes de prédiction. Abordée principalement sous l'angle du trafic routier, cette revue de la littérature nous a permis de mettre en lumière la diversité des méthodes existantes, mais aussi l'absence d'un réel consensus sur l'approche la plus adaptée. Nous nous sommes plus particulièrement intéressés aux méthodes exploitant le voisinage spatio-temporel des flux, ainsi qu'aux rares modèles capables de fournir des prédictions en présence de données incomplètes.

Par leur représentation intuitive des relations de causalité spatio-temporelles et leur aptitude intrinsèque à gérer les données manquantes, les réseaux bayésiens constituent des outils particulièrement adaptés à la prévision des flux à court terme. Une partie importante de cette thèse est consacrée à la description des algorithmes d'apprentissage et d'inférence relatifs à ces objets mathématiques. Dans le cadre des réseaux bayésiens à mélanges gaussiens, nous avons notamment développé un algorithme EM de division-fusion permettant d'optimiser automatiquement le nombre de composantes des modèles de mélanges gaussiens. Par la suite, nous avons introduit les réseaux bayésiens dynamiques et présenté un algorithme d'inférence capable de fournir des prédictions en temps réel (par la méthode du filtre bootstrap).

Afin de réunir les données de validation, de comptage par pesée et d'offre de transport au sein d'un même formalisme, nous avons défini un référentiel spatial commun par le biais d'une description topologique du réseau de transport public. Ce référentiel nous a permis de caractériser les relations spatiales entre les données, dont découle directement la structure de notre modèle de prévision des flux à court terme. Basé sur les réseaux bayésiens dynamiques, ce modèle repose sur des mécanismes de construction intuitifs, issus de la connaissance experte des interactions entre les flux de voyageurs. L'intégration de l'offre de transport, par l'intermédiaire des intervalles

de départ des trains, constitue l'un des intérêts majeurs de notre approche. Bien que notre étude se concentre sur le réseau ferré de la RATP, les principes de modélisation que nous avons développés s'exportent facilement à d'autres types de réseaux.

Dans le cadre de l'expérimentation réalisée sur la ligne 2 du métro, nous avons d'abord montré que l'augmentation de l'ordre du réseau bayésien dynamique a une influence positive sur les performances de prédiction. En d'autres termes, les processus qui régissent le comportement des flux ne sont pas markoviens d'ordre 1. Dans le cas des flux de départ, nous avons également observé que les performances obtenues en décrivant les distributions jointes locales par des modèles de mélanges gaussiens surpassent celles obtenues avec des distributions gaussiennes (bien que la différence s'estompe quand l'horizon de prédiction augmente). Ce résultat important illustre le caractère non linéaire des relations entre les variables, que la flexibilité des modèles de mélanges gaussiens permet de mieux représenter. Enfin, nous avons comparé notre approche à d'autres méthodes de prédiction, afin d'évaluer la pertinence des principes de construction développés dans ces travaux. Les résultats de cette comparaison témoignent de l'intérêt d'exploiter le voisinage spatio-temporel des flux, ainsi que du rôle fondamental de l'offre de transport dans la prédiction.

Perspectives

À l'exception de la moyenne historique, toutes les méthodes de prédiction que nous avons testées reposent sur le formalisme des réseaux bayésiens. Or afin d'évaluer notre approche de manière plus globale, il serait judicieux de confronter ses résultats avec ceux fournis par d'autres types de modèles. Une étude comparative pourrait par exemple être réalisée avec des méthodes basées sur les réseaux de neurones, telles que celle développée par Toqué *et al.* (2017) pour la prévision de flux de voyageurs multimodaux du réseau de transport d'Île-de-France.

Notre intérêt pour les réseaux bayésiens se justifie entre autres par leur capacité à gérer les données manquantes. Cette propriété est particulièrement utile dans un contexte de temps réel, où l'implémentation d'un algorithme d'imputation adapté n'est pas toujours aisée. Dans l'expérimentation menée sur la ligne 2, le taux d'incomplétude du jeu de données demeure relativement faible, ce qui limite ses effets sur les performances de notre modèle. À l'avenir, il conviendrait de tester des taux plus élevés, afin de mesurer leur impact sur la qualité de prédiction des flux et sur l'efficacité de l'apprentissage opéré par l'algorithme EM structurel.

Au cours de notre étude, nous ne nous sommes pas vraiment attardés sur les

perturbations qui affectent le réseau de transport public. Les données que nous avons utilisées ne contiennent pas d'incident d'exploitation majeur ni d'événement générant une affluence inhabituelle. Pourtant, c'est dans ces situations critiques pour l'opérateur de transport (car sortant du fonctionnement ordinaire de son réseau) qu'un modèle de prévision des flux présente le plus d'intérêt. Afin de valider notre approche, il serait donc pertinent d'analyser ses performances au regard de ce type de scénario.

En théorie, les modèles de mélanges gaussiens sont suffisamment flexibles pour représenter les patterns associés aux situations perturbées. Il convient néanmoins de disposer d'une quantité de données suffisante à leur apprentissage, ce qui peut être difficile en cas de phénomènes rarement observés. Une autre approche consisterait à nous départir de l'hypothèse de stationnarité du système pour permettre à la structure et aux paramètres du réseau bayésien dynamique d'évoluer au cours du temps. En mettant en œuvre des algorithmes d'apprentissage incrémental, le modèle serait capable de prendre en compte l'évolution du système et de s'adapter ainsi aux situations inhabituelles. L'étude de tels algorithmes constitue donc une piste intéressante à explorer dans la suite de nos recherches.

Grâce aux mécanismes de construction génériques que nous avons proposés, notre méthode de prévision des flux à court terme peut être facilement étendue à d'autres lignes et espaces du réseau de transport public, jusqu'à permettre la modélisation d'un vaste système multimodal. La modularité des réseaux bayésiens ouvre la voie à de nombreuses perspectives d'enrichissement du modèle via l'incorporation de nouvelles sources d'information : facteurs calendaires (jour de la semaine, vacances scolaires, etc.), conditions météorologiques, incidents d'exploitation, travaux, événements externes au réseau (manifestations culturelles ou sportives, mouvements sociaux, etc.), capacité des véhicules et bien d'autres. Par ailleurs, une connaissance plus fine des distances entre les flux permettrait à l'expert d'être mieux guidé dans la construction de la structure. Ce dernier serait en effet capable d'établir plus précisément les relations de causalité spatio-temporelles, réduisant le nombre d'arcs candidats sélectionnés et donc la complexité de l'apprentissage.

Les travaux réalisés dans le cadre de cette thèse sont essentiellement prospectifs. À l'heure actuelle, la RATP ne dispose pas de systèmes permettant de mesurer les flux de voyageurs en temps réel, ce qui limite les possibilités d'industrialisation de notre modèle. D'autre part, si les données de validation et de comptage par pesée permettent de mesurer les flux respectivement en périphérie du réseau de transport et à bord des trains, il existe peu d'information sur le comportement des flux piétons

à l'intérieur des stations (notamment au sein des grands pôles multimodaux). Afin d'exploiter pleinement le potentiel de ces travaux, il est donc nécessaire de faire évoluer les systèmes existants pour permettre la collecte de données en temps réel, mais aussi de déployer davantage de dispositifs à l'intérieur des espaces. Au-delà de notre seule démarche de prévision des flux à court terme, la capacité de la RATP à opérer de tels changements constitue un prérequis indispensable au développement d'un système de transport intelligent.

Bibliographie

Ahmed, M. S. et Cook, A. R. (1979). Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques. *Transportation Research Record*, 722:1–9.

Bilmes, J. A. (1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Rapport technique, International Computer Science Institute.

Bureau of Infrastructure, Transport and Regional Economics (2014). New traffic data sources – An overview. New Data Sources for Transport Workshop.

Cansado, A. et Soto, A. (2008). Unsupervised Anomaly Detection in Large Databases Using Bayesian Networks. *Applied Artificial Intelligence*, 22(4):309–330.

Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K. et Han, L. D. (2009). Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert Systems with Applications*, 36(3):6164–6173.

Celikoglu, H. B. (2013). An Approach to Dynamic Classification of Traffic Flow Patterns. *Computer-Aided Civil and Infrastructure Engineering*, 28(4):273–288.

Celikoglu, H. B. et Cigizoglu, H. K. (2007). Public transportation trip flow modeling with generalized regression neural networks. *Advances in Engineering Software*, 38(2):71–79.

Chandra, S. R. et Al-Deek, H. (2009). Predictions of Freeway Traffic Speeds and Volumes Using Vector Autoregressive Models. *Journal of Intelligent Transportation Systems*, 13(2):53–72.

Chen, C., Kwon, J., Rice, J., Skabardonis, A. et Varaiya, P. (2003). Detecting Errors and Imputing Missing Data for Single-Loop Surveillance Systems. *Transportation Research Record*, 1855:160–167.

Chen, C., Wang, Y., Li, L., Hu, J. et Zhang, Z. (2012). The retrieval of intra-day trend and its influence on traffic prediction. *Transportation Research Part C: Emerging Technologies*, 22:103–118.

Chen, K., Xu, L. et Chi, H. (1999). Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*, 12(9):1229–1252.

Chen, Q., Li, W. et Zhao, J. (2011). The use of LS-SVM for short-term passenger flow prediction. *Transport*, 26(1):5–10.

Chickering, D. M. (1996). Learning Bayesian Networks is NP-Complete. *In Learning from Data: Artificial Intelligence and Statistics V*, volume 22, pages 121–130. Springer.

Chow, C. K. et Liu, C. N. (1968). Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory*, 14(3):462–467.

Clark, S. (2003). Traffic Prediction Using Multivariate Nonparametric Regression. *Journal of Transportation Engineering*, 129(2):161–168.

Cooper, G. F. et Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9(4):309–347.

Cortes, C. et Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Davies, S. et Moore, A. (2000). Mix-nets: Factored Mixtures of Gaussians in Bayesian Networks with Mixed Continuous And Discrete Variables. *In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 168–175, Stanford, CA, USA.

Davis, G. A. et Nihan, N. L. (1991). Nonparametric Regression and Short-Term Freeway Traffic Forecasting. *Journal of Transportation Engineering*, 117(2):178–188.

Davis, G. A., Nihan, N. L., Hamed, M. M. et Jacobson, L. N. (1990). Adaptive Forecasting of Freeway Traffic Congestion. *Transportation Research Record*, 1287: 29–33.

Dean, T. et Kanazawa, K. (1989). A Model for Reasoning About Persistence and Causation. *Computational Intelligence*, 5(3):142–150.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Donat, R. (2009). *Modélisation de la fiabilité et de la maintenance par modèles graphiques probabilistes : application à la prévention des ruptures de rail*. Thèse de doctorat, Institut National des Sciences Appliquées de Rouen.

Dougherty, M. S. (1995). A review of neural networks applied to transport. *Transportation Research Part C: Emerging Technologies*, 3(4):247–260.

Dougherty, M. S. et Cobbett, M. R. (1997). Short-term inter-urban traffic forecasts using neural networks. *International Journal of Forecasting*, 13(1):21–31.

Driver, E. et Morrell, D. (1995). Implementation of Continuous Bayesian Networks Using Sums of Weighted Gaussians. *In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 134–140, Montreal, Canada.

Fei, X., Lu, C.-C. et Liu, K. (2011). A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. *Transportation Research Part C: Emerging Technologies*, 19(6):1306–1318.

Figueiredo, M. A. T. et Jain, A. K. (2002). Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396.

Fraley, C. et Raftery, A. E. (2007). Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *Journal of Classification*, 24(2):155–181.

Friedman, N. (1997). Learning Belief Networks in the Presence of Missing Values and Hidden Variables. *In Proceedings of the 14th International Conference on Machine Learning*, pages 125–133, Nashville, TN, USA.

Friedman, N. (1998). The Bayesian Structural EM Algorithm. *In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 129–138, Madison, WI, USA.

Friedman, N., Linial, M., Nachman, I. et Pe'er, D. (2000). Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4):601–620.

Friedman, N., Murphy, K. et Russel, S. (1998). Learning the Structure of Dynamic Probabilistic Networks. *In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 139–147, Madison, WI, USA.

Friedman, N., Nachman, I. et Peér, D. (1999). Learning Bayesian Network Structure from Massive Datasets: The « Sparse Candidate » Algorithm. *In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 206–215, Stockholm, Sweden.

Geiger, D. et Heckerman, D. (1994). Learning Gaussian Networks. *In Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 235–243, Seattle, WA, USA.

Geman, S. et Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Gordon, N. J., Salmond, D. J. et Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F - Radar and Signal Processing*, 140(2):107–113.

Hamed, M. M., Al-Masaeid, H. R. et Bani Said, Z. M. (1995). Short-Term Prediction of Traffic Volume in Urban Arterials. *Journal of Transportation Engineering*, 121(3):249–254.

Haworth, J. (2014). *Spatio-temporal forecasting of network data*. Thèse de doctorat, University College London.

Haworth, J. et Cheng, T. (2012). Non-parametric regression for space-time forecasting under missing data. *Computers, Environment and Urban Systems*, 36(6):538–550.

Heckerman, D., Geiger, D. et Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3):197–243.

Hong, W.-C. (2011). Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm. *Neurocomputing*, 74(12-13):2096–2107.

Huang, R. et Sun, S. (2013). Kernel Regression with Sparse Metric Learning. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 24(4):775–787.

Huang, W., Song, G., Hong, H. et Xie, K. (2014). Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning. *IEEE Transactions on Intelligent Transportation Systems*, 15(5):2191–2201.

Hulst, J. (2006). *Modeling physiological processes with dynamic Bayesian networks*. Mémoire de master, Delft University of Technology.

Isard, M. et Blake, A. (1998). A Smoothing Filter for Condensation. *In Proceedings of the 5th European Conference on Computer Vision*, volume 1, pages 767–781, Freiburg, Germany.

Ishak, S. et Alecsandru, C. (2004). Optimizing Traffic Prediction Performance of Neural Networks under Various Topological, Input, and Traffic Condition Settings. *Journal of Transportation Engineering*, 130(4):452–465.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. et Hinton, G. E. (1991). Adaptative mixtures of local experts. *Neural Computation*, 3(1):79–87.

Jarraya Siala, A. (2013). *Nouvelles paramétrisations de réseaux Bayésiens et leur estimation implicite - Famille exponentielle naturelle et mélange infini de Gaussiennes*. Thèse de doctorat, Université de Nantes - Faculté des Sciences de Sfax.

Jebara, T. et Pentland, A. (1999). Maximum Conditional Likelihood via Bound Maximization and the CEM Algorithm. *In Advances in Neural Information Processing Systems 11*, pages 494–500.

Jiang, X. et Adeli, H. (2005). Dynamic Wavelet Neural Network Model for Traffic Flow Forecasting. *Journal of Transportation Engineering*, 131(10).

Jiao, P., Li, R., Sun, T., Hou, Z. et Ibrahim, A. (2016). Three Revised Kalman Filtering Models for Short-Term Rail Transit Passenger Flow Prediction. *Mathematical Problems in Engineering*, 2016.

Jordan, M. I. et Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45.

Kamarianakis, Y., Kanas, A. et Prastacos, P. (2005). Modeling Traffic Volatility Dynamics in an Urban Network. *Transportation Research Record*, 1923:18–27.

Kamarianakis, Y. et Prastacos, P. (2005). Space–time modeling of traffic flow. *Computers & Geosciences*, 31(2):119–133.

Karlaftis, M. G. et Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399.

Kindermann, R. et Snell, J. L. (1980). *Markov Random Fields and Their Applications*. American Mathematical Society.

Kirby, H. R., Watson, S. M. et Dougherty, M. S. (1997). Should We Use Neural Networks or Statistical Models for Short-Term Motorway Traffic Forecasting? *International Journal of Forecasting*, 13(1):43–50.

Ko, Y., Zhai, C. et Rodriguez-Zas, S. (2009). Inference of gene pathways using mixture Bayesian networks. *BMC Systems Biology*, 3.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1143, Montreal, Canada.

Koller, D. et Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Kruegel, C., Mutz, D., Robertson, W. et Valeur, F. (2003). Bayesian Event Classification for Intrusion Detection. *In Proceedings of the 19th Annual Computer Security Applications Conference*, pages 14–23, Las Vegas, NV, USA.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19(2):191–201.

Leblond, V. et Garcia Castello, F. (2016). GLOBAL et IMPACT, deux outils complémentaires pour la planification des transports. *In Rencontres de la Mobilité Intelligente 2016*, Paris, France.

Lee, S. et Fambro, D. (1999). Application of Subset Autoregressive Integrated Moving Average Model for Short-Term Freeway Traffic Volume Forecasting. *Transportation Research Record*, 1678:179–188.

Leray, P. (2006). *Réseaux bayésiens : Apprentissage et diagnostic de systemes complexes*. Habilitation à diriger les recherches, Université de Rouen.

Leray, P. et François, O. (2004). Réseaux Bayésiens pour la Classification - Méthodologie et Illustration dans le cadre du Diagnostic Médical. *Revue d'Intelligence Artificielle*, 18(2):169–193.

Leray, P. et François, O. (2005). Bayesian Network Structural Learning and Incomplete Data. *In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 33–40, Espoo, Finland.

Levin, M. et Tsao, Y.-D. (1980). On Forecasting Freeway Occupancies and Volumes (Abridgment). *Transportation Research Record*, 773:47–49.

Li, L., Li, Y. et Li, Z. (2013a). Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies*, 34:108–120.

Li, Q., Qin, Y., Wang, Z., Zhao, Z., Zhan, M., Liu, Y. et Li, Z. (2013b). The Research of Urban Rail Transit Sectional Passenger Flow Prediction Method. *Journal of Intelligent Learning Systems and Applications*, 5(4):227–231.

Little, R. J. A. et Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Liu, H. (2012). *Bayesian Networks and Gaussian Mixture Models in Multi-Dimensional Data Analysis with Application to Religion-Conflict Data*. Mémoire de master, Arizona State University.

Lucas, P. J. F., van der Gaag, L. C. et Abu-Hanna, A. (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30(3):201–214.

Lv, Y., Duan, Y., Kang, W., Li, Z. et Wang, F.-Y. (2015). Traffic Flow Prediction With Big Data: A Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873.

Ma, Z., Xing, J., Mesbah, M. et Ferreira, L. (2014). Predicting short-term bus passenger demand using a pattern hybrid approach. *Transportation Research Part C: Emerging Technologies*, 39:148–163.

Min, W. et Wynter, L. (2011). Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, 19(4):606–616.

Moore, A. (1999). Very Fast EM-based Mixture Model Clustering using Multi-resolution kd-trees. *In Advances in Neural Information Processing Systems 11*, pages 543–549.

Murphy, K. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. Thèse de doctorat, University of California.

Naïm, P., Wuillemin, P.-H., Leray, P., Pourret, O. et Anna, B. (2011). *Réseaux bayésiens*. Eyrolles, troisième édition.

Neal, R. M. et Hinton, G. E. (1998). A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. *In Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers.

Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Prentice Hall.

Ng, S.-K. et McLachlan, G. J. (2004). Using the EM Algorithm to Train Neural Networks: Misconceptions and a New algorithm for Multiclass Classification. *IEEE Transactions on Neural Networks*, 15(3):738–749.

Ni, D., Leonard II, J. D., Guin, A. et Feng, C. (2005). A Multiple Imputation Scheme for Overcoming the Missing Values and Variability Issues in ITS Data. *Journal of Transportation Engineering*, 131:931–938.

Ni, M., He, Q. et Gao, J. (2017). Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media. *IEEE Transactions on Intelligent Transportation Systems*, 18(6):1623–1632.

Okutani, I. et Stephanedes, Y. J. (1984). Dynamic prediction of traffic volume through Kalman filtering theory. *Transportation Research Part B: Methodological*, 18(1):1–11.

Ormoneit, D. et Tresp, V. (1996). Improved Gaussian Mixture Density Estimates Using Bayesian Penalty Terms and Network Averaging. *In Advances in Neural Information Processing Systems 8*, pages 542–548.

Ortúzar, J. d. D. et Willumsen, L. G. (2011). *Modelling Transport*. John Wiley & Sons, quatrième édition.

Pan, P. et Schonfeld, D. (2011). Visual Tracking Using High-Order Particle Filtering. *IEEE Signal Processing Letters*, 18(1):51–54.

Park, B., Messer, C. J. et Urbanik II, T. (1998). Short-Term Freeway Traffic Volume Forecasting Using Radial Basis Function Neural Network. *Transportation Research Record*, 1651:39–47.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers.

Pearl, J. et Verma, T. S. (1991). A Theory of Inferred Causation. *In Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 441–452, Cambridge, MA, USA.

Pourret, O., Naïm, P. et Marcot, B. (2008). *Bayesian Networks: A Practical Guide to Applications*. John Wiley & Sons.

Qu, L., Li, L., Zhang, Y. et Hu, J. (2009). PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):512–522.

Queen, C. M. et Albers, C. J. (2009). Intervention and causality: forecasting traffic flows using a dynamic Bayesian network. *Journal of the American Statistical Association*, 104(486):669–681.

Queen, C. M. et Smith, J. Q. (1993). Multiregression Dynamic Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):849–870.

Quek, C., Pasquier, M. et Lim, B. B. S. (2006). POP-TRAFFIC: A Novel Fuzzy Neural Approach to Road Traffic Analysis and Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 7(2):133–146.

Robinson, J. W. et Hartemink, A. J. (2010). Learning Non-Stationary Dynamic Bayesian Networks. *Journal of Machine Learning Research*, 11:3647–3680.

Roos, J., Bonnevay, S. et Gavin, G. (2016). Short-Term Urban Rail Passenger Flow Forecasting: A Dynamic Bayesian Network Approach. *In Proceedings of the 15th IEEE International Conference on Machine Learning and Applications*, pages 1034–1039, Anaheim, CA, USA.

Roos, J., Bonnevay, S. et Gavin, G. (2017a). Dynamic Bayesian Networks with Gaussian Mixture Models for Short-Term Passenger Flow Forecasting. *In Proceedings of the 12th International Conference on Intelligent Systems and Knowledge Engineering*, Nanjing, China.

Roos, J., Bonnevay, S. et Gavin, G. (2017b). Prévision à court terme des flux de voyageurs du réseau ferré urbain : une approche par les réseaux bayésiens dynamiques. *In Actes de la 17ème conférence Extraction et Gestion des Connaissances*, pages 315–320, Grenoble, France.

Roos, J., Gavin, G. et Bonnevay, S. (2017c). A dynamic bayesian network approach to forecast short-term urban rail passenger flows with incomplete data. *Transportation Research Procedia*, 26:53–61.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Salojärvi, J., Puolamäki, K. et Kaski, S. (2005). Expectation Maximization Algorithms for Conditional Likelihoods. *In Proceedings of the 22nd International Conference on Machine Learning*, pages 752–759, Bonn, Germany.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Shachter, R. D. et Kenley, R. C. (1989). Gaussian influence diagrams. *Management Science*, 35(5):527–550.

Smith, B. L. et Demetsky, M. J. (1994). Short-term traffic flow prediction models - a comparison of neural network and nonparametric regression approaches. *In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1706–1709, San Antonio, TX, USA.

Smith, B. L. et Demetsky, M. J. (1997). Traffic Flow Forecasting: Comparison of Modeling Approaches. *Journal of Transportation Engineering*, 123(4):261–266.

Smith, B. L., Williams, B. M. et Oswald, R. K. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 10(4):303–321.

Song, L., Kolar, M. et Xing, E. P. (2009). Time-Varying Dynamic Bayesian Networks. *In Advances in Neural Information Processing Systems 22*, pages 1732–1740.

Soto, A., Zavala, F. et Araneda, A. (2007). An Accelerated Algorithm for Density Estimation in Large Databases Using Gaussian Mixtures. *Cybernetics and Systems*, 38(2):123–139.

Spiegelhalter, D. J. et Lauritzen, S. L. (1990). Sequential Updating of Conditional Probabilities on Directed Graphical Structures. *Networks*, 20(5):579–605.

Spirtes, P., Glymour, C. N. et Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag.

Stathopoulos, A. et Karlaftis, M. G. (2003). A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C: Emerging Technologies*, 11(2):121–135.

Štepánová, K. et Vavrečka, M. (2016). Estimating number of components in Gaussian mixture model using combination of greedy and merging algorithm. *Pattern Analysis and Applications*.

Sun, S. et Chen, Q. (2008). Kernel Regression with a Mahalanobis Metric for Short-Term Traffic Flow Forecasting. *In Proceedings of the 9th International Conference on Intelligent Data Engineering and Automated Learning*, pages 9–16, Daejeon, South Korea.

Sun, S., Huang, R. et Gao, Y. (2012). Network-Scale Traffic Modeling and Forecasting with Graphical Lasso and Neural Networks. *Journal of Transportation Engineering*, 138(11):1358–1367.

Sun, S., Zhang, C. et Yu, G. (2006). A Bayesian Network Approach to Traffic Flow Forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):124–132.

Sun, S., Zhang, C. et Zhang, Y. (2005). Traffic Flow Forecasting Using a Spatio-temporal Bayesian Network Predictor. *In Proceedings of the International Conference on Artificial Neural Networks*, pages 273–278, Warsaw, Poland.

Sun, Y., Leng, B. et Guan, W. (2015). A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. *Neurocomputing*, 166:109–121.

Sun, Y., Zhang, G. et Yin, H. (2014). Passenger Flow Prediction of Subway Transfer Stations Based on Nonparametric Regression Model. *Discrete Dynamics in Nature and Society*, 2014.

Thiesson, B., Meek, C. et Heckerman, D. (2001). Accelerating EM for Large Databases. *Machine Learning*, 45(3):279–299.

Toqué, F., Côme, E., El Mahrsi, M. K. et Oukhellou, L. (2016). Forecasting Dynamic Public Transport Origin-Destination Matrices with Long-Short Term Memory Recurrent Neural Networks. *In Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems*, pages 1071–1076, Rio de Janeiro, Brazil.

Toqué, F., Khouadjia, M., Côme, E., Trépanier, M. et Oukhellou, L. (2017). Short & Long Term Forecasting of Multimodal Transport Passenger Flows with Machine Learning Methods. *In Proceedings of the IEEE 20th International Conference on Intelligent Transportation Systems*, pages 560–566, Yokohama, Japan.

Trabelsi, G. (2013). *New structure learning algorithms and evaluation methods for large dynamic Bayesian networks*. Thèse de doctorat, Université de Nantes.

Tsamardinos, I., Brown, L. E. et Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1): 31–78.

Ueda, N., Nakano, R., Ghahramani, Z. et Hinton, G. E. (2000). SMEM Algorithm for Mixture Models. *Neural Computation*, 12(9):2109–2128.

van Hinsbergen, C. P. I. J., van Lint, J. W. C. et Sanders, F. M. (2007). Short Term Traffic Prediction Models. *In Proceedings of the 14th World Congress on Intelligent Transport Systems*, pages 5013–5030, Beijing, China.

van Lint, J. W. C. (2008). Online Learning Solutions for Freeway Travel Time Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 9(1):38–47.

van Lint, J. W. C., Hoogendoorn, S. P. et van Zuylen, H. J. (2005). Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C: Emerging Technologies*, 13(5-6):347–369.

van Lint, J. W. C. et van Hinsbergen, C. P. I. J. (2012). Short-Term Traffic and Travel Time Prediction Models. *Artificial Intelligence Applications to Critical Transportation Issues*, 22:22–41.

Verbeek, J. J., Vlassis, N. et Kröse, B. (2003). Efficient Greedy Learning of Gaussian Mixture Models. *Neural Computation*, 15(2):469–485.

Verma, T. S. et Pearl, J. (1988). Causal Networks: Semantics and Expressiveness. *In Proceedings of the 4th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, Minneapolis, MN, USA.

Vlahogianni, E. I., Golias, J. C. et Karlaftis, M. G. (2004). Short-term Traffic Forecasting: Overview of Objectives and Methods. *Transport Reviews*, 24(5):533–557.

Vlahogianni, E. I., Karlaftis, M. G. et Golias, J. C. (2005). Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. *Transportation Research Part C: Emerging Technologies*, 13(3):211–234.

Vlahogianni, E. I., Karlaftis, M. G. et Golias, J. C. (2007). Spatio-Temporal Short-Term Urban Traffic Volume Forecasting Using Genetically Optimized Modular Networks. *Computer-Aided Civil and Infrastructure Engineering*, 22(5):317–325.

Vlahogianni, E. I., Karlaftis, M. G. et Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43(1):3–19.

Vlassis, N. et Likas, A. (2002). A Greedy EM Algorithm for Gaussian Mixture Learning. *Neural Processing Letters*, 15(1):77–87.

Wang, Y. et Papageorgiou, M. (2005). Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transportation Research Part B: Methodological*, 39(2):141–167.

Weber, P., Medina-Oliva, G., Simon, C. et Iung, B. (2012). Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Engineering Applications of Artificial Intelligence*, 25(4):671–682.

Wei, Y. et Chen, M.-C. (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies*, 21(1):148–162.

Wermuth, N. (1980). Linear Recursive Equations, Covariance Selection, and Path Analysis. *Journal of the American Statistical Association*, 75(372):963–972.

Whitlock, M. E. et Queen, C. M. (2000). Modelling a Traffic Network with Missing Data. *Journal of Forecasting*, 19(7):561–574.

Whittaker, J., Garside, S. et Lindveld, K. (1997). Tracking and predicting a network traffic process. *International Journal of Forecasting*, 13(1):51–61.

Wickham, H. et Grolemund, G. (2016). *R for Data Science*. O'Reilly Media.

Williams, B. M. (2001). Multivariate Vehicular Traffic Flow Prediction. *Transportation Research Record*, 1776:194–200.

Williams, B. M., Durvasula, P. K. et Brown, D. E. (1998). Urban Freeway Traffic Flow Prediction Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models. *Transportation Research Record*, 1644:132–141.

Williams, B. M. et Hoel, L. A. (2003). Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results. *Journal of Transportation Engineering*, 129(6):664–672.

Xu, L. et Jordan, M. I. (1996). On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation*, 8(1):129–151.

Xu, L., Jordan, M. I. et Hinton, G. E. (1995). An Alternative Model for Mixtures of Experts. *In Advances in Neural Information Processing Systems 7*, pages 633–640.

Xue, R., Sun, D. J. et Chen, S. (2015). Short-Term Bus Passenger Demand Prediction Based on Time Series Model and Interactive Multiple Model Approach. *Discrete Dynamics in Nature and Society*, 2015.

Yin, H., Wong, S. C., Xu, J. et Wong, C. K. (2002). Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 10(2):85–98.

Yu, G., Hu, J., Zhang, C., Zhuang, L. et Song, J. (2003). Short-term Traffic Flow Forecasting Based on Markov Chain Model. *In Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 208–212, Columbus, OH, USA.

Zhang, B., Zhang, C. et Yi, X. (2004). Competitive EM algorithm for finite mixture models. *Pattern Recognition*, 37(1):131–144.

Zhang, G., Patuwo, B. E. et Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1):35–62.

Zhang, X., Liu, Q., Yang, W., Wei, N. et Dong, D. (2013). Wavelet Neural Network-based Short-Term Passenger Flow Forecasting on Urban Rail Transit. *Telkomnika*, 11(12):7379–7385.

Zhang, Y. et Liu, Y. (2009). Data Imputation Using Least Squares Support Vector Machines in Urban Arterial Streets. *IEEE Signal Processing Letters*, 16(5):414–417.

Zhang, Y. et Xie, Y. (2008). Forecasting of Short-Term Freeway Volume with v-Support Vector Machines. *Transportation Research Record*, 2024:92–99.

Zhang, Z., Chen, C., Sun, J. et Chan, K. L. (2003). EM algorithms for Gaussian mixtures with split-and-merge operation. *Pattern Recognition*, 36(9):1973–1983.

Zheng, F. et van Zuylen, H. J. (2013). Urban link travel time estimation based on sparse probe vehicle data. *Transportation Research Part C: Emerging Technologies*, 31:145–157.

Zheng, W., Lee, D.-H. et Shi, Q. (2006). Short-Term Freeway Traffic Flow Prediction: Bayesian Combined Neural Network Approach. *Journal of Transportation Engineering*, 132(2):114–121.

Zhong, M., Lingras, P. et Sharma, S. (2004). Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. *Transportation Research Part C: Emerging Technologies*, 12(2):139–166.

Zhu, Z., Peng, B., Xiong, C. et Zhang, L. (2016). Short-term traffic flow prediction with linear conditional Gaussian Bayesian network. *Journal of Advanced Transportation*, 50(6):1111–1123.