Analyse de réseaux temporels par des méthodes de traitement du signal : application au système de vélos en libre-service à Lyon

Ronan Hamon

ENS DE LYON

ED 52 - PHAST
Physique & Astrophysique de Lyon
Ecole doctorale

UNIVERSITÉ DE LYON

THÈSE

en vue de l'obtention du grade de :

Docteur de l'Université de Lyon,
délivré par l'École Normale Supérieure de Lyon

Discipline : **Physique**

Laboratoire de Physique

École Doctorale de Physique et d'Astrophysique de Lyon

présentée et soutenue publiquement le 30 septembre 2015 par :

Ronan HAMON

Analyse de réseaux temporels par des méthodes de traitement du signal : Application au système de vélos en libre-service à Lyon

Co-directeur de thèse : **Patrick FLANDRIN**
Co-directrice de thèse : **Céline ROBARDET**
Encadrant de thèse : **Pierre BORGNAT**

après avis de :

Mme Clémence MAGNIEN	Directrice de recherche, Université Pierre et Marie Curie	Rapporteuse
M. Cédric RICHARD	Professeur, Université de Nice Sophia-Antipolis	Rapporteur

Devant la commission d'examen formée de :

M. Pierre BORGNAT	Chargé de recherche, ENS de Lyon	Encadrant
Mme Maureen CLERC	Directrice de recherche, INRIA Sophia Antipolis	Examinatrice
M. Patrick FLANDRIN	Directeur de recherche, ENS de Lyon	Directeur
M. Michel HABIB	Professeur, Université Paris Diderot	Examinateur
Mme Clémence MAGNIEN	Directrice de recherche, Université Pierre et Marie Curie	Rapporteuse
M. Cédric RICHARD	Professeur, Université de Nice Sophia-Antipolis	Rapporteur
Mme Céline ROBARDET	Professeur, INSA Lyon	Directrice
Mme Marie VOGEL	Maître de conférences, ENS de Lyon	Examinatrice

« Je laisse Sisyphe au bas de la montagne ! On retrouve toujours son fardeau. Mais Sisyphe enseigne la fidélité supérieure qui nie les dieux et soulève les rochers. Lui aussi juge que tout est bien. Cet univers désormais sans maître ne lui paraît ni stérile ni futile. Chacun des grains de cette pierre, chaque éclat minéral de cette montagne pleine de nuit, à lui seul, forme un monde. La lutte elle-même vers les sommets suffit à remplir un cœur d'homme. Il faut imaginer Sisyphe heureux. »

– Albert Camus, *Le Mythe de Sisyphe*

Remerciements

Je tiens à exprimer mes remerciements les plus sincères à l'ensemble des personnes qui m'ont accompagné pendant cette thèse. Les rencontres que j'ai pu faire tout au long de ces trois années ont été précieuses, et m'ont permis de réaliser cette thèse dans des conditions de travail excellentes, notamment au sein du laboratoire de physique de l'ENS Lyon.

Je remercie tout particulièrement Pierre, Patrick et Céline. Leurs qualités scientifiques et humaines ont été très importantes pour la réalisation de ce travail et mon épanouissement scientifique.

Un grand merci.

Sommaire

Introduction

Les systèmes de vélos en libre-service sont devenus des éléments indispensables dans les offres de transport urbain des grandes villes mondiales. Comme tout système informatisé, ils génèrent des données volumineuses et complexes, dont l'utilisation est essentiellement limitée à la gestion et à l'exploitation du système. Les mouvements effectués par les usagers du système peuvent pourtant fournir des informations précieuses sur de nombreux aspects de la vie urbaine, par exemple sur la dynamique temporelle et spatiale des déplacements dans la ville, sur la place du vélo parmi les autres modes de transport, ou encore sur la répartition des inégalités territoriales et sociales dans l'espace géographique. Les travaux présentés dans ce manuscrit de thèse s'inscrivent dans ce contexte d'avalanche de données numériques, en proposant une méthode d'analyse innovante adaptée à l'étude des systèmes de vélos en libre-service. Avant d'entrer plus en détails sur ces travaux, il convient de s'intéresser à ce mouvement global d'extraction de connaissance dans des corpus de données numériques, connu sous le nom de *Big Data*, ou datamasse en français [72] en référence à la quantité de données à disposition.

La datamasse

Les récentes avancées technologiques ont bouleversé les rapports que nous entretenons avec le monde numérique. Le développement de l'Internet haut-débit ainsi que des réseaux mobiles permet désormais de rester connecté en permanence aux services en ligne (messagerie, réseau social, etc.). Des pans entiers de notre vie deviennent ainsi exclusivement numériques, c'est-à-dire représentés grâce à des données informatiques, et cette tendance est vouée à s'accroître au fil des ans, avec l'arrivée des objets connectés, notamment ceux permettant de relever et contrôler de nombreux paramètres biologiques tels que le rythme cardiaque, la température corporelle ou le nombre de calories dépensées.

La numérisation ne concerne pas uniquement les données personnelles : dans les entreprises, la collecte des données est devenue systématique grâce au faible coût de leur stockage. Dans la majorité des cas, cette collecte se fait sans objectif précis, sinon celui d'espérer, d'une manière ou d'une autre, pouvoir un jour valoriser ces données.

La datamasse peut se définir suivant trois caractéristiques, connues sous le nom des 3 V [85] :

— Volume, pour désigner la quantité de données à disposition ;
— Vitesse, pour décrire la fréquence à laquelle ces données sont générées ;
— Variété, pour caractériser les différents types de données qui existent.

Castellucia et al. [48] comparent les données aux matières premières, comme le fer ou le pétrole, qui ont constitué les ressources primaires des industries du 20^{e} siècle : de la même manière que le pétrole nécessite des recherches pour la prospection, l'extraction et l'exploitation, les données doivent être collectées, analysées et interprétées pour acquérir de la connaissance, quelle qu'en soit la fin. Le processus d'extraction de connaissances dans un corpus de données nécessite ainsi plusieurs étapes, chacune faisant appel à de nombreuses et diverses compétences.

Il est difficile de saisir la quantité phénoménale de données produites chaque seconde. Le recours à des infographies, comme le propose le site *The Internet in Real-Time* [1] permet, à travers la mise à jour en temps réel de compteurs de données produites par les principaux services Internet, de mesurer cette avalanche de données numériques. Ainsi, en l'espace d'une seconde, ce sont près de 220 000 messages qui sont envoyés via l'application Whatsapp, 23 000 minutes de vidéos échangées sur Skype, ou encore 52 000 *Likes* [2] sur Facebook. Toutes ces données sont à la fois diverses, complexes, et extrêmement nombreuses.

Le principal risque auquel on s'expose en cherchant à exploiter des données massives est de penser qu'il suffit d'appliquer des méthodes statistiques sur les données pour automatiquement obtenir de la connaissance. La datamasse permet seulement de trouver des corrélations, c'est-à-dire des liens statistiques entre différents phénomènes. À travers ces corrélations, il est alors possible de simplifier la réalité à travers un modèle, et d'utiliser cette simplification pour inférer des éléments de connaissance ou d'innovation. Cette étape d'inférence est difficilement réalisable par un algorithme car elle nécessite de la créativité et donc une action humaine, néanmoins la datamasse rend possible, par sa capacité à collecter, traiter et classer les données, ce processus. Un deuxième risque de la datamasse provient de l'automatisation de certaines tâches, et l'implicite confiance que l'on accorde aux algorithmes sans remettre en cause les résultats obtenus « mathématiquement ». Les nouveaux services basés sur la datamasse laissent de plus en plus d'importance aux algorithmes, et il est nécessaire de ne pas se laisser enfermer dans le modèle de pensée défini par l'algorithme (et donc de son concepteur) [40].

Le potentiel économique de ces données n'est plus à démontrer, devant les exemples de réussite de la nouvelle industrie qui a vu le jour ces dernières années, et dont le modèle économique repose sur leur traitement. Des entreprises comme Google, Facebook ou Twitter, pour ne citer que les plus connues, génèrent des profits records, tout en continuant à proposer leurs services gratuitement. La clé de leur réussite réside dans leur capacité à transformer les données personnelles de leurs utilisateurs, obtenues à travers l'utilisation du service, en publicité dont le contenu cible les intérêts, la personnalité et les attentes de chaque personne.

La datamasse n'est cependant pas uniquement dédiée à fournir de la publicité : l'affaire Snowden [67], du nom d'un ancien analyste de l'agence de renseignement de sécurité intérieure américaine (*National Security Agency* ou NSA) qui a divulgé des documents confidentiels sur les pratiques de l'agence, a montré que l'analyse de données massives était utilisée dans des proportions qu'il est difficile d'imaginer : le programme PRISM permet à la NSA d'accéder par exemple aux données issues du réseau social Facebook, aux conversations réalisées avec le logiciel de communication Skype ou aux services de Google (qui incluent entre autres un moteur de recherche et une messagerie) [53] : le volume de données a été estimé à 850 milliards d'enregistrements [102]. Si l'indignation qui a suivi ces révélations a apporté de nouvelles régulations, dont le récent *Freedom Act* voté aux États-Unis en juin 2015, censé limiter la surveillance de la population par la NSA [260], la récente loi sur le renseignement en cours de discussions en France, qui suscite la colère des associations de défense de droits numériques et civiques [177] à cause de l'introduction de « boîtes noires algorithmiques », montre que l'analyse de données a toujours le vent en poupe dans le domaine de la sécurité. Cette approche a néanmoins ses limites : en laissant de côté le débat sur le bien-fondé ou non d'une surveillance de masse [149], l'efficacité d'une telle procédure pour améliorer la traque d'individus dangereux est remise en cause par les spécialistes de l'analyse de données [128]. Dans le domaine proche de la criminologie, les limites des logiciels de prédiction d'actes de délinquance, notamment Predpol qui est utilisé dans quelques villes américaines dont Los Angeles et Atlanta, ont également été soulignées, tant du point de vue sociologique [24] que technique [116].

Le monde de la recherche commence également à s'emparer du phénomène de la datamasse [142, 131], même si paradoxalement, les succès sont moins visibles. On peut tout au plus noter les progrès

1. `http://pennystocks.la/internet-in-real-time/`
2. Une option proposée par Facebook pour exprimer, sans commentaire, un avis positif sur un contenu posté par un contact.

réalisés en épidémiologie [189], en citant par exemple la capacité de Google Flu Trends [113] à prévoir de manière très précise [62] la propagation du virus de la grippe, à travers les requêtes des utilisateurs sur le moteur de recherche. La communauté scientifique est plutôt concentrée sur le développement d'outils liés à la datamasse plutôt qu'à leur utilisation. Les travaux présentés dans ce manuscrit participent à ce mouvement : ils cherchent à proposer de nouvelles méthodes d'analyse des données, sans se soucier ni des aspects techniques liés à la collecte, au stockage ou au nettoyage de ces données, ni des aspects sociétaux et philosophiques que la datamasse implique. Il est néanmoins important d'avoir conscience de ces problèmes.

Le corpus de données à notre disposition est celui issu du système de vélos en libre-service à Lyon, et est constitué de tous les déplacements réalisés pendant l'année 2011. Son volume relativement faible, de l'ordre du million d'entrées, ne permettent pas de considérer ces travaux comme faisant partie du champ de la datamasse. Néanmoins, les travaux présentés dans cette thèse n'en sont pas éloignés, d'une part du fait de la diversité des données étudiées, et d'autre part car ils cherchent à étudier la dynamique de ces données. De plus, il paraît essentiel, si l'on souhaite étudier des problèmes de plus grande envergure, d'avoir des éléments de compréhension sur des jeux de données de petite taille. Ces données permettent ainsi de regarder les villes sous un angle quantitatif, et s'intègrent ainsi dans le concept des villes intelligentes.

Villes intelligentes

Les villes intelligentes (*smart cities* en anglais) sont un concept développé ces dernières années qui vise à proposer des solutions pour améliorer la gestion des villes, notamment grâce à l'utilisation des nouvelles technologies. L'espace urbain est vu comme une superposition complexe d'infrastructures regroupant les services municipaux comme les transports, la fourniture d'eau, d'énergie ou de systèmes de communications. Au-dessus de ces infrastructures, une nouvelle couche informationnelle est mise en place, permettant d'obtenir, à l'aide de capteurs, des informations en temps réel sur de nombreux aspects des systèmes de manière à optimiser la maintenance, informer les citoyens et, dans un deuxième temps, comprendre le fonctionnement de ces systèmes afin de les améliorer. Comme l'explique Jérémy Rifkin [202], les *smart grids*, ou réseaux d'énergie intelligents, proposent un nouveau paradigme dans la gestion de l'énergie, en changeant une approche dans laquelle une entreprise produit et fournit l'énergie aux consommateurs, pour passer à un autre modèle dans lequel chaque foyer produit de l'énergie et la distribue dans le réseau. Le concept de *Smart Cities* étend ce paradigme à l'échelle de la ville, en offrant aux habitants les moyens de devenir acteurs de la gestion de la ville. De par le volume conséquent des données générées, leur variété et leur vitesse d'acquisition, les questions reliées aux villes intelligentes s'inscrivent dans la datamasse.

Les premiers projets de villes intelligentes restent encore évasifs sur leurs finalités, en se contentant d'idées générales. La Figure 0.1 propose un exemple de ville intelligente, selon Bouygues Télécom Innovations [170]. Les capteurs disséminés dans les véhicules, les mobiliers urbains et sur les bâtiments enregistrent une quantité phénoménale d'informations, sans pour autant que l'objectif de cette collecte soit explicité. Une implémentation concrète des villes intelligentes a lieu dans l'agglomération du Grand Lyon, qui regroupe les communes de Lyon et de sa périphérie. La vision de la ville intelligente est décomposée en quatre idées [155] :

1. la prise en compte des enjeux environnementaux et contraintes énergétiques ;
2. le fonctionnement en réseau des acteurs entre eux ;
3. le passage de la propriété à l'usage ;
4. l'intégration des nouvelles technologies.

Ces principes, aux contours flous, se traduisent par la mise en place de plusieurs projets, comme la mise à disposition des données en accès libre (*Open Data*), des initiatives liées à l'amélioration du trafic (par

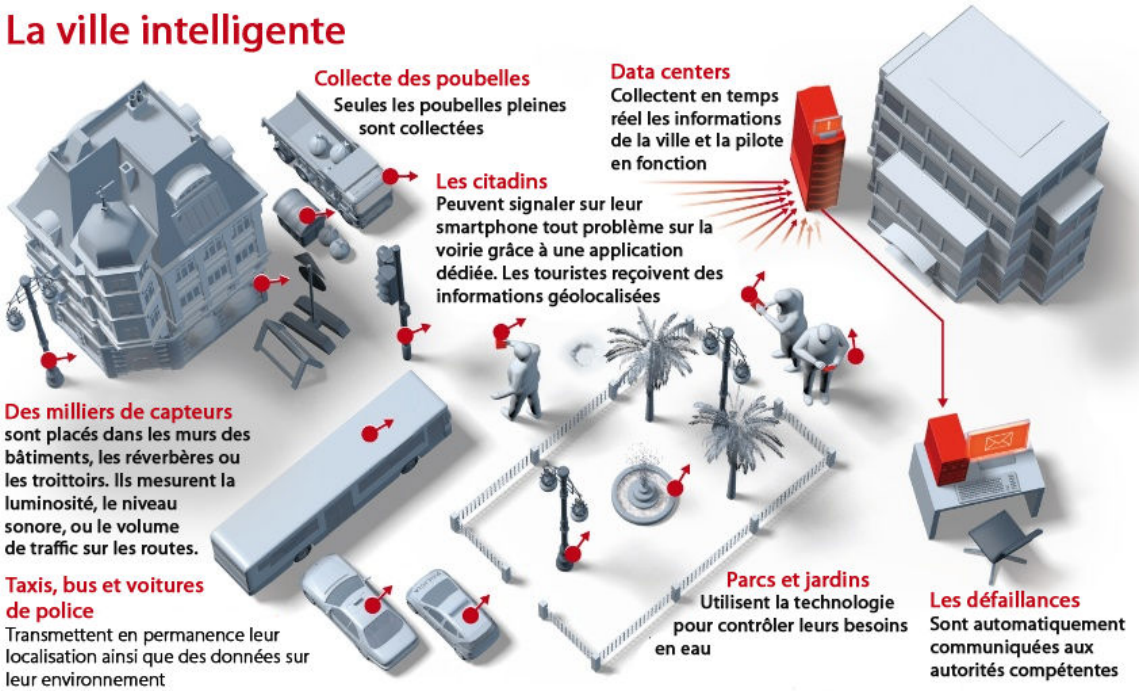


FIGURE 0.1 – Exemple de ville intelligente, selon Bouygues Télécom Innovations [170]

exemple Optimode, qui permet de prévoir son temps de parcours, en prenant en compte tous les modes de transport) ou la gestion de l'énergie (la mise en place de compteurs d'électricité intelligents). La quantité de données recueillies ainsi que la diversité des problèmes auxquels les villes sont confrontées laissent entrevoir des projets à portée beaucoup plus large, comme le montre le développement des recherches sur le sujet, par exemple à Lyon avec le dispositif de recherche et d'expérimentation sur la ville « Intelligences des Mondes Urbains » [3].

Ces recherches sur les villes intelligentes nécessitent un ensemble de compétences très vastes étant donné la diversité des problèmes rencontrés dans les villes. Il requiert également la mise au point de techniques d'extraction de connaissances adaptées aux données urbaines. Ces données se trouvent très souvent sous formes de réseaux, qu'ils soient de transports, d'énergie ou de communication. La Figure 0.2 affiche par exemple un plan des lignes fortes du réseau de transport lyonnais, laissant apparaître la structure en réseau : chaque station représente un nœud du graphe, et deux stations sont connectées si elles sont reliées soit par une ligne de bus, soit par une ligne de métro. Les capteurs forment également des réseaux, et peuvent s'échanger des informations en fonction de leur distance afin de mieux contrôler par exemple la pollution urbaine [131]. Cette structure particulière, que l'on retrouve dans de nombreux systèmes, est l'objet d'une nouvelle théorie en plein développement ces dernières années : la science des réseaux complexes.

Étude des réseaux complexes

La notion de réseau, qui est au cœur des travaux présentés dans cette thèse, consiste à considérer un ensemble d'entités, reliées entre elles par des relations de proximité. Dans le cas de la Figure 0.2, les entités représentaient les stations. Un autre exemple est celui des collaborations scientifiques : la Figure 0.3 représente un réseau, dans lequel les nœuds sont les chercheurs, et un lien existe entre deux chercheurs s'ils ont publié au moins un article ensemble. La localisation des chercheurs sur une carte permet de révéler d'une part où se concentrent les activités de recherche, et d'autre part quelles régions du monde entretiennent des collaborations de recherche. La liste des exemples de réseaux est longue,

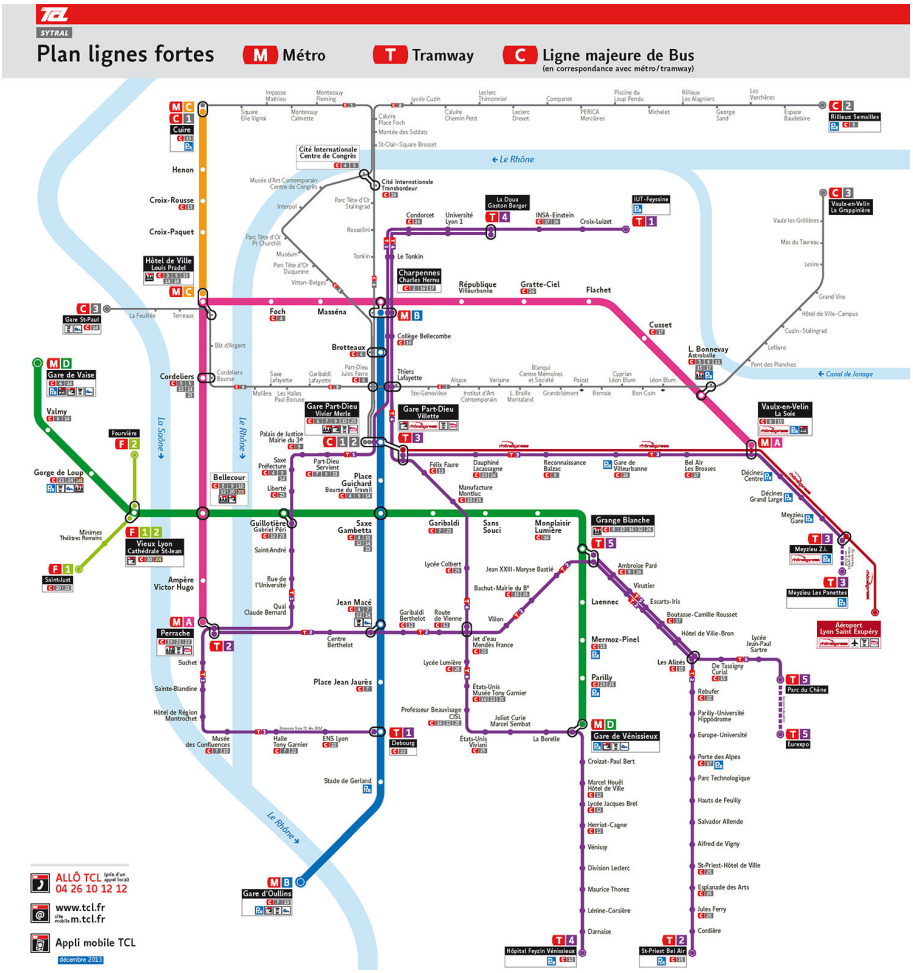


FIGURE 0.2 – Lignes fortes du réseau de transport lyonnais TCL [225]

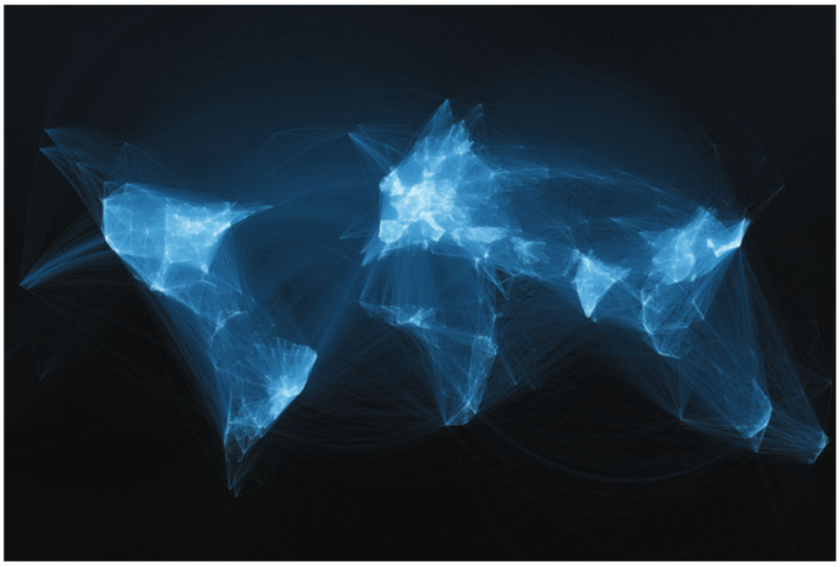


FIGURE 0.3 – Réseau des collaborations scientifiques (Extrait de [7])

et inclue des réseaux présents dans notre vie, comme les réseaux sociaux (Facebook, Linkedin, etc.), le réseau électrique ou encore Internet.

L'étude des réseaux complexes met en jeu de nombreuses disciplines des sciences exactes : la théorie des graphes permet de modéliser les réseaux à l'aide d'objets mathématiques, les graphes. L'informatique permet de mettre en place des algorithmes efficaces pour traiter des données en grande quantité. La physique apporte des concepts de méthode innovants, facilement adaptables aux réseaux. Enfin, l'analyse de réseaux complexes n'est pas possible sans considérer la discipline dans laquelle les données sont ancrées, et qui peut aller de la biologie à la psychologie, en passant par le transport, la sociologie ou encore l'histoire.

Une démarche similaire est adoptée dans ces travaux : les systèmes de vélos en libre-service peuvent également se représenter sous la forme d'un réseau, en considérant les entités comme les stations et les liens comme des flux de vélos. Ces données relationnelles peuvent tout d'abord être considérées soit comme des données tabulaires statiques, soit comme des données dynamiques agrégées en séries temporelles, permettant l'utilisation des techniques usuelles en fouille de données telles que la classification ou la régression linéaire. Néanmoins, si l'on veut pleinement exploiter les dimensions spatiale et temporelle des données, il semble essentiel de considérer à la fois la structure en réseau et la dynamique des relations, amenant ainsi à étudier des réseaux temporels, c'est-à-dire des réseaux dont la structure évolue au cours du temps.

Le manque d'outils adaptés pour étudier ces objets appelle à de nouvelles méthodes, dont le développement est la motivation des travaux présentés dans cette thèse. La méthode proposée se base sur le parallèle entre l'analyse de réseaux temporels et de signaux temporels, objets d'étude de la discipline du traitement du signal. Les similitudes dans la description de l'évolution d'un réseau temporel et celle d'un signal permettent de considérer une dualité entre ces deux types d'objets. Ainsi, l'analyse dans le domaine des signaux, à l'aide d'outils de traitement du signal bien établis, permet de caractériser le réseau temporel correspondant. Ce changement de représentation nécessite plusieurs étapes, utilisant des outils de théorie des graphes, de statistique et de traitement du signal. De plus, l'interprétation des résultats sur le réseau obtenu à partir des données sur les systèmes de vélos en libre-service requiert également des compétences en sociologie, en transport et en géographie. Cette multidisciplinarité témoigne des compétences nécessaires à l'étude des réseaux complexes.

Plan

Le manuscrit est divisé en quatre chapitres. Le Chapitre 1 présente le système de vélos en libre-service lyonnais, et expose les différents travaux entrepris à travers des méthodes d'analyse de données classiques. Il justifie également le passage vers un réseau temporel, et appelle ainsi au développement de nouvelles méthodologies pour son analyse. Le Chapitre 2 introduit les réseaux, sans prendre en compte la dimension temporelle. Afin de transformer des réseaux, à la dimension élevée, en des signaux unidimensionnels, une indexation adéquate des entités du réseau est nécessaire. Ce chapitre répond à ce problème en proposant, à l'aide d'outils de la théorie des graphes, un algorithme d'étiquetage des nœuds du réseau. Le Chapitre 3 introduit la dualité entre réseaux et signaux, en étudiant les transformations d'une représentation à une autre, tant sur des aspects théoriques qu'empiriques. Il met ainsi en exergue la forte connexion entre structure du réseau et propriétés fréquentielles des signaux. Cette dualité est illustrée à travers le traitement d'un réseau par les outils du traitement du signal à des fins de débruitage. Le Chapitre 4 étend la dualité aux réseaux temporels, et propose une méthode d'extraction automatique des structures les plus pertinentes. À travers des analyses empiriques, le bien-fondé de la méthode pour observer l'évolution temporelle du réseau temporel est illustré. Deux méthodes de décomposition de réseau temporel sont également discutées, de manière à repérer dans un réseau temporel, les structures les plus pertinentes à la fois au niveau du réseau mais également à l'échelle temporelle. Elles sont ensuite

appliquées sur le réseau temporel obtenu sur le système de vélos en libre-service présenté au Chapitre 1, permettant de mettre en regard la méthodologie et l'application et de conclure sur l'approche proposée.

CHAPITRE 1

Analyse par les données des systèmes de Vélos en Libre-Service : le cas lyonnais

Résumé –

Ce premier chapitre présente les travaux réalisés sur le système de vélos en libre-service à Lyon, le système Vélo'v. La Section 1 présente le contexte des systèmes de vélos en libre-service. La Section 2 se concentre sur le système Vélo'v, à travers une description du système, du territoire sur lequel il est implanté, et des activités annuelle et hebdomadaire. La Section 3 discute du travail interdisciplinaire sur l'établissement d'une typologie des usagers du système Vélo'v. La Section 4 présente enfin quelques éléments reliés à la classification des stations.

Sommaire

1 Les systèmes de vélos en libre-service

1.1 Présentation et historique des systèmes de vélos en libre-service (VLS)

Présentation Les systèmes de partage de vélos sont devenus ces dix dernières années des éléments incontournables dans les politiques de transport urbain, comme en témoigne l'explosion récente du nombre de vélos en circulation dans les grandes villes du globe. En proposant un accès abordable au vélo, ils doivent participer activement à la mise en place d'alternative aux véhicules motorisés pour les déplacements urbains et contribuent à la réduction de la pollution de l'air, du niveau de bruit et des problèmes de congestion touchant les grandes métropoles mondiales. Une des clés du succès du partage de vélos est sa simplicité d'utilisation. Dans la majorité des systèmes, l'utilisateur a la possibilité de retirer ou de déposer un vélo dans une des nombreuses stations réparties sur toute la ville, le tout de manière entièrement automatisée. Moyennant un abonnement journalier ou hebdomadaire, disponible immédiatement à la station à l'aide d'une carte de paiement, ou annuel, chaque trajet donne droit à une période d'utilisation gratuite ou de coût très faible, généralement de 30 minutes à 1 heure, au-delà de laquelle le prix augmente sensiblement, incitant à des trajets courts. Le nombre de trajets est néanmoins illimité dans la période d'abonnement, permettant d'enchaîner plusieurs trajets de suite, sous réserve de pouvoir déposer son vélo dans une station.

Selon le site « Bike sharing map » [163], qui recense et surveille en temps réel les systèmes de partage de vélos autour du monde, le nombre de villes ayant mis en place un système VLS est passé de 13 en 2004 à 855 en 2014, et la barre symbolique du million de vélos partagés en circulation a tout récemment été franchie au début de l'année 2015. Si les systèmes actuels les plus importants sont majoritairement situés en Chine, qui comprend à elle seule 237 villes avec un système VLS et représente environ 80 % de la flotte totale de vélos à travers le monde, les systèmes VLS sont également très implantés dans les pays occidentaux, en Europe et en Amérique du Nord, et plus particulièrement en France, qui comprend la deuxième flotte mondiale en nombre de vélos et s'est révélée être un pays précurseur dans le développement du partage de vélos. La Table 1.1 recense une sélection de quelques villes mondiales, classées par le nombre de vélos en circulation [250, 163] : la part des villes chinoises est très importante (14 villes sur les 20 premières), même si les grandes villes européennes ont été les premières à développer des systèmes à très grandes échelles, par exemple à Paris avec le système Vélib.

Historique Malgré son succès ces dernières années, l'idée du partage de vélos est relativement ancienne et a germé à Amsterdam en 1965 sous l'impulsion du groupe contestataire et libertaire Provo [188]. Bien avant les préoccupations actuelles pour l'environnement, l'objectif était déjà de promouvoir le vélo comme alternative à la voiture et d'améliorer la qualité de vie dans la ville. Le projet, baptisé « Vélos Blancs »(*Witte Fietsen*) ne vit cependant pas le jour, à la fois par le manque d'implication des institutions et par l'échec d'un projet pilote qui conduisit au vol des 50 vélos mis en circulation. Il fut cependant à l'origine d'initiatives ultérieures, comme à la Rochelle où des « Vélos Jaunes » furent mis en circulation en 1974 par la Mairie, qui connurent néanmoins le même sort [27]. L'intérêt des villes pour la mise en place de partage de vélos en complément des moyens de transport urbain traditionnels incita à de nouvelles expérimentations et, dans les années 90, une deuxième génération de système VLS [71] apparut à Copenhague comprenant environ 1000 vélos répartis dans la ville. Des difficultés de gestion freinèrent l'expansion du système, notamment liées à l'anonymat de la location. Il fallut attendre la fin des années 90 et le fort développement des technologies numériques, notamment des systèmes de communication, pour permettre au partage de vélos d'être une solution envisageable pour les municipalités.

Le développement d'une troisième génération de systèmes VLS permit en effet l'identification des vélos et des usagers de manière centralisé, facilitant à la fois la location pour l'usager et la maintenance pour l'exploitant. Un premier système VLS basé sur des cartes magnétiques vit le jour sur le campus de Portsmouth en 1996, avant que le programme « Vélo à la carte » à Rennes, développé par ClearChannel,

#	Pays	Ville	# de stations	# de vélos	Année de mise en service
1	Chine	Hangzhou	2965	78000	2008
2	Chine	Taiyuan	1262	41000	2012
3	France	Paris	1229	19000	2007
...	...	...	...	...	
8	Royaume-Uni	Londres	743	9800	2010
...	...	...	...	...	
17	Espagne	Barcelone	424	6000	2007
18	Canada	Montréal	411	5120	2009
19	Belgique	Bruxelles	346	4115	2009
20	France	Lyon	346	3200	2005
21	États-Unis	Washington	344	2800	2010
...	...	...	...	...	
29	Espagne	Valence	276	2400	2010

TABLE 1.1 – Liste d'une sélection de villes ayant adopté un système VLS en 2014, classées en fonction du nombre de stations. Les chiffres sont indicatifs, le nombre de stations et de vélos pouvant varier sensiblement au cours du temps [250, 163].

n'apparaisse en 1998. L'engouement pour les systèmes VLS dans les années 2000, dans un contexte de prise de conscience environnementale et d'amélioration de la qualité de vie dans la ville, provoqua le succès du partage de vélos que l'on connaît aujourd'hui, avec des motivations similaires au projet des « Vélos Blancs » néerlandais, mais avec une technologie permettant une mise en œuvre à grande échelle. La quatrième génération a des contours encore floues, mais pourrait concerner la mise à disposition de vélos à assistance électrique, comme proposé à Madrid depuis 2014, ou la mise en place de politique d'auto-régulation à l'aide d'application sur smartphones [71].

Modèle économique Malgré le coût généralement faible pour l'usager, les systèmes VLS restent chers pour les collectivités. Il existe plusieurs types de modèles économiques pour la fourniture d'un service de partage de vélos, le plus répandu étant celui proposé par un des deux principaux afficheurs publicitaires mondiaux que sont JCDecaux et ClearChannel. Dans ce type de financement, le service de partage de vélos est couplé à la gestion de l'affichage publicitaire urbain ainsi qu'à l'équipement en mobilier urbain. Le premier système VLS automatisé mis en place en France à Rennes par ClearChannel en 1998 a ainsi été mis en place suivant ce modèle. JCDecaux a néanmoins été un des acteurs forts du développement des systèmes VLS en France, en déployant dans plusieurs grandes villes françaises (Lyon, Paris, Marseille, Toulouse entre autres) son système de partage de vélos par le biais de sa filiale Cyclocity. Dans les deux cas, l'entreprise s'occupe de l'aménagement du système (mise en place l'infrastructure matérielle et logicielle) de l'exploitation (gestion des paiements, des abonnements, du centre d'appel, du système de rotations des vélos) et de la maintenance (réparation des vélos, entretien des stations). En échange, l'entreprise perçoit les revenus publicitaires liés aux panneaux qu'il exploite, induisant ainsi un manque à gagner pour les collectivités. Selon plusieurs études [43, 200], le coût moyen pour un système VLS oscillerait entre 2000 € et 3000 € par vélo et par an, en incluant les frais d'investissement, d'exploitation et de maintenance ainsi que les coûts, largement sous-estimés pour les premiers systèmes, liés au vandalisme [198]. La viabilité économique des systèmes VLS semble ainsi compromise pour les villes moyennes, entraînant des discussions sur la possible fermeture du service, comme à Valence ou Pau [197] ou des fermetures effectives comme à Aix-en-Provence ou dans d'autres systèmes à l'étranger [163].

Un autre modèle économique assez répandu consiste à déléguer la gestion des vélos en libre-service par une régie de transport, une entreprise publique souvent également en charge des autres types de

transport en commun. Ce système présente également des limites financières, même si les sources de financement sont plus diverses. À Rennes, le partage de vélo est subventionné par la collectivité et fait partie de l'offre de transport STAR fournie par Rennes Métropole, laissant la gestion à un exploitant qui n'est pas spécialiste des systèmes VLS. À Montréal, le système Bixi, auto-financé par les abonnements des utilisateurs, suscite des doutes sur sa viabilité économique [10] malgré des tarifs d'abonnement supérieurs à la moyenne (autour de 80 $ pour 6 mois d'utilisation par an, et 5 $ la journée). Quant au système londonien, basé sur la même infrastructure que Bixi, il est sponsorisé par une entreprise privée qui lui donne son nom, avec les risques que cela amène lorsqu'un contrat arrive à sa fin : le système, sponsorisé par Barclays jusqu'à 2013, est maintenant financé par Santander [187] à des conditions jugées désavantageuses pour la ville.

Face à ces coûts importants, certaines collectivités proposent de nouvelles approches de location de vélos, par exemple à Strasbourg où des locations longue-durées, de 1 jour à 1 an, permettant de limiter les risques liés à la dégradation, ainsi que les coûts de fonctionnement [107].

Cette rapide présentation des systèmes de vélos en libre-service montre à la fois l'engouement actuel des municipalités pour ce système de transport propre, économique pour l'usager et incitant à la mobilité urbaine. Néanmoins, il n'est pas sans limite, notamment financière, et appelle ainsi à des recherches actives afin d'une part, analyser le fonctionnement des systèmes VLS et d'autre part, l'améliorer.

1.2 Vue d'ensemble des études sur les systèmes VLS

Parallèlement à l'explosion des systèmes VLS dans le monde, les recherches sur le partage de vélos ont connu un intérêt florissant de la part des chercheurs en transport et au-delà. La synthèse de la littérature sur les systèmes de partage de vélos (*bike sharing systems* en anglais) réalisée par Fishman et al, d'abord en 2013 [92] puis en 2015 [91] permet de se rendre compte à la fois de la profusion de publications scientifiques sur ce sujet et de la diversité des thèmes de recherche abordés. Ces études examinent ainsi les multiples questions que pose ce nouveau mode de transport, qu'elles soient techniques – est-ce que le système fonctionne correctement ? est-il optimal ? est-il améliorable et comment ? – sociales – qui utilise les VLS et pourquoi ? comment améliorer l'accès aux VLS ? – en rapport avec le domaine du transport – quel est l'impact des systèmes VLS sur les autres modes de transport ? – ou bien traitant les aspects économiques, comme savoir ce que l'activité des systèmes VLS nous dit sur l'activité économique d'une ville. Cet engouement s'explique par la nécessité, tant pour les décideurs politiques que les sociétés gestionnaires de tels systèmes, de connaître et décrire les systèmes VLS pour développer et accompagner l'utilisation du partage de vélo dans la ville.

De plus, à la différence des systèmes de transport traditionnels nécessitant la mise en place d'infrastructure souvent lourdes et coûteuses – par exemple à l'aide de capteurs ou d'enquêtes sur le terrain – afin d'obtenir des données partielles sur le système, les technologies employées dans les systèmes VLS de troisième génération actuellement en activité permettent d'accéder rapidement à des données numériques sur de multiples aspects du partage de vélos. Ces données, fournies par les opérateurs – la plateforme « JCDecaux developer » [4] permet par exemple d'accéder à l'état des stations de tous les systèmes gérés par Cyclocity en temps réel – ont favorisé la collaboration entre les spécialistes des questions socio-économiques, qui jusque-là traitaient ces problèmes, et les chercheurs en analyse de données (traitement du signal, fouille de données, etc.), afin de trouver des éléments de réponse aux problèmes cités ci-dessus, dans des jeux de données pouvant être conséquents et nécessitant des techniques sophistiquées.

Les systèmes VLS comme mode de transport Au-delà des enquêtes réalisées par des organismes institutionnels [43, 105, 194] ou les opérateurs eux-mêmes [148, 232] consistant en des descriptions qualitatives ou quantitatives de l'utilisation du système et de sa gestion, la place du partage de vélos

Ville	Analyse spatio-temporelle	Régulation	Prévision des flux	Classification des stations	Autres
Barcelone	[136, 182]			[99, 100, 211]	
Dublin					[258]
Hangzhou			[252]		
Londres	[182]			[39, 147, 211, 235, 261]	[183]
Lyon	[161, 35, 165, 182]		[36, 35]	[37, 39]	[134, 190]
Montréal	[182]		[84]		
Paris	[61, 182]	[175]		[39, 61, 195]	
Vienne	[182]	[241]		[241]	
Washington	[182]	[216]		[261]	[86]

TABLE 1.2 – Publications traitant de l'analyse de données pour l'étude des systèmes VLS, classées par ville et par thème de recherche, sur une sélection des villes les plus étudiées.

parmi les modes de déplacements urbains a été l'objet de nombreuses études, cherchant à comprendre les raisons du succès des systèmes VLS. Elles concernent l'analyse d'un système en particulier, tel que celui de Montréal [169], Hangzhou [218], Bangkok [237], Londres [180] ou Dublin [236], ou des études comparatives entre plusieurs systèmes, en Amérique du Nord, Europe occidentale et Asie [153, 167, 186, 217].

La cohabitation avec les autres modes de transport et en particulier, les effets des systèmes VLS sur la réduction de l'usage de la voiture ont été particulièrement scrutés : les agglomérations urbaines cherchent en effet à promouvoir le déplacement en vélo, dans le cadre de politiques de maîtrise de la consommation d'énergie et des émissions de gaz à effet de serre issues principalement du transport motorisé [71]. Plusieurs études ont montré le faible impact du partage de vélos sur l'usage de la voiture [148, 168, 173, 257], à travers des enquêtes terrains : les usagers des systèmes VLS sont en très grande majorité des utilisateurs des modes de transport doux (transports en commun, marche, etc.). Le rôle des systèmes VLS dans la promotion du vélo en ville a également été étudié à travers le passage entre vélo partagé et vélo personnel [49] ou des facteurs influençant la pratique du vélo en partage [141, 196]. Les bénéfices sur la santé de la pratique du vélo induits par l'utilisation des systèmes VLS ont aussi fait l'objet de plusieurs études [79, 93, 205, 251], et l'exemple londonien [251] illustre l'impact global positif de la pratique du vélo en ville sur certaines catégories de population.

La caractérisation des usagers, et leur motivations, sont également à l'origine de nombreuses études, principalement conduites par des chercheurs en sociologie. Ces travaux ont été réalisés soit à l'aide d'enquêtes individuelles, comme à Montréal [16, 101], ou bien de méthodes d'analyse de données, comme à Londres [183]. Parmi les thèmes abordés, on peut citer la compréhension des comportements mis en jeu dans la relation entre les usagers et le partage de vélos [33], l'étude des caractéristiques sociales des utilisateurs des systèmes VLS tels que le genre [22] ou le niveau social [123], ainsi que les facteurs de différenciation entre les cyclistes réguliers possédant leur propre vélo et les utilisateurs des systèmes VLS [42].

Analyse de données VLS L'analyse des systèmes VLS par les données est également à l'origine de nombreuses études, souvent multidisciplinaires. Les systèmes VLS sont caractérisés par des analyses temporelles et spatiales, via par exemple la classification des stations de vélos, comme à Paris [61, 174, 195], Barcelone [99, 100, 136] ou Londres [147, 235] pour ne citer que les principaux. Des approches comparatives sur plusieurs villes ont également été mise en œuvre [39, 182, 211, 261], et rendent compte de la forte similarité entre tous les systèmes, notamment sur la répartition spatio-temporelle de l'activité. La plupart de ces études mettent en évidence des relations entre l'usage des vélos, l'heure de la journée et une description géographique et socio-économique de la ville.

Beaucoup d'efforts ont également été déployés, notamment par des groupes de recherche issus des sciences exactes, pour prévoir la demande afin de planifier le système et d'en anticiper le développe-

ment (où localiser les stations dans la ville, combien d'emplacements par station, etc.), à la fois par des approches théoriques [209, 151, 104, 70, 55, 38, 146] que par des approches pilotées par les données comme à Lyon [35, 37], à Hangzhou [252] ou à Montréal [84]. Ce dernier problème a d'ailleurs fait récemment l'objet d'un concours, en marge d'une conférence sur la fouille de données, témoignant de la popularité de cette thématique [5]. D'autres problématiques, liées au bon fonctionnement du système au quotidien (maintenance des vélos, équilibrage des stations, etc.) ont également été étudiées, une fois encore par le développement de modèles théoriques [214, 246, 242, 243, 216, 44, 143, 26, 180, 174] et par l'analyse de données réelles [174, 241, 216].

Ces données posent également le problème de leur visualisation, prenant en compte à la fois les dimensions temporelle et spatiale de l'activité. Dans le cadre du *Hubway Data Visualization Challenge* [6], plusieurs contributions ont proposé, par le biais une nouvelle fois d'un concours, des solutions originales permettant de visualiser dans le temps et dans l'espace les données. Un outil a également été développé pour représenter les déplacements à Lyon de manière adaptée et propice à l'analyse [190].

La Table 1.2 donne un aperçu des publications traitant de l'analyse de données pour l'étude des systèmes VLS, classées par ville et par thème de recherche, pour une sélection de villes.

2 Description du système Vélo'v

2.1 Un système de partage de vélos performant

Vélo'v [156] est le système de vélos en libre-service mis en place dans la ville de Lyon et de Villeurbanne à partir de mai 2005 par la communauté urbaine du Grand Lyon. Le système comprend actuellement 348 stations, avec environ 4000 vélos en circulation. En 2014, plus de 8.3 millions de locations ont été réalisées par les 58 000 abonnés annuels et les 773 410 utilisateurs ponctuels utilisant le système sur une journée ou une semaine [157]. Chaque vélo est partagé en moyenne 6 fois par jour, ce qui en fait un des systèmes les plus actifs d'Europe, et un des premiers succès des systèmes VLS. Il est en effet le premier système de grande envergure à avoir été mis en place dans le monde, et a notamment ouvert la voie à la mise en service de systèmes VLS à très grande échelle, tel que le système Vélib à Paris, actuel plus important service d'Europe [91]. En mai 2015, le système fête ses 10 ans de mise en service en organisant comme pour chaque anniversaire une ascension en Vélo'v de la colline de Fourvière [1]. Cet événement, témoigne de l'ancrage du système Vélo'v dans le quotidien des habitants du Grand Lyon et de sa popularité.

La Figure 1.1a montre une photographie d'une station située dans le 5^{e} arrondissement de Lyon : les vélos sont disponibles à la location tandis que les bornettes libres peuvent accueillir des vélos. La borne d'accueil permet de s'abonner et de s'identifier afin de louer un vélo et consulter son compte. La Figure 1.1b montre un exemple de vélo qu'il est possible de louer dans une des stations du système Vélo'v. Ses caractéristiques techniques (vélo lourd, stable et robuste) sont adaptées à la circulation urbaine et à l'utilisation massive par tout type de personne. Comme la plupart des systèmes VLS, Vélo'v est accessible à partir d'un abonnement soit courte durée (1 jour ou 1 semaine) disponible directement sur la borne de n'importe quelle station, moyennant la possession d'une carte bancaire. Il est également possible de s'abonner pour un an, nécessitant une inscription plus complète accessible en ligne ou par courrier. Dans les deux cas, l'abonnement permet l'utilisation gratuite d'un vélo pendant trente minutes ; une fois cette période écoulée, l'utilisateur peut continuer sur le même vélo et payer par tranche de demi-heure jusqu'à la restitution du vélo, ou alors restituer le vélo et en emprunter un autre pour lequel trente minutes d'utilisation gratuite sont de nouveau disponibles. Le succès du système Vélo'v tient en partie à la politique tarifaire du système, très avantageuse pour l'utilisateur. En mai 2015, le ticket journée est fixé à 1.50€ et le ticket semaine à 5€, alors que l'abonnement longue-durée de 1 an coûte 25€, avec des réductions pour les étudiants et les personnes bénéficiaires du RSA. Il est de plus possible de coupler

(a) (b)

FIGURE 1.1 – (a) Station Vélo'v 5002 - Saint Just / Compagnon de la Chanson. Les vélos sont disponibles à la location tandis que les bornettes libres peuvent accueillir des vélos. La borne d'accueil permet de s'abonner et de s'identifier afin de louer un vélo et consulter son compte [32] (b) Exemple de vélo disponible à la location dans le système Vélo'v [31].

son abonnement Vélo'v avec l'abonnement des transports en commun lyonnais proposé par TCL [225], permettant de bénéficier de 30 minutes supplémentaires de gratuité par trajet.

2.2 Le territoire urbain du Grand Lyon

Le Grand Lyon comprend 58 municipalités et plus de 1.3 millions d'habitants, dont la moitié vit dans une des deux grandes villes de l'agglomération, Lyon (485 000 habitants) et Villeurbanne (145 000 habitants). Le centre-ville – Lyon, et dans une moindre mesure, Villeurbanne – a des caractéristiques démographiques distinctives par rapport aux autres grandes villes françaises [12] : en moyenne la population est plus jeune et plus qualifiée (66 % de la population a moins de 45 ans, par rapport à 58 % à Marseille), et le nombre de foyers avec une seule personne est important (1 sur 2 à Lyon par rapport à 1 sur 3 en moyenne en France). La population active est constituée à 80 % de cadres, cadres supérieurs et employés. Selon l'INSEE [129] en 2010, les étudiants représentent 18 % de la population entre 15 et 64 ans, soit une proportion importante si l'on compare avec Paris (10.5 %) et Marseille (10 %).

La répartition démographique à l'intérieur du Grand Lyon est également hétérogène : la ville de Lyon comprend la plus forte proportion de diplômés et de cadres par rapport aux autres communes. Néanmoins, l'aire résidentielle de la ville n'est pas homogène et de fortes disparités sociales existent. À l'échelle du Grand Lyon, et particulièrement pour les villes près du centre-ville, le territoire est marqué par une dissymétrie Est/Ouest, les populations de faible niveau social étant principalement situées à l'est de Lyon. La carte de la Figure 1.2 affiche une comparaison dans chaque arrondissement entre la taille de la population active de 15 à 75 ans, potentiellement intéressée par le service Vélo'v, et le nombre de stations disponibles. Les cas du 8^{e} et 9^{e} arrondissements soulignent les inégalités économiques à l'intérieur de la ville : les arrondissements les plus pauvres sont aussi ceux pour lesquels l'accès au système Vélo'v est le plus difficile.

Les stations du système Vélo'v, qui peuvent accueillir entre 10 et 40 vélos, sont principalement situées dans les villes de Lyon, divisée en 9 arrondissements, et de Villeurbanne [1]. La distance entre deux stations est en moyenne de 255 mètres, et au plus de 850 mètres [165]. Cette distance tend à augmenter avec la récente extension du réseau vers les communes périphériques. La répartition spatiale

1. Par la suite, Lyon désigne l'aire correspondant aux villes de Lyon et de Villeurbanne, sauf mention contraire explicite.

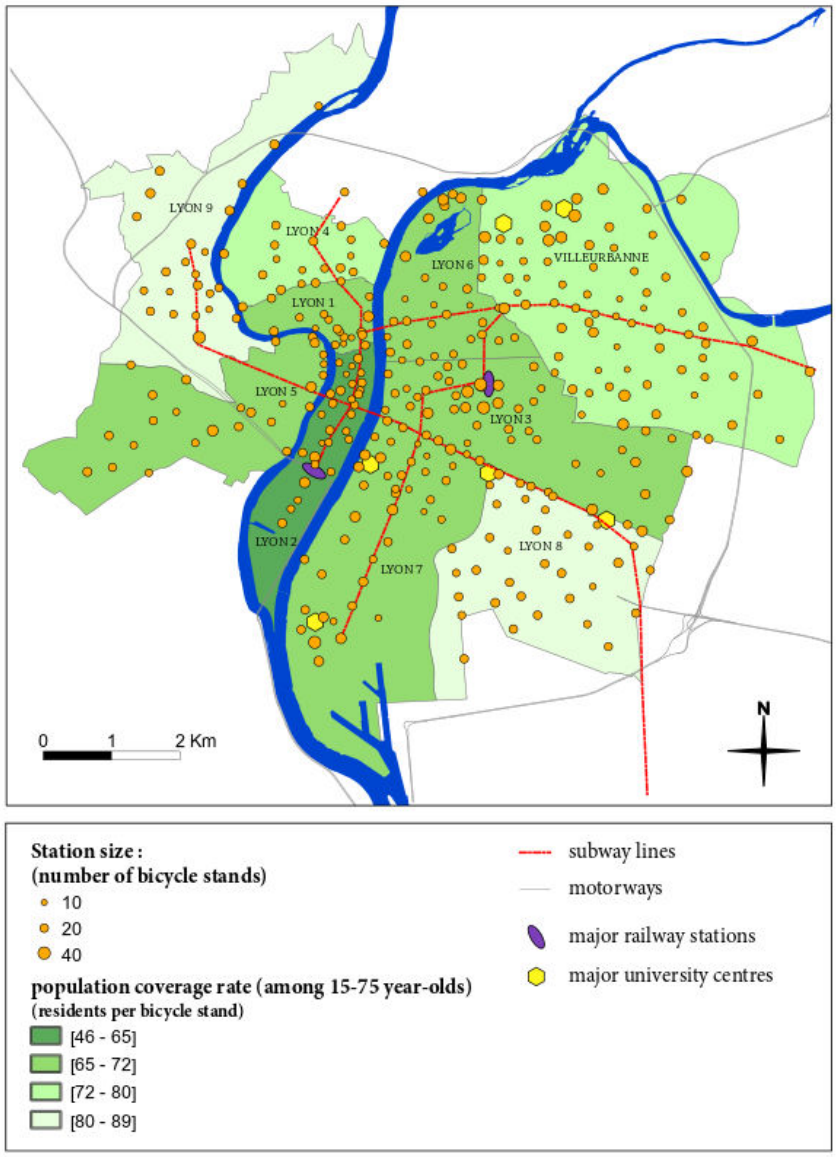


FIGURE 1.2 – Carte des stations Vélo'v à Lyon en 2011, avec la population par arrondissement et les principaux lieux d'intérêts. (Extraite de [240])

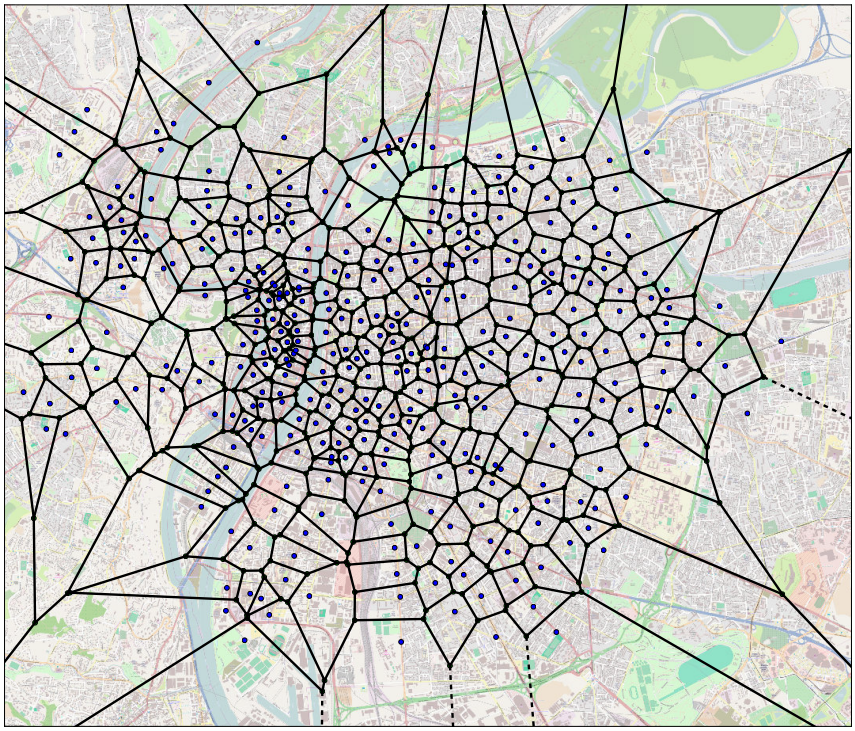


FIGURE 1.3 – Maillage des stations Vélo'v : les points bleus représentent les stations, et les lignes noires le diagramme de Voronoï.

des stations est hétérogène : les stations sont en majorité dans le centre-ville de Lyon, près des universités et près des nœuds de transport majeurs (stations de métros, tramways, gares), soulignant l'intégration du réseau Vélo'v avec le réseau de transport lyonnais TCL [225]. La Figure 1.3 affiche le maillage des stations Vélo'v : les points bleus représentent les stations, et les lignes noires le diagramme de Voronoï [2] correspondant. Elle montre la disparité dans la couverture en stations de l'aire urbaine : les arrondissements de l'ouest (4e et 5e) souffrent de contraintes topologiques dues à l'altitude (collines de la Croix-Rousse et de Fourvière).

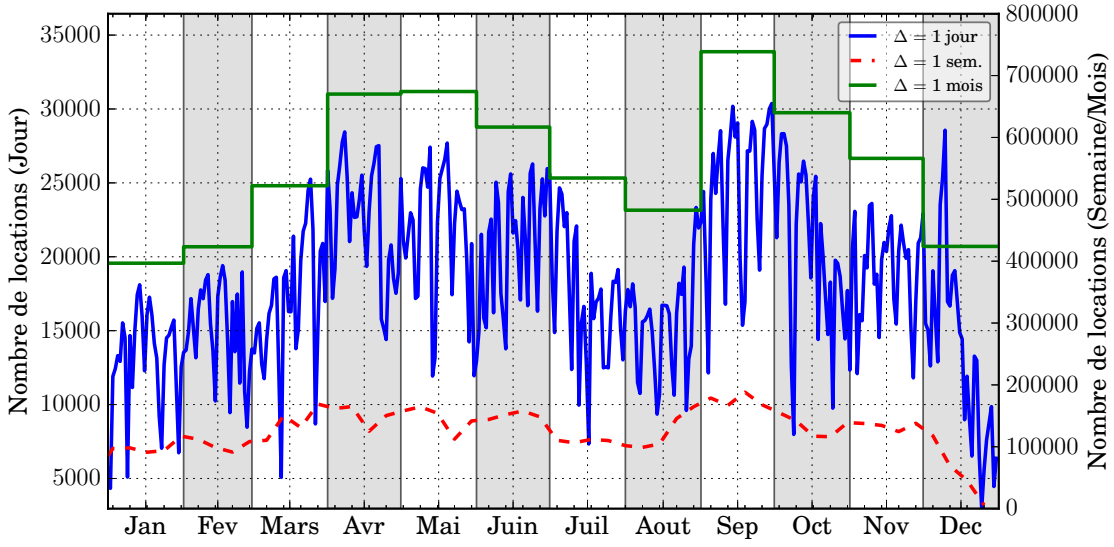
2.3 Description quantitative du système Vélo'v

2.3.1 Analyse spatio-temporelle

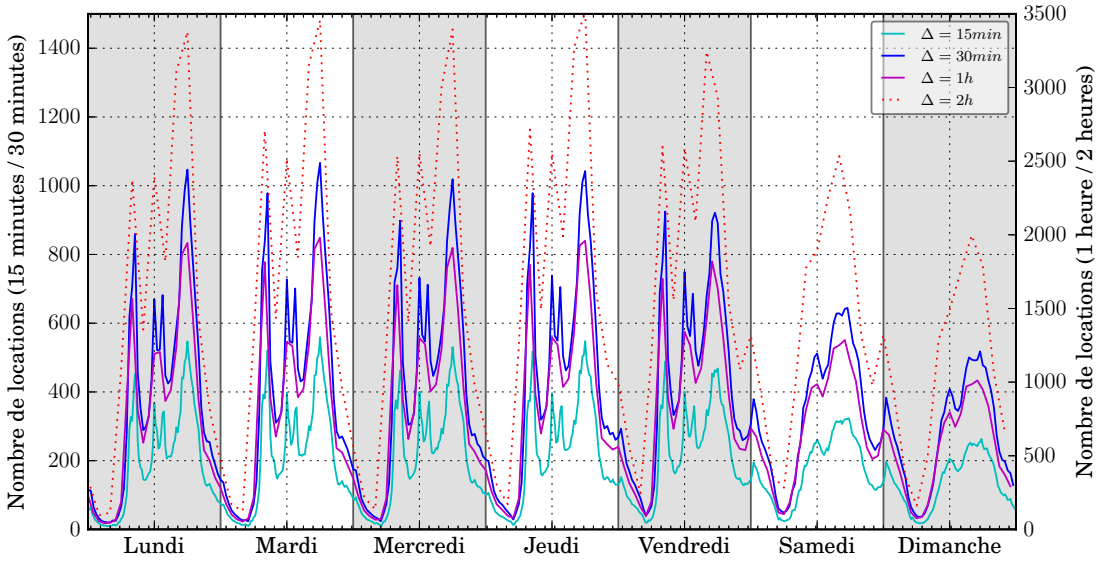
Les premières analyses quantitatives du réseau Vélo'v ont été publiées dans [161] sur une base de données partielle des données de 2005 à 2007. Les tendances observées se retrouvent par la suite sur les années suivantes [35, 134] sur des jeux de données plus complets, et permettent de dresser un premier aperçu de l'utilisation du système. Il s'avère ainsi que le système Vélo'v est principalement utilisé pour de faibles distances, de l'ordre de quelques kilomètres, et une durée de parcours inférieure à 30 minutes, correspondant à la limite d'utilisation gratuite du vélo.

Une description de l'activité du système par le biais du nombre de locations a également été réalisée, mise en place ici pour l'année 2001. La Figure 1.4a affiche le nombre de locations réalisées pour le système Vélo'v en agrégeant les locations par jour, par semaine et par mois. L'agrégation par jour des déplacements met en évidence une évolution du nombre de locations non-stationnaires dans le temps, dont les fluctuations liées à l'activité journalière cache la tendance liée au cycle des saisons. Cette tendance est visible plus distinctement sur la courbe obtenue lorsque l'agrégation se fait à la semaine, et de

2. Chaque zone contient les points de la carte dont la station la plus proche est celle associée à la zone.



(a) Nombre de locations sur l'année



(b) Nombre moyen de locations sur la semaine

FIGURE 1.4 – Nombre de locations pour le système Vélo'v pour l'année 2011 pour différentes périodes d'agrégations.

manière encore plus claire lorsque l'agrégation se fait au mois. L'étude des rythmes sur l'année permet ainsi de mettre en évidence une utilisation plus intense du système Vélo'v au mois de septembre, suivi des mois d'avril et mai. Le mois d'août est en revanche le mois pendant lequel le système est le moins utilisé, avec les mois de décembre, janvier et février.

Le rythme des saisons a ainsi une influence notable sur l'utilisation du système Vélo'v. Cela peut tout d'abord s'expliquer par les conditions météorologiques : la température, qui peut descendre sous les $0\,^{\circ}\mathrm{C}$ à Lyon en hiver, mais surtout la pluie, dissuadent les utilisateurs à se déplacer en vélo, alors qu'au contraire le printemps avec des températures plus douces et des jours ensoleillés, favorise la pratique du vélo. La baisse de l'utilisation pendant les mois d'été s'explique, plus que par les fortes chaleurs, par les vacances scolaires. L'hypothèse d'un service massivement utilisé par les étudiants se confirme en observant le regain en septembre du nombre de locations, le mois pendant lequel le système est le plus utilisé. L'arrivée de nouveaux étudiants pour la rentrée scolaire, combinée avec le faible coût de l'abonnement pour les étudiants (actuellement à 15 € l'année, à comparer au 28 € de l'abonnement mensuel aux transports en commun) provoque cet engouement, qui semble s'estomper les mois suivants.

Les rythmes sur la semaine donnent également des indications intéressantes sur l'utilisation du système Vélo'v. La Figure 1.4b affiche le nombre moyen de locations réalisées sur la semaine pendant l'année 2011, en agrégeant les locations par quart d'heure, par demi-heure, par heure et par tranche de deux heures. Le motif temporel met en évidence deux pics importants pour les jours de semaine, situés le matin et en fin d'après-midi, et un ou deux pics d'intensité plus faible entre midi et 14h, suivant le choix de la période d'agrégation. Les jours de week-end ont quant à eux des comportements différents, avec des pics plutôt situés en fin de matinée et en fin d'après-midi.

Le nombre de locations de vélos est ainsi un bon indicateur des rythmes urbains qu'il est courant d'observer à l'échelle d'une ville : les pics du matin et du soir, situés respectivement vers 9h et vers 17h, sont ainsi révélateurs des trajets domicile-travail. Le pic du midi, qui se décompose en deux pics, à midi et à 14h, montre les trajets liés à la pause déjeuner. Ce schéma, qui se répète quasiment à l'identique sur les cinq jours de la semaine, semble d'intensité légèrement plus faible le mercredi, témoignant probablement des rythmes scolaires. Enfin, les pics du week-end laissent entrevoir une activité tournée vers les loisirs. Il est enfin intéressant de noter que ce motif n'est pas spécifique à Lyon mais se retrouve dans de nombreux systèmes de vélos en libre-service [39], avec des décalages dans l'heure des pics en fonction des pratiques culturelles.

Une analyse spatiale est ensuite conduite afin de détecter des motifs spatiaux dans l'usage des vélos, permettant de se rendre compte de la non-uniformité spatiale de l'utilisation du système Vélo'v. Des motifs de déplacements connus sont observables, comme les déplacements de zone d'habitation vers les campus le matin et inversement, ou des trajets vers les zones d'activité comme le centre-ville pendant la soirée.

La Figure 1.5 affiche une carte des stations les plus actives sur la semaine, pour laquelle chaque point est une station, et la couleur représente le nombre moyen de vélos par semaine qui arrive et qui part de la station. La localisation des principaux centres d'intérêt de Lyon, en complément de la carte à la Figure 1.2, est également affichée.

Pendant les trois pics des jours de semaine, l'activité se concentre dans le quartier Part-Dieu, dans lequel se situe la gare, un centre commercial et le quartier des affaires, confirmant ainsi le caractère professionnel des trajets réalisés pendant ces périodes. Les stations situées autour des campus universitaires, comme par exemple celui de la Doua à Villeurbanne, sont également sollicitées pendant ces moments de la journée. Une activité importante est également présente dans la Presqu'île, qui regroupe les zones commerciales du centre-ville. La présence de nombreux restaurants, théâtres et bars, notamment dans le quartier de l'Opéra (en haut de la zone située sur la carte) explique également la forte activité nocturne

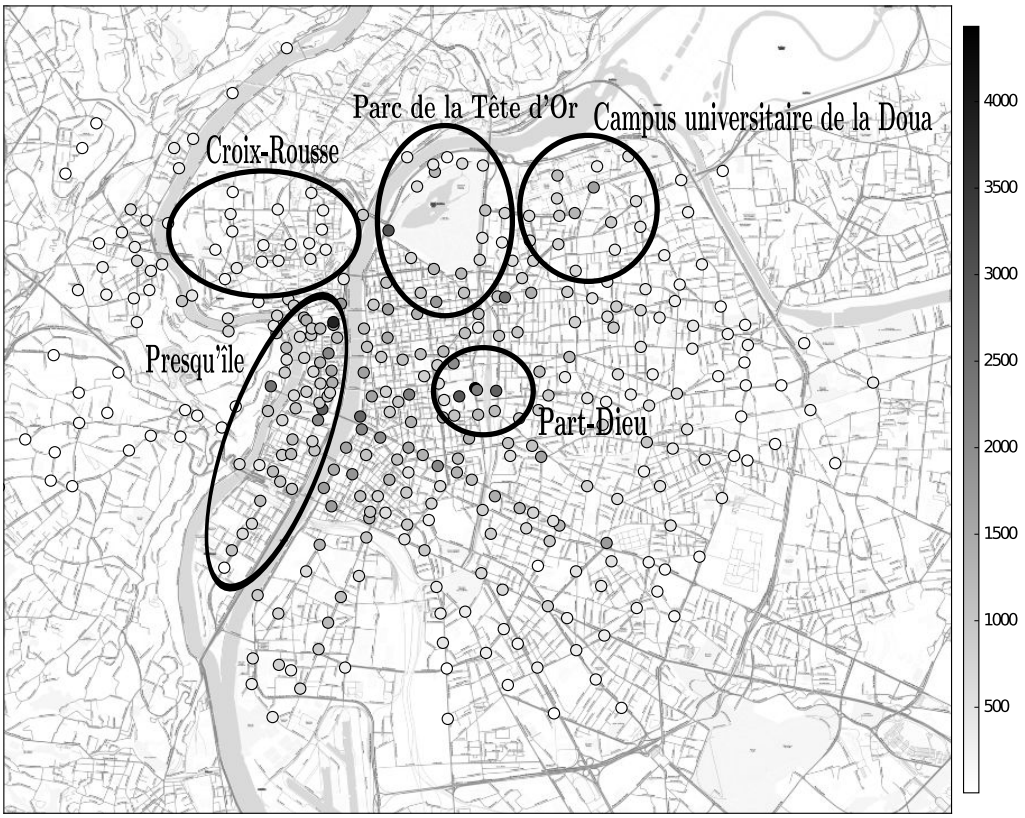


FIGURE 1.5 – Carte des stations les plus actives. Chaque point est une station, et la couleur représente le nombre moyen de vélos par semaine qui arrive et qui part de la station. Les zones encerclées correspondent aux principaux centres d'intérêt de Lyon.

de cette zone. Enfin, les activités du week-end sont différentes, avec une augmentation des trajets depuis et vers les zones de loisir, comme le Parc de la Tête d'Or au nord de Lyon.

Il est intéressant également de noter les zones dans lesquelles l'activité est faible. De manière globale, les stations situées à la périphérie de la ville ont une activité faible, voire quasiment inexistante. Cela s'explique à la fois par le fait que ces stations sont excentrées, et donc difficilement atteignables et avec peu d'activités commerciale ou culturelle. Un autre facteur à prendre en compte à Lyon est la présence de collines, dont celle de la Croix-Rousse mentionnée sur la carte. De manière peu surprenante, les utilisateurs sont plus enclins à descendre les vélos qu'à les remonter, laissant les stations en haut des collines vides. La régulation mise en place par l'exploitant du système, et qui consiste à remonter les vélos à l'aide de camions, ne permet pas d'enrayer ce phénomène.

2.3.2 Prédiction des flux entrants et sortants

Une autre approche s'est concentrée sur l'élaboration d'un modèle linéaire pour expliquer les flux entrants et sortants dans chaque station. Dans [35], un modèle linéaire est mis en place pour tenter d'expliquer le nombre de vélos loués en moyenne par jour à l'aide de variables météorologiques (température, pluviométrie), du nombre d'utilisateurs enregistrés, du nombre de vélos disponibles, d'un marqueur pour indiquer un jour férié ou des vacances et d'un marqueur pour indiquer un jour de grève. Cette étude a permis de mettre en évidence des phénomènes prévisibles, tels que l'influence de la météo sur les locations, ainsi que l'utilisation du Vélo'v comme un mode de transport quotidien, comme en témoigne la baisse du nombre de locations pendant les jours fériés et les vacances. De plus, l'abonnement des usagers augmente le nombre de locations, en facilitant l'accès au service. Les travaux réalisés dans [166] poursuivent

cette ambition d'expliquer les flux entrants et sortants, en introduisant des variables socio-économiques sur chaque quartier de la ville de Lyon fournis par l'INSEE [129]. D'un point de vue méthodologique, l'utilisation de la parcimonie a permis de sélectionner les variables socio-économiques les plus pertinentes, en supposant que les flux observés sont une observation tronquée de la demande (tronquée car lorsque la station est vide des flux sortants ne sont plus possibles, de même lorsque la station est pleine pour les flux entrants). Cependant, la méthode introduite dans [166] devient inefficace du fait du nombre de stations et de variables à considérer.

Pendant cette thèse, une extension de cette approche a été considérée, et est proposée dans l'Annexe A de ce manuscrit. Cette approche a depuis été poursuivie dans [231].

2.4 Données sur le système Vélo'v

Les travaux présentés dans cette thèse ont été réalisés dans le cadre du projet ANR Vél'innov [2], visant à proposer une approche pluridisciplinaire de l'étude du système Vélo'v. Dans le cadre de ce projet, un partenariat a été mis en place avec Cyclocity, la filiale de JCDecaux qui s'occupe de la gestion de nombreux systèmes VLS européens, dont Lyon et Paris. Il a permis d'avoir à disposition plusieurs jeux de données portant sur les trajets effectués à Lyon, d'abord pour la période allant de mai 2005 à fin 2007, puis récemment pour l'année 2011, période que nous allons considérer par la suite.

Il existe plusieurs types de données disponibles sur les systèmes VLS. Outre des données statiques sur la description du réseau, la plupart des systèmes propose des interfaces pour récupérer de façon automatique l'état des stations, avec une résolution temporelle fine (de l'ordre de la minute). La mise en place d'accord avec les sociétés gestionnaires de systèmes VLS permet également d'accéder à des données sur les mouvements.

Données « Disponibilités » Les données « Disponibilités » sont récupérables en ligne publiquement sur l'interface mis en place par les opérateurs. Pour le système Vélo'v, une connexion au site « JCDecaux Developer » [4] donne accès aux informations sur la disponibilité des stations avec une précision temporelle de l'ordre de la minute.

Données « Mouvements » La mise en place d'un partenariat avec JCDecaux a permis de disposer, dans le cadre du projet ANR Vél'innov, de données sur les mouvements réalisés par les utilisateurs. Ces données, non publiques pour des raisons de confidentialité, proposent les informations suivantes pour chaque mouvement :

— date de départ et d'arrivée du trajet ;
— station de départ et d'arrivée du trajet ;
— identifiant anonyme du client ;
— numéro du vélo ;
— type de trajet.

Les données contiennent également des informations sur les utilisateurs, contenant des informations sur l'âge, le genre et le code postal, uniquement pour les titulaires d'un abonnement longue-durée.

La collecte et le filtrage de ces données ont été une contribution importante de cette thèse. À partir des données brutes, une base de données a été mise en place, de manière à éliminer les erreurs d'enregistrement. Ce travail n'est pas détaillée par la suite.

3 Typologie des usagers Vélo'v par leur pratique du vélo partagé

3.1 Les usagers du système Vélo'v

Les usagers sont les principaux acteurs des systèmes VLS et, si leur étude se fait à travers leurs trajets, il est néanmoins possible et intéressant de les caractériser d'un point de vue sociologique. Grâce

Durée d'abnt	# de mvts	% du total de mvts	# moy. de mvts par abnt
1 jour	1 169 362	18.0	1.7
1 semaine	960 565	14.8	9
1 an	4 363 500	67.2	86

TABLE 1.3 – Nombre de mouvements par type d'abonnement de l'usager, pour l'année 2011 (**mvts** : mouvements ; **abnt** : abonnement).

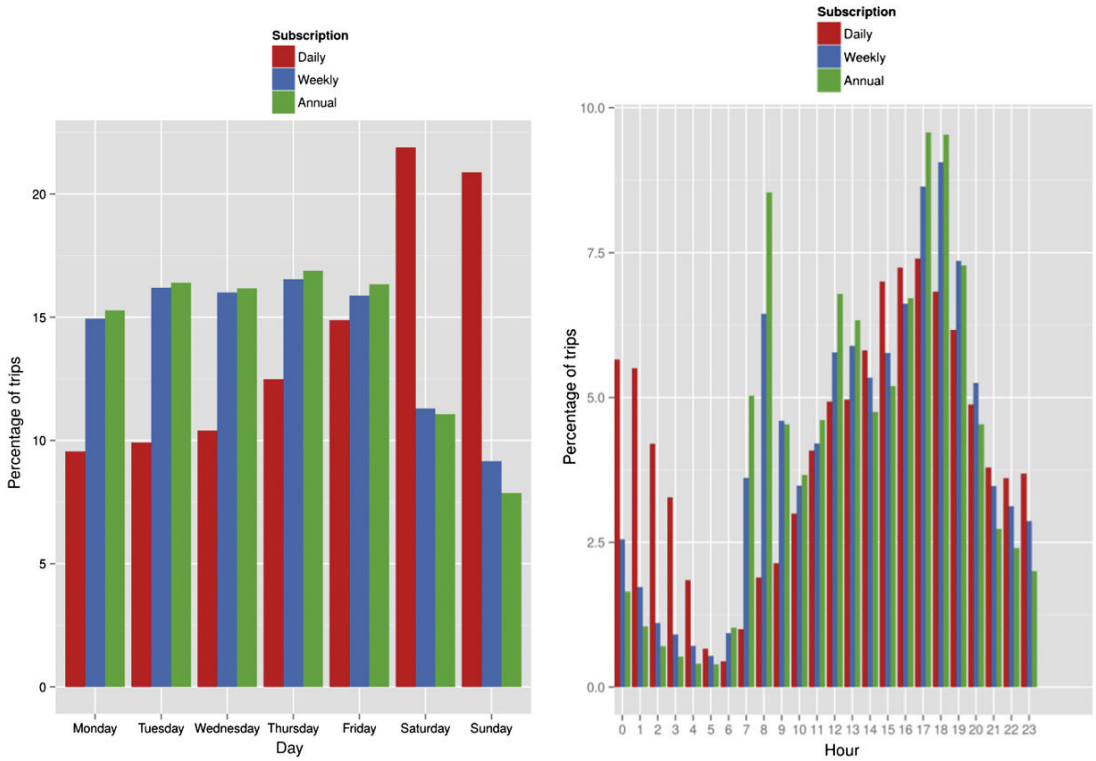
aux données à notre disposition, chaque mouvement est associé à un client, lui-même associé à un type d'abonnement. Il est ainsi possible d'établir une différentiation entre les motifs temporels des abonnés journaliers, hebdomadaires et annuels, et de mettre en place une comparaison entre les différents types d'usagers afin de caractériser leur pratique du partage de vélos.

3.1.1 Comparaison entre les abonnés annuels, hebdomadaires et journaliers

La Table 1.3 donne le nombre de mouvements réalisés par les utilisateurs pour chaque type d'abonnement : ceux réalisés avec un abonnement annuel représentent environ deux tiers des mouvements réalisés en 2011. Les informations supplémentaires disponibles sur les abonnés annuels (collectées lors de l'inscription) permettent de caractériser leur pratique du partage de vélos en fonction de données démographiques et spatiales. Par la suite, une comparaison entre les abonnés annuels, hebdomadaires et journaliers est réalisée, ainsi qu'une étude sur les abonnés annuels.

L'étude des rythmes temporels sur l'année par catégorie d'abonnement met en évidence des disparités dans la pratique du partage de vélos. Les abonnés annuels ont un usage plus régulier sur toute l'année alors que les abonnés courte-durée ont une utilisation plus saisonnière du partage de vélos : la différence entre le mois le plus pratiqué et le mois le moins pratiqué est de l'ordre 50 % pour les abonnés annuels, alors qu'elle est de 80 % pour les autres usagers.

La Figure 1.6a affiche les rythmes sur la semaine pour les trois types d'abonnements. Une différentiation est assez nette entre les usagers annuels et hebdomadaires d'un côté, dont l'utilisation est plutôt régulière les jours de semaine mais chute d'environ un tiers le samedi et de moitié le dimanche, et les usagers journaliers, qui ont une utilisation plutôt concentrée sur la fin de semaine. Les rythmes journaliers, affichés sur la Figure 1.6b, présentent également des différences entre les types d'abonnements : les rythmes d'utilisation des abonnés annuels et hebdomadaires présentent des pics d'utilisation le matin autour de 8h, le soir vers 18h ainsi qu'entre midi et 14h. Au contraire, l'utilisation du système Vélo'v par les abonnés journaliers est plutôt concentrée en fin de journée et pendant la nuit. Ces résultats montrent ainsi des motifs temporels d'utilisation différents en fonction du type d'abonnement : les usagers avec un abonnement hebdomadaire ou annuel sont plutôt des utilisateurs réguliers, qui utilisent le système comme un moyen de transport quotidien, pour des trajets professionnels. Au contraire, l'abonnement journalier est plutôt réservé à des trajets ponctuels, hors des cycles traditionnels des activités urbaines. La forte proportion des mouvements réalisés la nuit laisse à penser que ces trajets sont reliés à des activités nocturnes de loisir, et compensent l'absence de transport en commun après minuit. Cette différentiation dans l'utilisation du système par les différents types d'abonnement est renforcée par la durée moyenne des trajets, plus élevée (22.8 minutes) pour les abonnés annuels et hebdomadaires que pour les abonnés journaliers (13.5 minutes).



(a) Moyenne sur la semaine

(b) Moyenne sur la journée

FIGURE 1.6 – Pourcentage des trajets réalisés par les abonnés annuels, hebdomadaires et journaliers pendant l'année 2011 (Extraite de [240]).

3.1.2 Répartition géographique, genre et âge des abonnés

Le nombre d'abonnés [3] du système Vélo'v atteignait approximativement 50 000 personnes en 2011. La carte 1.7 illustre la répartition géographique des abonnés dans l'aire urbaine du Grand Lyon et le département du Rhône, obtenue à partir du code postal. Il en ressort que les abonnés vivent en très grande majorité dans l'aire géographique couverte par le système Vélo'v : 84.2 % des abonnés vivent dans le centre-ville, 7.3 % vivent en dehors des limites du système Vélo'v mais dans l'aire urbaine du Grand Lyon, et enfin 8.5 % vivent en dehors du Grand Lyon, en majorité dans le département du Rhône. Ces résultats contredisent ceux obtenus pour la ville de Londres [183, 22] où les utilisateurs tendent à vivre éloignés de l'aire couverte par le système VLS Barclays. De manière générale, la proportion d'abonnés décroît avec la distance entre leur résidence et le centre-ville de Lyon. Il y a néanmoins des exceptions pour les communes situées à l'ouest et au nord-ouest de Lyon, en raison de la meilleure intégration de ces communes avec le système Vélo'v et des profils socio-professionnels des populations concernées, plutôt situés dans les catégories élevées.

Une légère sous-représentation des femmes (44 % des abonnés) est visible dans la population des abonnés Vélo'v. Cette sous-représentation se vérifie pour toutes les classes d'âge considérées, à ceci près que le ratio entre femmes et hommes peut varier significativement : il est ainsi plus faible pour les abonnés entre 31 et 45 ans et entre 14 et 19 ans. La répartition des abonnés par classe d'âge est donnée par la Table 1.4. Les classes d'âge 18-24 ans et 25-34 sont sur-représentées parmi les abonnés Vélo'v, par rapport à la population lyonnaise : ces deux groupes représentent environ 60 % des abonnés, alors

3. Le terme « abonnés » désigne par la suite les usagers titulaires d'un abonnement annuel, en opposition aux usagers ponctuels, disposant d'un abonnement soit hebdomadaire, soit journalier.

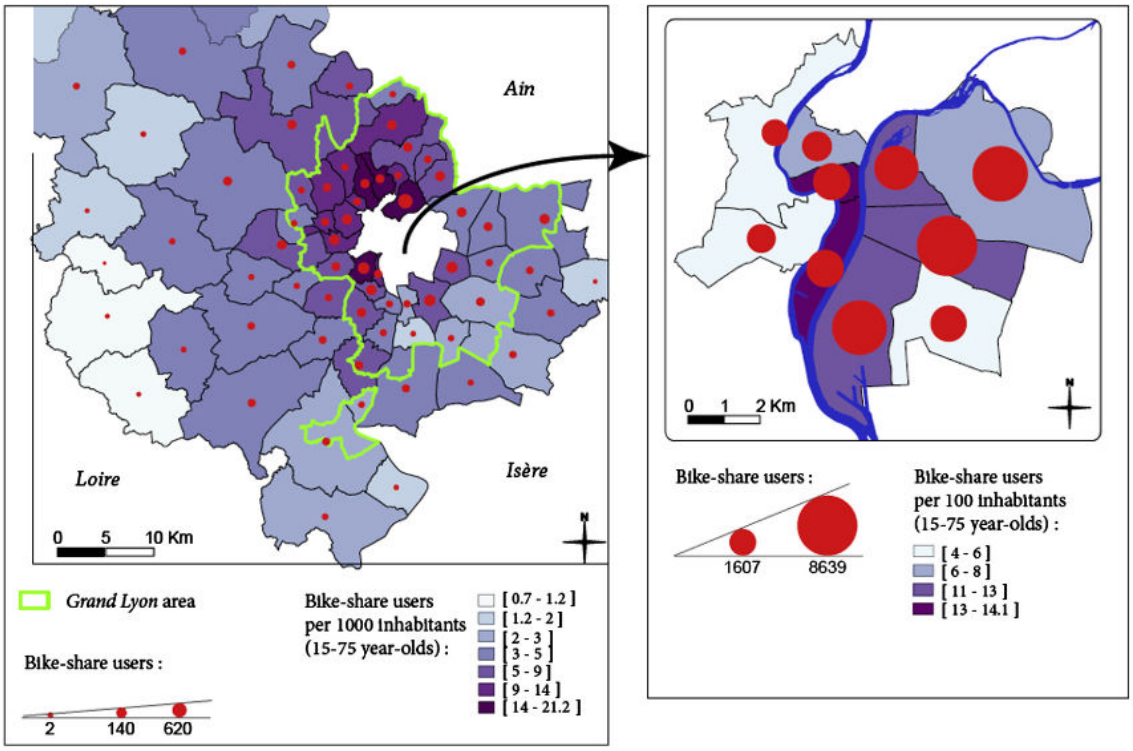


FIGURE 1.7 – Répartition des abonnés Vélo'v dans l'aire urbaine du Grand Lyon et au-delà. (Extraite de [240])

Âge	% parmi les abonnés Vélo'v	% parmi les utilisateurs de TCL	% parmi la population du Grand Lyon
5-17	2.7	20.1	18
18-24	28.9	21.2	11.1
25-34	31.1	16.9	14.9
35-49	23.7	17.7	22.1
50-64	11.8	14.6	20.8
65+	1.7	9.5	13.1

TABLE 1.4 – Comparaison des effectifs par âge des abonnés Vélo'v avec les abonnés TCL et la population du Grand Lyon, pour l'année 2011.

qu'il ne représente que 26 % de la population du Grand Lyon. L'âge médian des abonnés est de 30 ans, ce qui confirme la relative jeunesse des utilisateurs du système Vélo'v.

La répartition géographique diffère également suivant le genre des abonnés : les abonnés femmes vivent en majorité dans le centre-ville de Lyon, et sont moins nombreuses à vivre en dehors du centre. Cela peut s'expliquer à la fois par l'importante mobilité des hommes âgés de 35 à 49 ans, plus enclins à parcourir de longue distance, ainsi que par le fait que les étudiants, en majorité masculins, enregistrent généralement l'adresse de leurs parents vivant en dehors du centre-ville lors de leur inscription. Dans le centre du Grand Lyon, les femmes sont plus nombreuses dans les 1er et 7e arrondissements. Certains quartiers sont en revanche très déséquilibrés : l'exemple de Villeurbanne, où 60 % des abonnés sont des hommes, est révélateur, et peut s'expliquer une nouvelle fois par la présence d'un campus scientifique où le nombre d'étudiants masculins est plus élevé par rapport à la population de la ville.

3.1.3 Comparaison avec les usagers des transports en commun

L'abonnement Vélo'v peut être couplé avec d'autres cartes d'abonnement de transport en commun, permettant d'estimer la proportion des abonnés qui utilisent les autres moyens de transport. Ainsi, 52.4 % des abonnés combinent leur abonnement Vélo'v avec un abonnement TCL, 4.6 % avec un abonnement TER (trains régionaux), et seulement 1 % avec une carte Parking. Ces chiffres, bien qu'approximatifs – combiner un abonnement Vélo'v avec un autre abonnement ne requiert pas, d'un point de vue technique, l'utilisation de cet autre abonnement et réciproquement, l'utilisation d'un autre moyen de transport peut se faire sans nécessairement combiner l'abonnement avec son abonnement Vélo'v – permet toutefois de se rendre compte que les abonnés Vélo'v sont en majorité des personnes enclines à utiliser les transports en commun, en complément ou au détriment du partage de vélos.

Afin de caractériser les abonnés longue-durée, une classification des utilisateurs disposant d'un abonnement annuel est proposée, de manière à établir une typologie des usagers basée sur l'intensité et la régularité de leur pratique. Ce travail à fait l'objet d'une publication dans la revue *Journal of Transport Geography* grâce à une collaboration avec Julien Barnier, Isabelle Mallon et Marie Vogel du centre Max Weber, Luc Merchez du laboratoire Environnement, Ville, Société, et Patrice Abry et Guillaume Lozenguez du laboratoire de Physique [240].

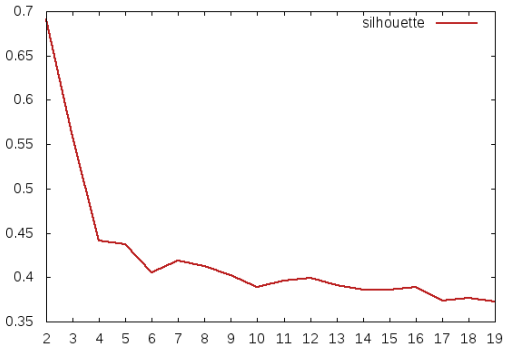
3.2 Définition d'un profil par usager

Un vecteur d'attributs, appelé profil, est attribué à chaque utilisateur, et est construit de manière à refléter l'intensité et la régularité de sa pratique sur la semaine et pendant l'année. Pour cela, 21 attributs sont définis, les 8 premiers correspondant à l'activité hebdomadaire alors que les 13 autres correspondent à l'activité annuelle. Le profil de l'utilisateur i, noté $\boldsymbol{x}^i = [x_1^i, \ldots, x_{21}^i]$ est calculé de la façon suivante :

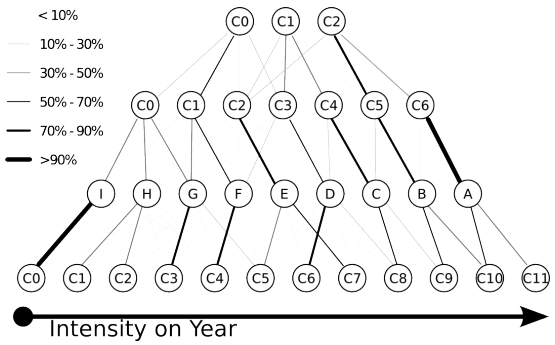
- x_1^i : nombre moyen de trajets réalisés par semaine, calculé sur toutes les semaines pendant lesquelles l'utilisateur i a réalisé au moins un trajet ;
- $x_2^i, ..., x_6^i$: nombre moyen de trajets réalisés les jours de semaine, calculé sur les jours pendant lesquels l'utilisateur i a au moins un trajet, et triés par ordre croissant d'intensité ;
- x_7^i : nombre moyen de trajets réalisés le samedi, calculé sur les jours pendant lesquels l'utilisateur i a au moins un trajet ;
- x_8^i : nombre moyen de trajets réalisés le dimanche, calculé sur les jours pendant lesquels l'utilisateur i a au moins un trajet ;
- x_9^i : nombre total de trajets réalisés dans l'année par l'utilisateur i ;
- $x_{10}^i, ..., x_{21}^i$: nombre de trajets réalisés par l'utilisateur pour chaque mois de l'année, triés par ordre croissant d'intensité.

Du fait du tri des jours de la semaine et des mois par ordre croissant d'intensité, les profils sont définis sans prendre en compte quels jours et quels mois sont les plus intenses, mais quelle est la répartition de l'intensité de la pratique sur la semaine et l'année. Seule une distinction entre jours de la semaine et week-end est faite, l'utilisation étant sensiblement différente pendant ces deux périodes. La matrice des profils est notée $\boldsymbol{X} = (\boldsymbol{x_i})_{i=1,...,n}$, avec n le nombre d'utilisateur, égal à 50480 dans notre étude.

Les attributs sont normalisés de la façon suivante : x_1 et x_9, représentent l'intensité de la pratique sur la semaine et sur l'année, et sont normalisés par 1.5 fois la distance interquartile sur tous les utilisateurs, qui est une façon classique en statistique de tenir compte des valeurs aberrantes dans la normalisation. Les autres attributs représentent la régularité de la pratique et sont normalisés de manière à ce que la somme des jours de la semaine (variables $x_2, \ldots, x_8$) et la somme des mois (variables $x_{10}, \ldots, x_{21}$) soient égales à 1 Ce choix de normalisation est discuté et comparé avec d'autres méthodes dans l'Annexe B.



(a) Silhouette moyenne des classes obtenues après application d'un algorithme à K-moyennes, pour K variant de 2 à 20



(b) Hiérarchie des classifications en explorant les filiations entre les classes, pour 3, 7, 9 et 12 classes.

FIGURE 1.8 – (Extraites de [240]).

3.3 Visualisation des profils dans le plan factoriel

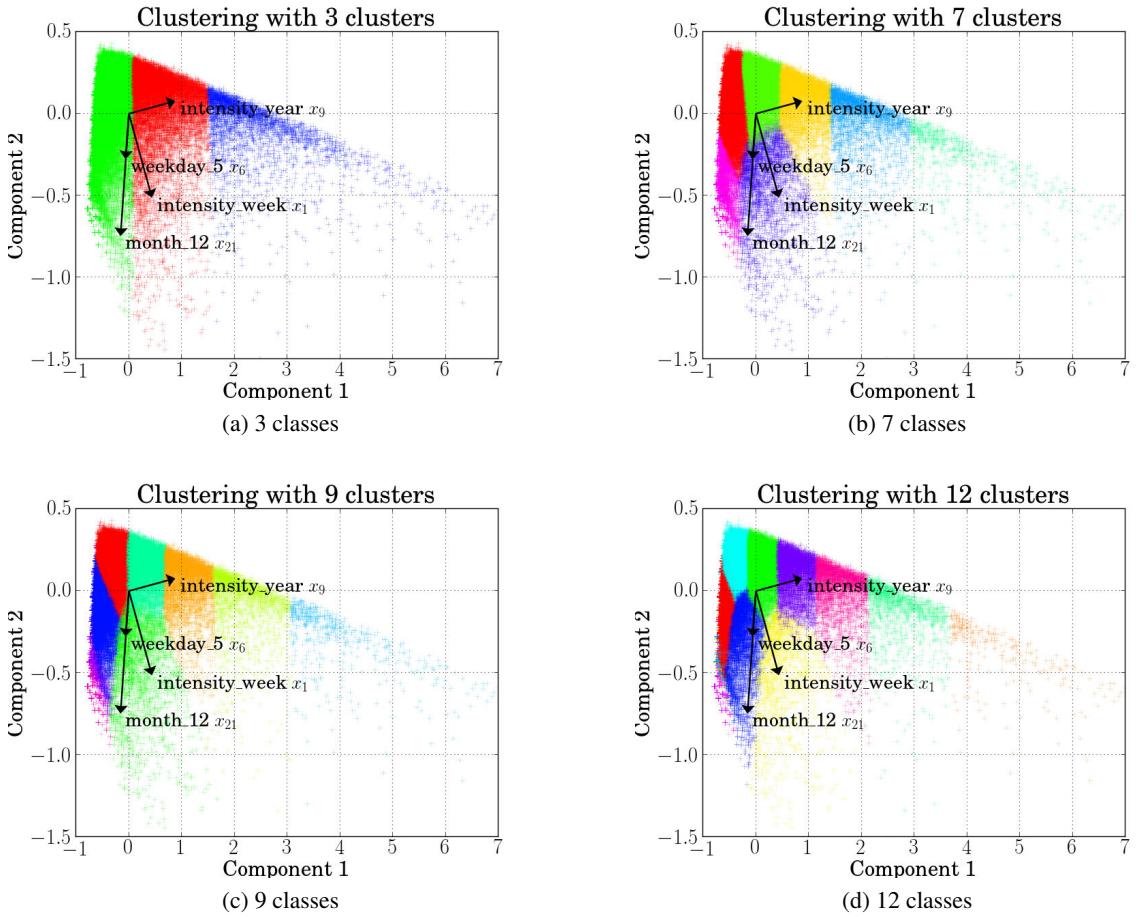
Ces profils sont visualisés dans le premier plan factoriel obtenu après une ACP (analyse en composantes principales) de la matrice $\boldsymbol{X}$ (voir la Figure 1.9), les deux premiers axes expliquant 85 % de la variance totale des données. Les attributs 1 et 9 (intensité sur la semaine et sur l'année) sont dominants pour la première composante, alors que les attributs 6 et 21 (pourcentage de mouvements réalisés sur le jour le plus intense et le mois le plus intense) constituent avec l'attribut 1 le second axe. Cette visualisation permet de mettre en évidence des corrélations entre les attributs représentant l'intensité et ceux dérivant la régularité.

Ces corrélations ne sont néanmoins pas gênantes pour la suite : une méthode des K-moyennes [158] est utilisée, couplée avec une évaluation statistique et une analyse des résultats minutieuse. Étant donné que l'objectif de ce travail est de créer et d'interpréter une typologie des utilisateurs, sans forcément trouver une classification pré-existante et bien-définie, il est préférable, pour des raisons d'interprétation, de garder les attributs originaux malgré leur corrélation, plutôt que d'utiliser les variables obtenues après une ACP, décorrélées mais plus complexes à interpréter.

3.4 Choix du nombre de classes

Le choix du nombre de classes se fait à posteriori, en comparant les classifications obtenues, pour un nombre de classes K variant de 2 à 20. La comparaison s'effectue en regardant les silhouettes de chaque classification [208], et plus particulièrement la silhouette moyenne : plus la valeur de la silhouette moyenne est élevée, plus la classification est supposée bonne. En calculant la valeur de la silhouette moyenne pour K variant de 2 à 20, on observe sur la Figure 1.8a comme attendu que d'une part la valeur décroît en fonction de K, et d'autre part qu'il n'y a pas de coupure franche entre différentes valeurs de K, ce qui confirme l'absence de typologie pré-existante dans les données. À la lecture du graphique, les valeurs de K les plus adaptées seraient 2 ou 3. Cependant, ce faible nombre de classes ne nous permet pas d'obtenir une typologie suffisamment riche pour être exploitable. Les valeurs de 3, 7, 9, 12, et 16 semblent être des valeurs pertinentes à retenir, dont seules les 4 premières sont retenues afin de ne pas tomber dans l'excès inverse en complexifiant inutilement l'interprétation.

Afin de choisir parmi ces quatre valeurs, des partitions des données en 3, 7, 9 et 12 classes sont comparées, en explorant les filiations entre les classes, c'est-à-dire en regardant quel pourcentage d'utilisateur dans une classe pour une valeur de K donnée se retrouve dans une classe obtenue avec une valeur de K plus faible. Cette hiérarchie entre les classes mène à une représentation multi-classification présentée à la Figure 1.8b. Sur la Figure 1.9 se trouve les classes obtenues pour un choix du nombre de classes



(a) 3 classes (b) 7 classes

(c) 9 classes (d) 12 classes

FIGURE 1.9 – Classes obtenues pour différents choix du nombre de classes, affichées en couleur dans le plan factoriel principal de la matrice $\boldsymbol{X}$. (Extraite de [240]).

fixé respectivement à 3, 7, 9 et 12. Les classes sont affichées en couleur dans le plan factoriel principal de la matrice $\boldsymbol{X}$. D'après ces résultats, le choix retenu est de 9 classes. En effet, le choix de 3 classes réduit la typologie à une description des utilisateurs à forte, moyenne et faible intensité, sans autre distinction, alors que le choix de 7 classes ne permet pas de faire de différentiation entre les utilisateurs avec une faible activité. D'un autre côté, étendre la classification à 12 classes permettrait une différentiation entre les utilisateurs actifs le samedi et ceux actifs le dimanche, avec cependant un coût d'interprétation plus élevé, du fait de deux nouvelles classes. Le choix de 9 classes semble ainsi un choix raisonnable, établissant un bon compromis entre richesse de la typologie et simplicité de l'interprétation des classes.

3.5 Analyse de la typologie obtenue

Les profils moyens des 9 classes obtenues sont affichés sur la Figure 1.10. Les classes sont triées par ordre décroissant de l'intensité de l'usage, et nommés de A à I. Elles sont d'abord différentiées par l'intensité de l'usage selon l'année. Les classes A, D et F sont complètement différentiées par l'intensité de la pratique, avec un profil similaire pour la régularité de la pratique sur la semaine (jours de semaine plus intense que les jours de week-end) ainsi qu'un profil équilibré de l'intensité sur les mois. Ces profils regroupent ainsi des utilisateurs réguliers à la fois dans la semaine et dans l'année. Au contraire, la classe I décrit des usagers dont le motif d'utilisation du partage de vélos couvre seulement un jour de la semaine par semaine, alors que la classe H contient des utilisateurs actifs surtout les jours de week-end. Enfin,

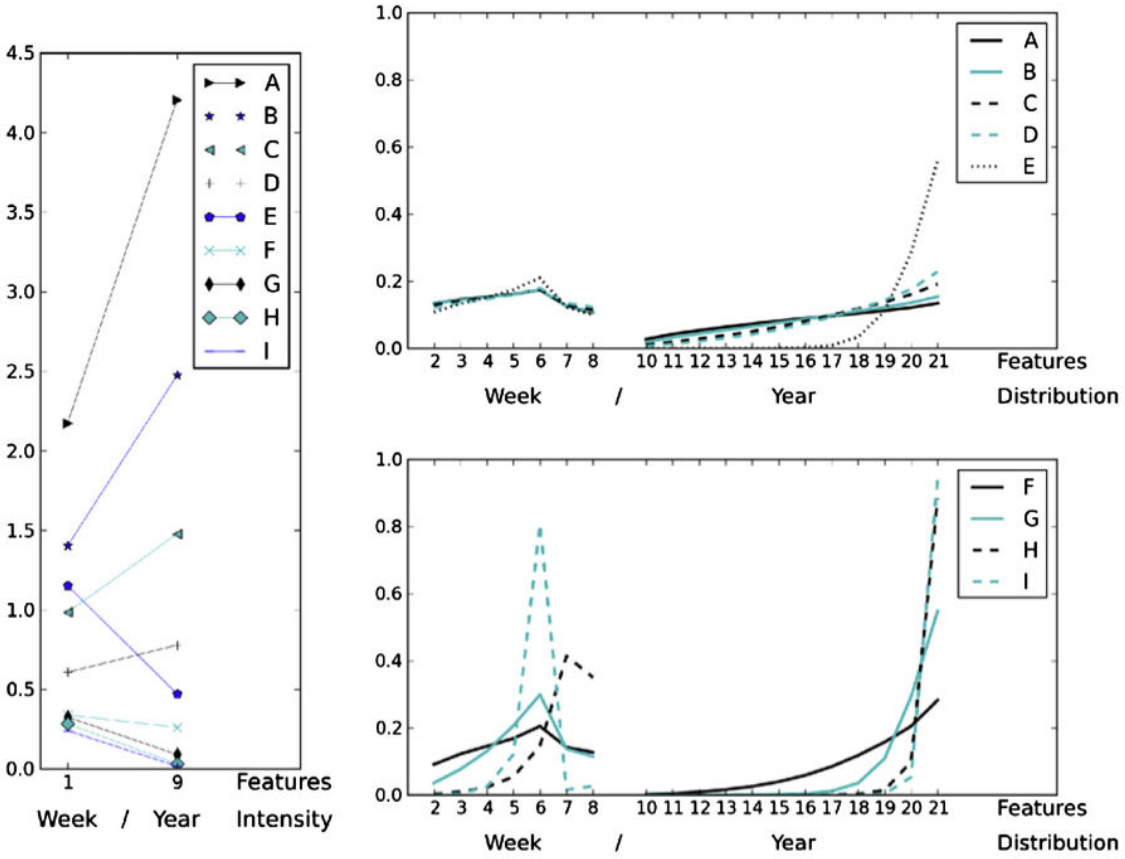


FIGURE 1.10 – Profils moyens des 9 classes obtenues. Les classes sont triées par ordre décroissant de l'intensité de l'usage, et nommés de A à I. (Extraite de [240]).

les classes G et E regroupent des utilisateurs avec une faible régularité sur l'année, que ce soit avec une intensité moyenne (classe E) ou faible (classe G).

Une description de chacune des classes est donnée dans la Table 1.5. Tout d'abord, le contraste entre le nombre de trajets réalisés dans l'année apparaît clairement entre les classes, de 2.8 trajets par an en moyenne pour les utilisateurs de la classe I à plus de 690 pour ceux de la classe A. Les quatre classes les plus actives représentent 29 % des utilisateurs et comprend pour presque 75 % des trajets. Au contraire, les classes H et I regroupent 10 % des utilisateurs mais totalisent moins de 0.5 % des trajets : ces utilisateurs ont ainsi très peu utilisé leur abonnement (classe I) ou bien de manière sporadique, principalement le week-end (classe H). Deuxièmement, la distribution des trajets sur la semaine et sur l'année permet de distinguer des comportements allant d'une utilisation unique du partage de vélos vers une utilisation quotidienne. La classe E illustre l'intérêt de prendre en considération des variables représentant la régularité de la pratique : bien que l'intensité de la pratique soit assez élevée, l'activité est concentrée sur une petite partie de l'année.

3.6 Discussions sur les utilisateurs Vélo'v

Bien que la classification obtenue mette en évidence 9 classes d'utilisateurs, il est possible de simplifier l'interprétation de ces classes en les regroupant elles-mêmes en 4 catégories, qui mettent en avant les forts contrastes dans la pratique du vélo en libre-service. En utilisant des variables telles que le genre, l'âge ou le code postal, ainsi que d'autres attributs liés à l'utilisation, une description sociologique des catégories peut être réalisée. De plus, en intégrant à ces 9 classes d'abonnés les caractéristiques des abon-

C	$\|C\|$	N_u	N_m	$R_{H/F}$	A_m	N_t	D_F	D_H
A	526 (1%)	693.8	11.4	0.25	31	364 965 (8.4 %)	0.47	1.49
B	2029 (4%)	408.1	11.1	0.38	28	828 018 (2.53 %)	2.53	5.19
C	4288 (8%)	243.8	10.3	0.55	28	1 045 328 (24.0%)	6.80	9.83
D	7925 (16%)	128.3	9.8	0.66	31	1 017 166 (23.3%)	14.2	16.9
E	2790 (6%)	77.7	3.3	0.61	24	216 830 (5.0 %)	4.76	6.13
F	16 250 (32%)	43.1	7.9	0.82	30	699 701 (16.0%)	32.8	31.7
G	11 509 (23%)	14.9	3.3	1.03	30	171 121 (0.3%)	26.3	20.0
H	2586 (5%)	5.1	1.5	1.12	32	13 210 (0.3%)	6.15	4.31
I	2577 (5%)	2.8	1.2	1.08	36	7162 (0.2%)	6.01	4.39
Total	50 480	86.44	6.6	0.79	30	4 363 501 (100%)	100	100

TABLE 1.5 – Description des classes obtenues. C : Classe de l'utilisateur ; $|C|$ Nombre d'utilisateurs ; N_u Nombre moyen de trajet par utilisateur ; N_m : Nombre moyen de mois actifs par utilisateur ; $R_{F/H}$: Ratio Femmes/Hommes ; A_m : Âge médian ; N_t : Nombre total de mouvements ; D_F : Répartition des femmes (%) ; D_H : Répartition des hommes (%).

nés hebdomadaires et journaliers décrites dans la Section 3.1.1, il est possible d'avoir une discussion sur l'ensemble des utilisateurs du système Vélo'v. Ces quatre catégories sont les suivantes :

Utilisateurs extrêmes Cette catégorie regroupe les utilisateurs des classes A et B, dont la pratique du VLS est intensive avec une grande régularité sur la semaine et sur l'année. Elle regroupe moins de 5 % des utilisateurs mais représente 27.4 % des trajets. Les utilisateurs sont clairement masculins (ratio femmes/hommes de 0.25) et ont un âge médian de 31 ans. Ces observations – les utilisateurs sont le plus souvent des hommes jeunes – sont cohérentes avec celles réalisées sur les cyclistes en France [121], bien que moins prononcées. L'absence d'informations plus précises ne permet pas de les caractériser plus finement, et notamment de savoir, comme précédemment établi pour des cyclistes intensifs [133], s'ils possèdent une voiture et si la pratique du vélo relève d'une conviction personnelle. Néanmoins, leur degré de pratique (chaque utilisateur fait en moyenne entre 400 et 700 trajets par an) laisse penser qu'ils utilisent exclusivement Vélo'v pour leurs déplacements, ceux-ci étant réguliers sur l'année, malgré des conditions météorologiques parfois difficiles à Lyon en hiver.

Utilisateurs assidus Cette catégorie regroupe les utilisateurs intensifs et réguliers des classes C et D (environ 30 % des utilisateurs). La classe E est également ajoutée à ce groupe, étant donné que les utilisateurs de cette classe sont aussi actifs que ceux de la classe C (23 trajets par mois), malgré une pratique concentrée sur quelques mois de l'année. Ces utilisateurs ont un usage établi du système pendant l'année (entre 128 et 243 trajets par an suivant la classe) ou pendant une courte période de l'année. Sociologiquement, cette catégorie est plutôt masculine (le ratio femmes/hommes est autour de 0.6) et comprend deux groupes d'âge, le premier avec un âge médian de 24 ans, regroupant principalement des usagers de la classe E, et le deuxième avec un âge médian autour de 30 ans, regroupant principalement les utilisateurs des classes C et D. L'âge médian relativement jeune par rapport aux autres classes des utilisateurs du premier groupe, et la brièveté de la période d'utilisation, laisse à penser que ces utilisateurs sont des étudiants arrivés en cours d'année à Lyon, ce qui est confirmé par le fait que 47 % d'entre eux ont fait leur premier trajet pendant le dernier trimestre de l'année.

Utilisateurs multimodaux Cette catégorie regroupe les utilisateurs de la classe F. L'intensité de leur pratique du Vélo'v est faible mais régulière, ils effectuent en moyenne 43 trajets par an, distribués sur 8 mois. Elle regroupe 32 % des abonnés annuels, 31 % des hommes et 32 % des femmes, avec

un ratio femmes/hommes de 0.82 très proche de celui observé sur l'ensemble des abonnés. Cette catégorie correspond ainsi à une classe des utilisateurs standards du système Vélo'v. Il est ainsi possible de prendre comme hypothèse que ces utilisateurs utilisent le partage de vélos comme un moyen de transport parmi d'autres, que ce soit en complément d'un autre (multimodalité) ou en remplacement (intermodalité).

Utilisateurs sporadiques Cette catégorie est composée des utilisateurs de la classe G, regroupant les comportements avec une pratique faible du partage de vélos et étalée sur 3 mois de l'année, ceux de la classe H, avec une activité faible principalement le week-end, et ceux de la classe I, regroupant les utilisateurs avec seulement 1 ou 2 trajets dans l'année. Elle regroupe environ 16 600 personnes (33 % du total), qui sont pour la plupart des femmes (le ratio femmes/hommes est égal à 1.05). Le faible coût de l'abonnement annuel explique en partie le grand effectif de cette classe, composée d'utilisateurs curieux du système mais dont la période de découverte n'a pas mené à une utilisation ultérieure.

À cette typologie des abonnés annuels, il est possible d'ajouter une cinquième catégorie, regroupant les abonnés hebdomadaires et journaliers. Les pratiques des utilisateurs titulaires d'un abonnement sur la semaine peuvent être reliées à celles des membres de la catégorie des utilisateurs sporadiques, et plus précisément de la classe G, alors que les abonnés journaliers ont une activité concentrée sur le week-end et les périodes nocturnes.

Les catégories sont genrées de manière très distinctes : les fortes intensités de pratique se retrouvent en grande partie chez des individus masculins, alors que les catégories avec une intensité moyenne présentent une répartition beaucoup plus équitable entre individus masculins et féminins. Au contraire, la catégorie des utilisateurs avec une intensité de pratique faible voire inexistante est plutôt féminine. Il n'y a pas de grandes différences d'âge entre les différentes catégories, à l'exception des sous-groupes composés des utilisateurs de la classe E, qui sont significativement plus jeunes (24 ans) que la moyenne des utilisateurs, et ceux de la classe I, qui sont un peu plus âgés (36 ans).

Les catégories d'usagers peuvent également être regardées spatialement dans l'espace géographique lyonnais, à l'aide de cartes de répartition des groupes par arrondissement, qui ne sont visibles ici que pour les catégories « Utilisateurs multimodaux » et « Utilisateurs extrêmes » à la Figure 1.11. La première chose à noter sur ces cartes est que la résidence n'est pas un facteur pertinent pour distinguer spatialement les catégories : il n'y a pas de sur-représentation d'un ou plusieurs groupes dans un arrondissement en particulier, ce qui signifie que l'utilisation du Vélo'v n'est pas reliée au lieu d'habitation de l'usager. Quelques exceptions apparaissent néanmoins, par exemple pour les utilisateurs extrêmes et assidus, dont la présence est relativement plus nombreuse dans les 2e, 3e, 6e et 7e arrondissements ainsi qu'à Villeurbanne. Cette observation peut néanmoins être contestée, du fait de la grande taille de ces arrondissements qui ne permet pas une analyse fine de la répartition spatiale des usagers. Les déplacements dans l'espace ne permettent pas non plus de distinguer plusieurs catégories d'utilisateurs. Si l'on se concentre sur les flux agrégés au niveau des arrondissements, il n'y a pas de différence significative entre les quatre catégories. Des différences apparaissent néanmoins lorsque l'on compare avec les abonnés journaliers : les déplacements effectués par ces abonnés courte-durée semblent répartis de manière plus homogène, que ce soit les flux à l'intérieur d'un arrondissement ou les flux entre les arrondissements.

De ce point de vue, Lyon contraste avec la situation de Londres, où la répartition spatiale des usagers est beaucoup plus hétérogène [22], notamment entre les zones résidentielles et les zones commerciales et industrielles.

Cette typologie en 4 ou 5 classes rappelle celles proposées par d'autres auteurs [133, 42, 192, 76] sur des typologies de cyclistes, dans lesquelles l'intensité et la régularité de la pratique sur l'année soulignaient les différents publics dans l'utilisation du vélo comme moyen de transport. Dans ces classifications, les utilisateurs des systèmes VLS s'intègrent dans les groupes des cyclistes occasionnels, qui

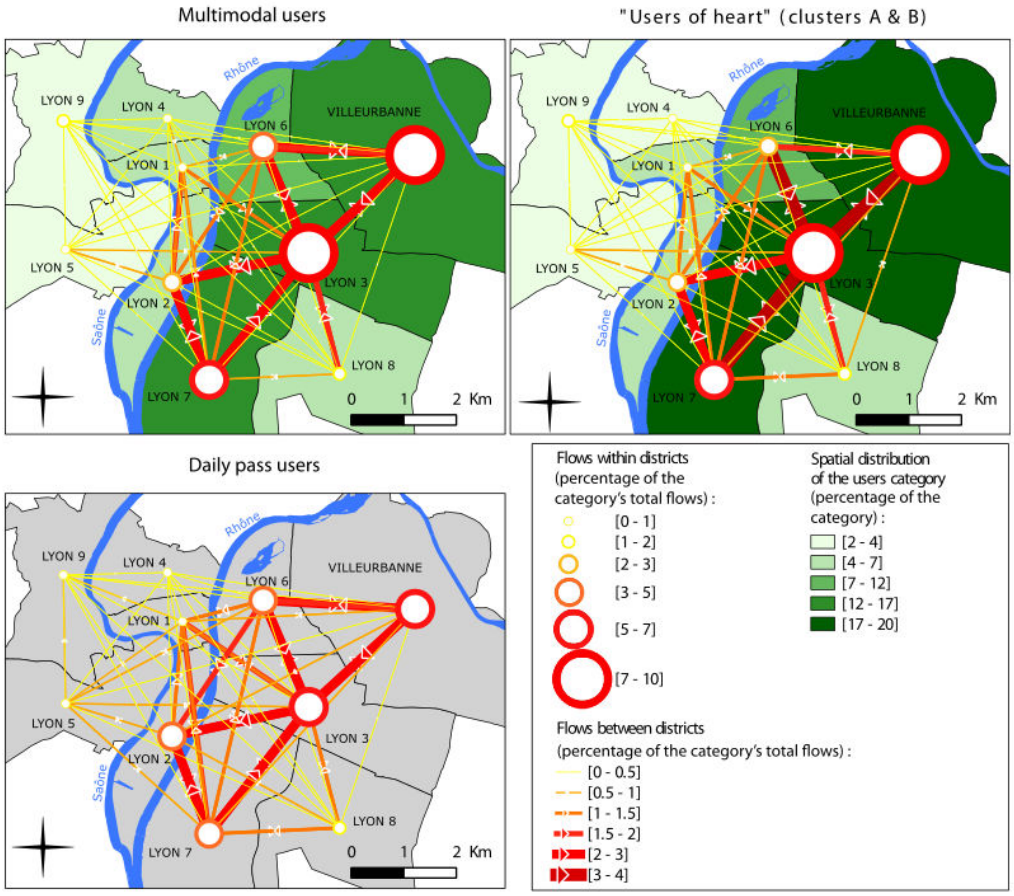


FIGURE 1.11 – Flux de déplacements entre chaque arrondissements pour trois catégories d'usagers. Lorsque l'information est disponible, la répartition spatiale des usagers est également représentée. La catégorie « Multimodal users » correspond à « Utilisateurs multimodaux », la catégorie « Users of heart » à « Utilisateurs extrêmes » et enfin les « Daily pass users » correspondent aux utilisateurs journaliers. (Extraite de [240]).

utilisent le vélo comme un moyen de transport parmi les autres, et n'hésitent pas à changer selon les conditions.

3.7 Limites et discussions

En accédant à des bases de données sur les utilisateurs et leurs déplacements, une étude sociologique a pu être mise en place en utilisant des techniques classiques d'analyse de données. Cette approche contraste avec les approches traditionnelles en sociologie, dans lesquelles les études sont essentiellement qualitatives.

Plusieurs limites de cette étude peuvent être soulevées. Tout d'abord, la composition du jeu de données fait que les informations démographiques sur les utilisateurs Vélo'v ne concernent que les abonnés annuels, qui représentent 67 % des utilisateurs. Malgré l'extension de la typologie aux abonnés journaliers et hebdomadaires, il n'est pas possible de connaître la régularité et l'intensité de la pratique des ces utilisateurs : de nombreux indices laissent ainsi penser qu'ils ont une utilisation différente du système.

La deuxième limitation de cette étude tient à la nature des données : s'il a été possible de caractériser finement les schémas de la pratique des VLS pour les abonnés annuels, ces données ne donnent pas les motivations liées à l'utilisation du vélo en libre-service. Il est possible ,à partir de l'heure et de la

répartition géographique d'un trajet, d'émettre des hypothèses réalistes sur sa nature, c'est-à-dire si c'est un trajet plutôt professionnel ou plutôt pour des loisirs. Cette approche est néanmoins insuffisante, et ne donne de toute façon aucune indication sur les intentions de l'utilisateur, à savoir pourquoi il a utilisé le vélo en partage plutôt qu'un autre moyen de transport et inversement, pourquoi il n'utilise pas le système Vélo'v lorsqu'il se déplace avec d'autres moyens de transport.

Cette étude est la première étape d'un programme de recherche plus large, permettant la mise en place d'entretiens qualitatifs avec des usagers et des non-usagers du système Vélo'v. Les résultats de la typologie obtenue permettent de concevoir d'une part les questionnaires, mais également les profils sociologiques des personnes interrogées. Ces entretiens vont ainsi permettre de valider les hypothèses obtenues à partir de la typologie sur les mécanismes entraînant l'utilisation ou non du système Vélo'v.

4 Classifications des stations par les trajets

Un thème de recherche fécond dans la littérature sur les systèmes de vélos en libre-service est la classification des stations [99, 147, 61]. Il est en effet intéressant de pouvoir regrouper des stations, en se basant sur des indicateurs tels que les flux entrants et sortants, ou sur le nombre de vélos disponibles : en cherchant des caractéristiques communes aux stations d'un même groupe, l'objectif est alors de trouver les causes qui expliquent le comportement observé des stations.

Dans cette section, deux approches de classification de stations sont abordées, issues de travaux toujours en cours de réalisation. Le premier exemple détaille une méthode pour obtenir un profil d'activité pour chaque station, ainsi qu'un résultat de classification à partir de ces profils. Le deuxième exemple utilise la notion de réseau pour décrire le système Vélo'v, permettant de tirer profit des méthodes de la théorie des réseaux.

4.1 Contrainte de capacité

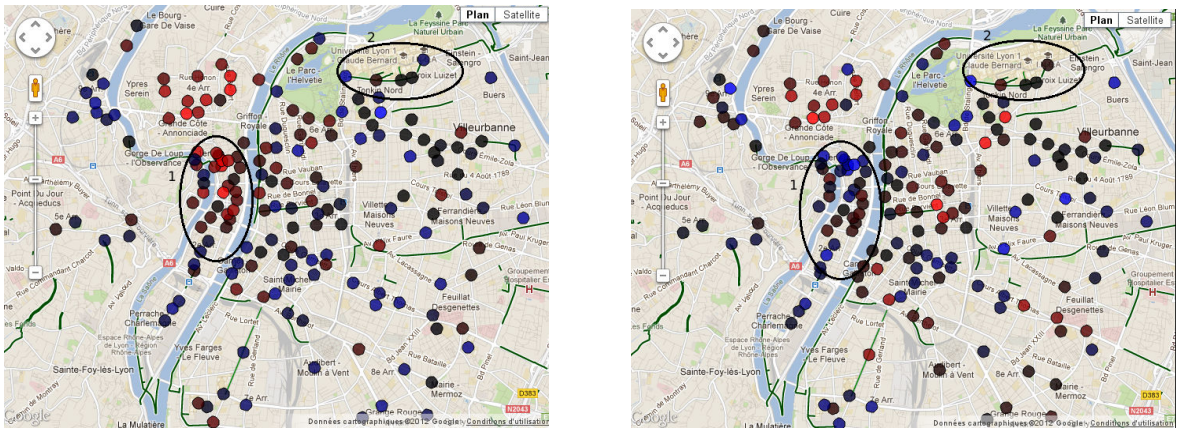
Du fait de l'inégale répartition géographique des activités et des populations, les systèmes VLS sont sujets à des effets indésirables qui perturbent leur bon fonctionnement : certaines zones de la ville, pour des raisons socio-économique et géographique, ont tendance à attirer et à retenir les gens, alors que d'autres sont au contraire des zones que les usagers quittent. Une des manifestations les plus évidentes de ces dysfonctionnements est la présence de stations vides ou pleines : les stations étant de capacité limitée, les stations les plus attractives se remplissent, n'offrant plus d'espace disponible pour déposer un vélo, alors que les stations les moins attractives se vident. La régulation des systèmes de vélos en libre-service est ainsi un enjeu majeur dans la maintenance et l'exploitation de tels systèmes, mais qui est complexe à cause de l'opposition entre les différents objectifs. Du point de vue des usagers, la régulation doit permettre d'éviter de se retrouver devant une station sans vélo lorsque l'on souhaite en louer un, ou bien dans un cas plus problématique, de se retrouver avec un vélo dans une station pleine sans possibilité de le déposer. Du point de vue de l'exploitant, l'objectif est de pouvoir mettre en place une régulation peu coûteuse, tout en garantissant un fonctionnement minimal du système.

Une approche descriptive du problème de la régulation est proposée dans l'Annexe C, qui discute d'une méthode de détection des moments pendant lesquels la contrainte de capacité joue, c'est-à-dire que la station est soit vide, soit pleine. Ces moments peuvent être facilement obtenus si l'on dispose de données sur la disponibilité des stations, ce qui n'est pas le cas pour les années avant 2013. De plus, ces données peuvent ne pas donner des résultats précis, si l'on considère par exemple le cas dans lequel la station contient un unique vélo hors d'usage : bien que ce vélo puisse être emprunté, et que la station n'est pas considéré comme vide, elle l'est dans les faits.

Il est intéressant d'utiliser ces informations pour réaliser des indicateurs statistiques afin de quantifier l'importance de la contrainte de capacité de manière à la fois globale sur l'ensemble du système et

État	7h-9h	12h-14h	18h-20h	Total
Vide (en %)	11	11	10	7
Plein (en %)	10	6	6	6

TABLE 1.6 – Pourcentage moyenné sur les stations du temps pendant lequel le système est en contrainte de capacité (c'est-à-dire soit pleine, soit vide) pour différentes périodes de la journée des jours de semaine.



(a) Jours de semaine de 7h à 9h

(b) Jours de semaine de 18h à 20h

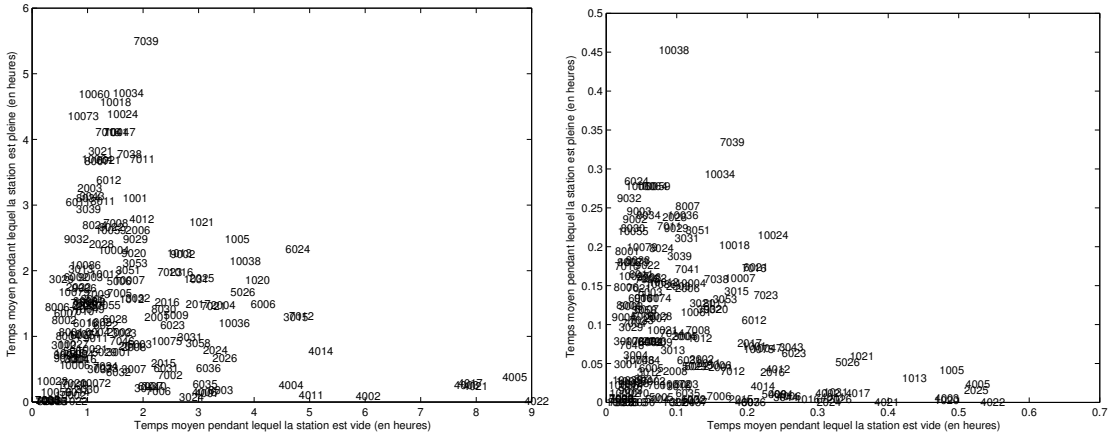
FIGURE 1.12 – Carte de Lyon avec affichage des stations Vélo'v : la couleur du cercle indique le type de contrainte de capacité qui domine (bleu : station pleine, rouge : station vide). Plus la couleur est prononcée, plus le temps pendant lequel la contrainte de capacité a agi est important. La zone 1 délimite les stations de la Presqu'île, la zone 2 le campus de la Doua.

également station par station. Elles permettent également d'établir des profils de stations, et de pouvoir regrouper les stations entre elles en fonction de leur comportement vis-à-vis de cette contrainte.

4.1.1 Statistiques sur l'ensemble du système

La Table 1.6 regroupe les pourcentages du temps pendant lequel le système a été en contrainte de capacité pour les jours de semaine. On remarque que sur les périodes du matin, du midi et du soir, la contrainte de capacité agit plus souvent que sur le reste de la journée.

L'information peut également être affichée sur une carte afin de se rendre compte de l'évolution spatiale de ces contraintes de capacité. Les Figures 1.12a et 1.12b représentent ces temps sur une carte pour les jours de semaine, respectivement de 7h à 9h, et de 18h à 20h. Les stations sont représentées par des cercles de couleur, indiquant le type de contrainte de capacité qui domine (bleu : station pleine ; rouge : station vide). Plus la couleur est prononcée, plus le temps pendant lequel la contrainte de capacité a agi est important. La zone 1 notée sur les cartes indique les stations de la Presqu'île. On remarque une couleur rouge dominante le matin, indiquant une majorité de stations vides, et une couleur bleue le soir, indiquant une majorité de stations pleines. La zone 2 indique les stations proches du campus de la Doua. On remarque des stations plutôt pleines le matin et plutôt vides le soir correspondant aux trajets « Domicile - École ». Une des raisons qui pourraient expliquer la faible intensité des couleurs est que le campus comprend des zones d'habitation, notamment des résidences étudiantes, qui font que les flux ne sont pas que dans un sens le matin et le soir. On retrouve des comportements similaires à ceux trouvés dans [165].



(a) Journée (b) Matin

FIGURE 1.13 – Affichage des stations suivant leurs temps pendant lesquels les stations sont vides ou pleines, pour deux périodes différentes

4.1.2 Statistiques sur les stations

Profils statiques des stations Un profil statique pour chaque station peut être réalisé en considérant les temps moyens pendant lesquels les stations sont pleines et vides. La Figure 1.13a trace en fonction de ces attributs les stations, permettant de mettre en évidence trois comportements différents des stations :

— des stations très souvent vides et rarement pleines en bas à droite du graphique ;
— des stations très souvent pleines et rarement vides en haut à gauche du graphique ;
— des stations équilibrées en bas à gauche.

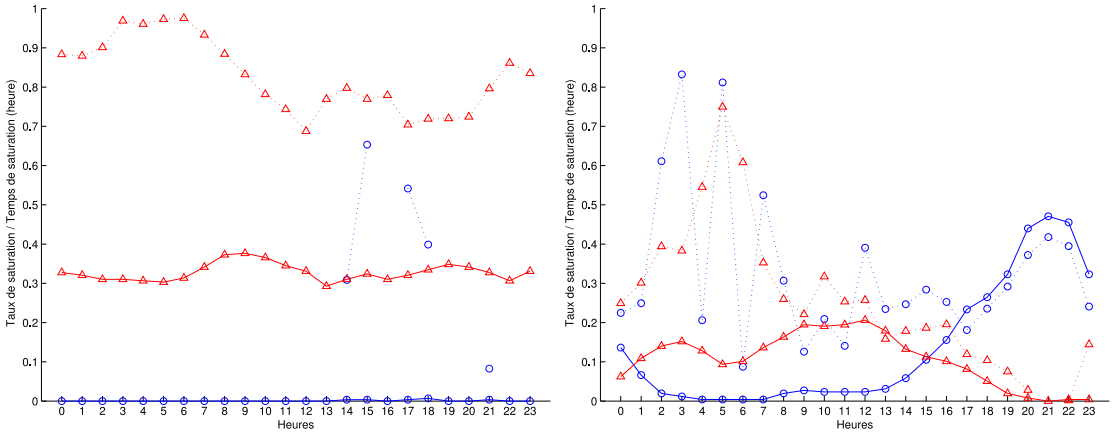
L'absence de station en haut à droite du graphique indique que les stations ne sont jamais à la fois très souvent vides et très souvent pleines.

Un graphe similaire peut être tracé en se concentrant sur la période du matin (Figure 1.13b). Si l'on analyse plus attentivement ce dernier graphe, deux groupes de station se distinguent pour être très souvent vides le matin (autour de 50 % de temps pendant lequel la contrainte de capacité a été activée) : un groupe de stations se situant sur la colline de la Croix-Rousse, dû au fait que ces stations sont en altitude et ont du mal à se remplir, et un groupe de stations du centre-ville, qui peut s'expliquer par la présence d'une forte densité de population à cet endroit de la ville avec une population qui a tendance à se déplacer le matin pour motifs professionnels ou scolaire. Inversement, les stations qui sont très souvent pleines se situent près des campus universitaires.

Profils dynamiques des stations Des profils dynamiques des stations peuvent être réalisés pour chaque station en calculant, pour des intervalles de 1h, les valeurs suivantes :

— la proportion de jours pendant lesquels la station s'est retrouvée soit dans l'état vide, soit dans l'état plein ;
— la proportion du temps pendant lequel la station est restée sous contrainte de capacité lorsque la station était pleine ou vide.

Concrètement, on obtient des profils comme ceux affichés sur la Figure 1.14 pour deux stations. Pour chaque intervalle d'une heure, la courbe pleine affiche la proportion de jours pendant lesquels la station s'est retrouvée soit dans l'état vide, soit dans l'état plein (par exemple, une valeur de 0.5 indique que pour la moitié des jours pris en compte dans l'étude, la station a été soit pleine soit vide). La courbe en pointillés affiche la proportion du temps pendant lequel la station est restée sous contrainte de capacité, lorsque la station était pleine ou vide (par exemple, une valeur de 0.5 indique que lorsque la station a été pleine ou vide, elle l'a été en moyenne pendant la moitié du temps, c'est-à-dire pendant 30 minutes). La



(a) Station « Place de la Croix-Rousse »

(b) Station « Opéra »

FIGURE 1.14 – Proportion de jours pendant lesquels la station s'est retrouvée soit dans l'état vide, soit dans l'état plein (représenté par un trait plein) et proportion du temps pendant lequel la station est restée sous contrainte de capacité, lorsque la station était pleine ou vide. (représenté par un trait en pointillés), pour deux stations du système Vélo'v. Une distinction est réalisée suivant que la station ait été soit vide (en rouge avec des triangles) soit pleine (en bleu avec des ronds).

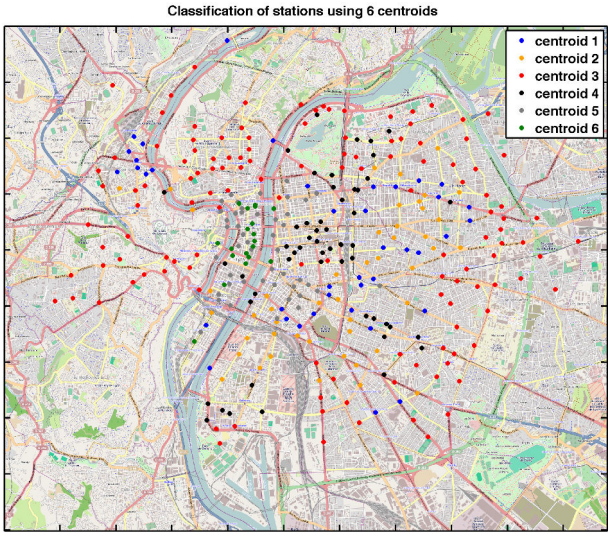
couleur permet de faire une distinction entre les moments pendant lesquels la station est pleine (en bleu avec des ronds) et les moments pendant lesquels la station est vide (en rouge avec des triangles).

La Figure 1.14b affiche le profil de la station « Place de la Croix-Rousse », située sur la colline de Croix-Rousse. Le pourcentage d'heure pendant lequel la station a été vide est très élevé quelle que soit l'heure de la journée (trait plein rouge avec des triangles autour de 60 %) : cela s'explique par le flux entrant très faible, à cause de l'altitude. Si l'on regarde le pourcentage moyen du temps où la station est restée vide (trait en pointillés rouge avec des triangles), on remarque qu'il est très élevé, ce qui signifie que la station est très souvent vide, avec une légère inflexion pendant la journée : elle est due à la présence de la régulation opérée par l'exploitant du système, et qui permet de remonter régulièrement des vers les stations en haut de la colline.

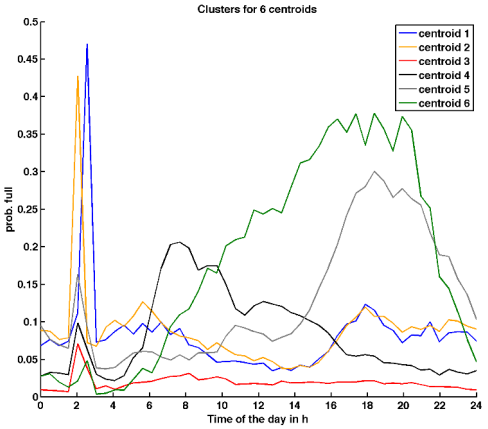
La Figure 1.14a affiche le profil de la station « Opéra », située sur la Presqu'île. On remarque un changement de tendance pendant la journée : le matin la station est globalement plutôt vide (trait rouge au-dessus du trait bleu) puis à partir de 15 heures la tendance s'inverse et la station a tendance à être plus souvent pleine que vide (trait bleu avec des ronds au-dessus du trait rouge avec des triangles). Cette inversion s'explique par le grand nombre d'activités qui composent le lieu, à la fois commerciales (boutiques du centre-ville), culturelles (opéras, théâtres, cinémas, etc.) et festives (restaurants, bars, etc.).

À partir de ces profils, obtenus pour chaque station, il est possible d'utiliser un algorithme classique de classification, par exemple l'algorithme des K-moyennes [158]. Cette étape n'a pas été réalisée pour ces données, car il n'a pas été jusque-là possible de valider proprement la méthode de détection des moments de contrainte de capacité. Cette validation nécessite en effet de disposer des jeux de données sur les mouvements et sur les disponibilités des stations, pour des intervalles de temps identiques, afin de comparer les périodes obtenues avec l'état des stations pendant ces périodes. La récente acquisition des données de déplacements pour l'année 2013 permet d'envisager comme perspective future ce travail de validation. L'obtention de profils dynamiques fiables pourraient également permettre une classification des stations.

Afin d'illustrer néanmoins un résultat de classification de stations, un exemple est proposé par la suite, portant sur des travaux réalisés en collaboration avec Maximilien Thess, en stage au laboratoire



(a) Carte des stations Vélo'v, la couleur indique la classe.



(b) Profil moyen de chaque classe

FIGURE 1.15 – Résultats d'une classification en utilisant comme profil des stations la probabilité pour une station d'être pleine au cours du temps.

de Physique en 2014. Les profils des stations sont obtenus en calculant une probabilité pour une station d'être pleine pour chaque intervalle de temps de 1h sur la journée, en se basant sur les données de disponibilités des stations. Les détails sur les calculs de ces profils ne sont pas évoqués dans ce manuscrit.

La Figure 1.15 affiche les résultats de cette classification, obtenue en utilisant la méthode des K-moyennes, avec $K = 6$. La classification permet de regrouper des stations avec des comportements proches. Par exemple, la classe verte regroupe des stations dont les probabilités d'être pleine en fin de journée sont élevées. Ces stations sont situées sur la Presqu'île, qui est un lieu avec une activité nocturne importante, ce qui explique que le risque de trouver une station pleine en fin de journée soit plus élevé qu'à un autre moment.

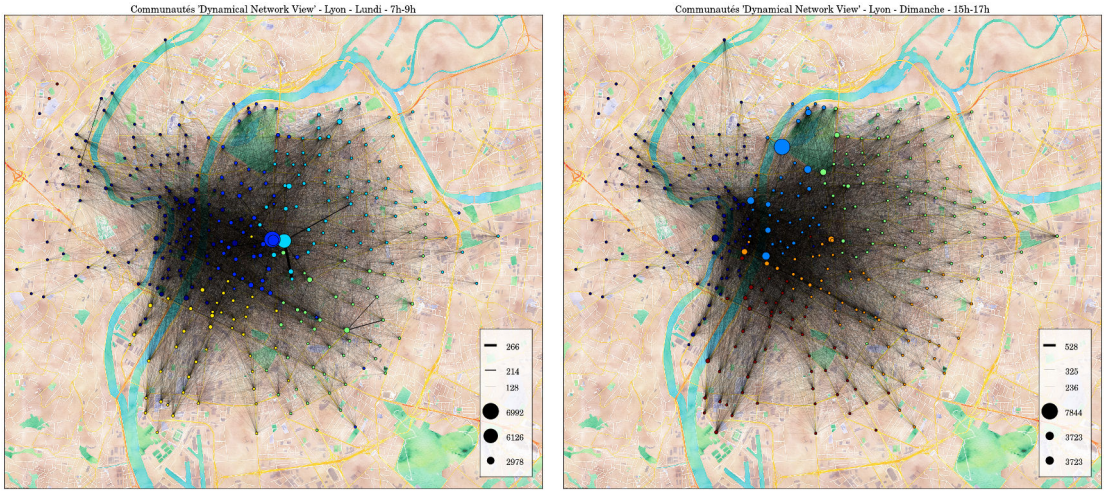
Cette illustration donne une idée des résultats que l'on peut obtenir avec ce type de classification, et les interprétations qu'elles permettent de faire. De manière générale, les classifications obtenues restent en cohérence avec les résultats connus sur les stations.

4.2 Détection de communautés dans un réseau

Une approche introduite d'abord par Cerf et al. [51] puis reprise par Borgnat et al. [36, 37] a été de considérer le système de Vélo'v comme un réseau, reprenant ainsi la notion de réseau de transport développé dans l'introduction de ce manuscrit. Parmi les résultats obtenus, il a été possible de comparer la structure du réseau, basée sur les flux de vélos entre les stations, au tissu géographique lyonnais, afin d'étudier les différences, en regardant des stations éloignées mais avec un comportement proche, et les similitudes, c'est-à-dire les stations formant un groupe géographique cohérent en termes d'activité.

L'idée de ces travaux introduits dans [37] consiste donc à représenter le système Vélo'v comme un réseau, c'est-à-dire comme un ensemble de relations entre les stations. Pour chaque paire de stations, le lien les unissant correspond au flux de vélos entre les deux stations. Une fois le réseau construit, une méthode de détection de communautés est appliquée (dans ces travaux la méthode de Louvain [29]) afin de repérer dans ce réseau des groupes de stations s'échangeant des vélos de manière privilégiée.

La Figure 1.16 affiche un exemple de classification de stations à partir d'une détection de communautés sur le réseau obtenu à partir des données Vélo'v. La couleur des nœuds indique la communauté, alors que la taille indique le nombre de vélos arrivant ou partant de ce nœud. Les communautés sont ancrées dans des zones géographiques, témoignant de la localité des déplacements.



(a) Lundi de 7h à 9h (b) Dimanche de 15h à 17h

FIGURE 1.16 – Exemple de classification de stations à partir d'une détection de communautés sur le réseau obtenu à partir des données Vélo'v. La couleur des nœuds indique la communauté, alors que la taille indique le nombre de flux arrivant ou partant de ce nœud.

Cette approche, qui est traitée ici de manière évasive, permet néanmoins de souligner la pertinence du réseau comme représentation des données Vélo'v.

5 Conclusions et perspectives

Ce premier chapitre a permis d'entrer dans le cœur de l'analyse de données réelles à travers l'application sur le système de vélos en libre-service à Lyon. Les différentes approches exposées révèlent d'une part la diversité des méthodes et leur technicité : chaque méthode requiert en effet un apprentissage théorique important afin d'en maîtriser les subtilités, que ce soit des outils de classification comme exposés dans les Sections 3 et 4, de statistique comme les travaux sur les modèles de régression linéaire présentés en Annexe A ou de manière générale d'analyse de données temporelles et spatiales. D'autre part, ce chapitre a permis de se rendre compte que la technique mathématique et informatique n'est pas suffisante pour obtenir de la connaissance. Elle permet tout au plus d'acquérir de l'information qui prend du sens uniquement lorsqu'elle est mise en regard avec une approche adaptée au système étudié, en l'occurrence dans ce chapitre avec les domaines du transport, de la géographie et de la sociologie.

Un dernier point qui retient notre attention consiste à analyser le système Vélo'v sous l'angle des réseaux. Dans la lignée des travaux présentés dans la Section 4.2, une représentation naturelle des données consiste à considérer chaque station comme interagissant avec les autres stations du système par le biais des déplacements de vélos. Ces travaux se sont limités à l'étude de réseaux statiques, à travers l'agrégation de l'activité sur des périodes d'intérêts. Si l'on veut pleinement exploiter la dimension spatio-temporelle de ces données, il est nécessaire d'ajouter une dimension temporelle à ce réseau, et ainsi de considérer des réseaux temporels, c'est-à-dire des réseaux dont la structure évolue au cours du temps. L'absence d'outils adaptés pour étudier ces objets motive le développement d'une méthode originale, qui est l'objet des chapitres 2, 3 et 4.

CHAPITRE 2

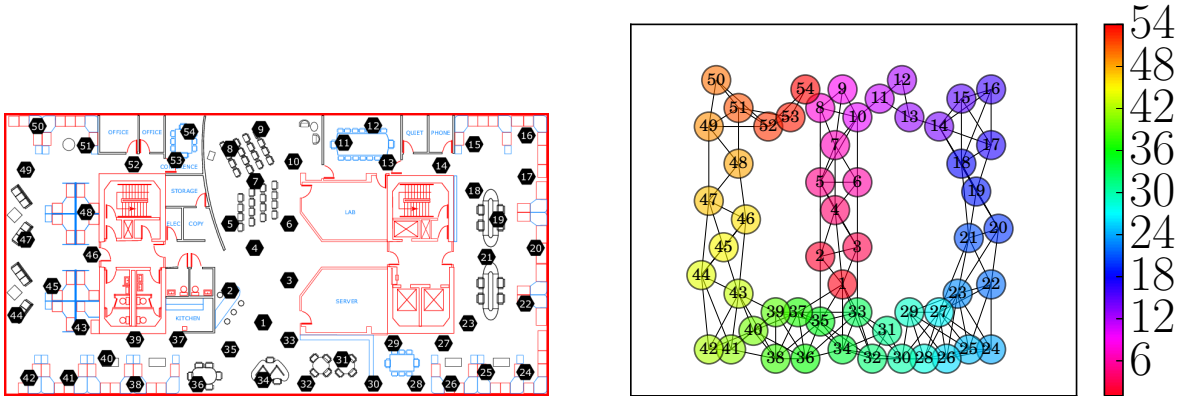
Étiquetage des nœuds du graphe en cohérence avec la structure

Résumé – Dans ce chapitre, une méthode pour numéroter les nœuds en cohérence avec la structure du graphe est proposée. Cette méthode se révèle en effet indispensable dans les chapitres suivants, dans lesquels un graphe est représenté sous la forme de signaux. Afin de garantir des signaux lisses, il est primordial de disposer d'un étiquetage des nœuds en cohérence avec la structure du graphe. La Section 1 énonce le problème, à travers deux illustrations. Des rappels sur les notions de graphes et réseaux qui seront utilisées dans la suite de ce manuscrit sont ensuite donnés dans la Section 2. La Section 3 situe le problème qui nous intéresse dans le cadre général des problèmes d'étiquetage des graphes. La description de l'heuristique proposée est présentée dans Section 4, puis son implémentation algorithmique et sa complexité dans la Section 5. La validation de l'heuristique est réalisée dans les Sections 6 et 7 sur des graphes standards et des réseaux complexes. Le chapitre se conclue par quelques perspectives.

Sommaire

1 Énoncé du problème

Dans de nombreuses applications, la structure d'un réseau complexe donne des indications précieuses pour la compréhension des relations sous-jacentes entre les différents nœuds du réseau. Plus précisément, il est souvent utile de considérer les nœuds dans un ordre cohérent avec la structure du graphe. Les nombreux travaux sur la détection de communautés dans un réseau [97] en est ainsi un exemple marquant, et révèle l'intérêt d'avoir des informations sur la structure : la mise en évidence de groupes de nœuds fortement connectés entre eux permet d'expliquer de nombreux comportements, par exemple dans les réseaux sociaux [111]. La présence de communautés n'est néanmoins qu'un seul type d'organisation parmi la multitude de structures possibles dans les réseaux [178]. Dans de nombreux cas, la topologie du réseau est inconnue et ne peut pas être caractérisée explicitement. L'optimisation d'une fonction globale pour mieux révéler les communautés quand la topologie du réseau ne correspond pas à ce type de structures n'aide pas à la compréhension de la structure du graphe. Des approches basées sur des mesures de proximité ont été proposées [65] et ont la capacité d'identifier qu'un réseau n'est pas structuré en communautés. En proposant une technique pour ordonner les nœuds d'un graphe en cohérence avec la structure du graphe, nous rendons possible d'identifier une classe plus large de structures de graphe.



(a) Plan du laboratoire avec la position des 54 capteurs.

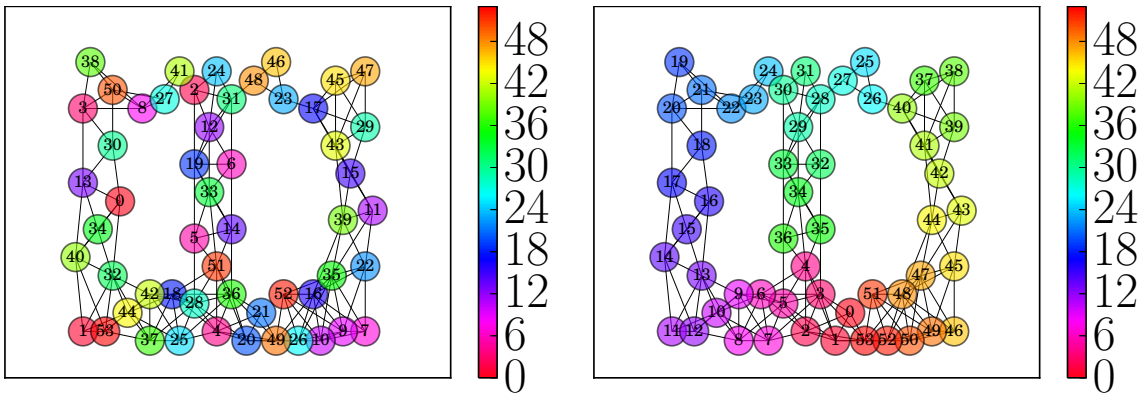
(b) Réseau de capteurs correspondant, prenant en compte la distance entre les capteurs et les contraintes topologiques du laboratoire. La couleur indique également l'identifiant du nœud.

FIGURE 2.1 – Exemple de réseau de capteurs obtenu à partir des données issues de l'expérience réalisée au Intel Berkeley Lab entre le 28 février et le 5 avril 2004 [30]. L'identifiant des nœuds est donné par un entier de 1 à 54, en reprenant l'identifiant initial des capteurs.

Ce problème d'étiquetage des nœuds d'un graphe a pour principale motivation la transformation de graphes en signaux : lorsque l'on passe d'un graphe, qui est un objet avec une dimension élevée, en des signaux unidimensionnels, une indexation adéquate des entités du réseau est nécessaire. Cette motivation est plus largement discutée dans le Chapitre 3. Néanmoins, l'intérêt d'avoir un étiquetage en cohérence avec la structure du graphe dépasse ce seul cadre d'application : on peut citer par exemple les problèmes d'inférence distribuée [137] ou de diffusion [52], afin de sélectionner l'ordre des nœuds lors de mises à jour asynchrones, ou de visualisation de réseaux [28], afin de représenter les nœuds dans un espace de

faible dimension. Une illustration est proposée, issue du domaine du traitement du signal sur graphe, qui a connu un développement très important ces dernières années [222, 210].

Considérant un ensemble de nœuds avec une structure définie, un signal est assigné à chaque nœud. Cette situation représente par exemple un réseau de capteurs, dans lequel chaque nœud du graphe est un capteur, qui peut communiquer avec les autres capteurs du système selon un réseau de communication établi. Un exemple de réseau de capteurs est construit à partir de l'expérience réalisée au Intel Berkeley Lab entre le 28 février et le 5 avril 2004 [30]. 54 capteurs, répartis dans plusieurs pièces du laboratoire (voir la Figure 2.1a), ont enregistré la température, l'humidité et la luminosité à intervalles réguliers pendant toute la durée de l'expérience. À partir des positions des capteurs dans l'espace, un réseau de capteurs est construit en prenant en compte la distance physique entre les capteurs, ainsi que d'obstacles entre les capteurs, comme les murs. La Figure 2.1b représente le réseau obtenu, en indiquant l'identifiant de chaque nœud à la fois par un entier entre 1 et 54 et à l'aide d'un code couleur, en utilisant les identifiants initiaux de chaque capteur : deux couleurs proches signifient que les deux identifiants sont également proches, en considérant un étiquetage cyclique, c'est-à-dire que 1 et 54 sont aussi proches que 1 et 2. L'utilisation de la couleur permet de mettre en évidence la forte relation entre structure et identification des nœuds : de manière naturelle, les capteurs ont été numérotés en suivant la configuration spatiale du laboratoire, probablement dans l'ordre de posage des capteurs. Avoir une séquence ordonnée des nœuds en fonction de la structure du graphe donne ainsi l'opportunité d'inférer la proximité géographique d'un nœud à partir d'une comparaison de leur étiquette.



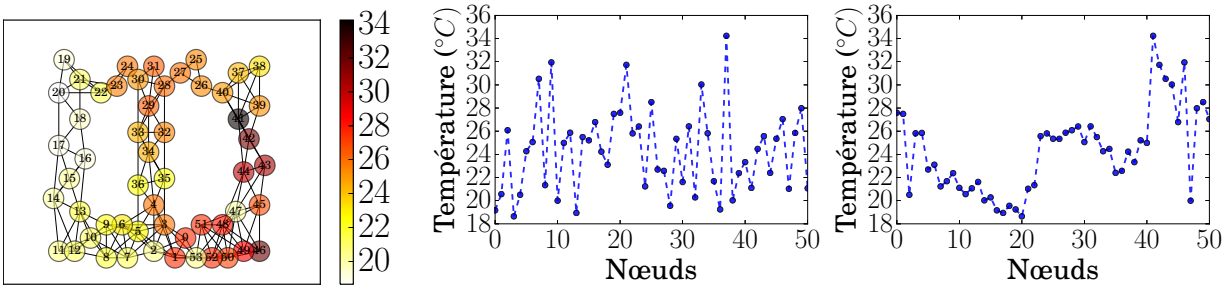
(a) Étiquetage aléatoire

(b) Étiquetage en cohérence avec la structure du graphe

FIGURE 2.2 – Représentation du réseau de capteurs Intel Berkeley Lab, en utilisant le même code couleur que la Figure 2.1b. Deux exemples d'étiquetage des nœuds sont proposés, l'un aléatoire, l'autre obtenu en utilisant la méthode présentée dans la Section 4.

Une première question qui peut se poser consiste à se demander dans quelle mesure il est possible de retrouver un ordre des nœuds cohérent avec la structure, dans le cas où les capteurs n'ont pas cette numérotation naturelle, comme dans le cas de la Figure 2.2a, où les nœuds sont ordonnés de façon aléatoire. L'ordre présenté à la Figure 2.2 est obtenu en utilisant l'heuristique discutée à la Section 4, et est différent de l'ordre initial, ici construit à la main par ceux qui déploient les capteurs, de la Figure 2.1b, bien que reflétant de la même façon la structure du réseau de capteurs.

L'ajout d'un signal sur chacun des nœuds permet d'introduire une autre question à laquelle un ordre cohérent avec la structure du graphe permet de répondre. À partir du réseau de capteurs présenté à la Figure 2.1b, on associe à chaque nœud un signal, ici choisi comme étant la température relevée par le capteur dans un intervalle de temps fixé. La Figure 2.3a affiche le réseau dans lequel la couleur du nœud indique la valeur de la température, en utilisant l'étiquetage présenté à la Figure 2.2b. On peut se rendre compte visuellement que la température n'est pas homogène le long du réseau : certains capteurs relèvent



(a) (b) Ordre aléatoire (c) Ordre cohérent avec la structure du graphe

FIGURE 2.3 – Introduction d'un signal sur les nœuds du réseau de capteurs Intel Berkeley Lab. (a) Représentation du réseau de capteurs. L'identifiant de chaque nœud est le même qu'à la Figure 2.2b, alors que la couleur du nœud indique la valeur de la température relevée (b-c) Représentation du signal en fonction du nœud du graphe pour différents étiquetages du graphe.

des températures différentes de celles des capteurs proches spatialement. Une visualisation de ce signal en fonction des nœuds dans un espace euclidien permet de se rendre compte des discontinuités spatiales de la température, à condition que l'ordre dans lequel les nœuds sont considérés pour l'indexation soit cohérent avec la structure du graphe. Ainsi, lorsque l'ordre des nœuds est choisi suivant l'étiquetage défini à la Figure 2.2b, les différentes évolutions spatiales de la température, visibles sur la Figure 2.3c, sont facilement identifiables : température quasi-constante entre les nœuds 44 et 9 (l'étiquetage est choisi ici cyclique, c'est-à-dire que les deux extrémités du signal sont raccordées), et une décroissance de la température pour les nœuds de 10 à 40, ponctuée d'irrégularités pour les nœuds 15, 16 et 22. Au contraire, utiliser un ordre aléatoire des nœuds ne permet pas de mettre en évidence ces évolutions spatiales de la température, qui ne sont visibles que lorsque le réseau peut être simplement visualisé, comme dans cet exemple où le réseau est petit et planaire, ce qui ne constitue pas la majorité des réseaux construits à partir de données réelles.

Une méthode pour ordonner les nœuds d'un graphe en cohérence avec la structure du graphe est proposée par la suite. Auparavant, quelques notions sur les graphes et les réseaux, utiles pour la suite du chapitre, sont rappelées.

2 Rappels sur les graphes et les réseaux

Cette section donne des éléments de théories des graphes et des réseaux nécessaires à la bonne compréhension des chapitres 2, 3 et 4. Elle est en grande partie basée sur les ouvrages de Clark et Holton [59] et de Mark Newman [178], auxquels le lecteur pourra se référer pour avoir plus de détails. Elle introduit également les notations utilisées tout au long de la thèse.

La théorie des graphes est le sous-domaine des mathématiques discrètes qui étudient les graphes, des objets mathématiques utilisés pour représenter des relations entre des entités. Elle propose des définitions et des résultats mathématiques pour l'étude de ces objets à travers la résolution de problèmes combinatoires. À titre d'exemple, on peut citer parmi les résultats importants de la théorie des graphes le théorème des quatre couleurs, qui indique qu'il est possible de colorer n'importe quelle carte découpée en régions connexes avec seulement 4 couleurs, de sorte que deux régions limitrophes n'aient pas la même couleur. La théorie des graphes fournit ainsi le cadre mathématique pour la preuve de ce théorème à l'aide d'outils souvent sophistiqués [112], mais également des algorithmes permettant d'obtenir cette coloration, pour un graphe donné.

La théorie des réseaux se différencie de la théorie des graphes en ceci qu'elle se concentre plus particulièrement sur les systèmes physiques pouvant se représenter sous la forme de graphes. Elle fait ainsi le lien entre la théorie des graphes et d'autres disciplines, comme la physique, l'informatique, la biologie, la sociologie ou l'économie, à la fois par l'étude de données issues de ces domaines (réseaux sociaux, réseau Internet, etc.) mais également par l'utilisation d'outils extérieurs à la théorie des graphes, comme ceux de la mécanique statistique.

2.1 Représentations d'un graphe

Graphe simple Un graphe $G = \{V, E\}$ est un objet mathématique qui consiste en un ensemble fini de nœuds V (aussi appelés sommets) et un ensemble de liens E (aussi appelés arêtes), où les éléments de E sont des paires non-ordonnées d'éléments de V. Le nombre de nœuds est noté $n = |V|$, alors que le nombre de liens est noté $m = |E|$. Les éléments de V sont implicitement associés à des entiers, allant de 1 à n : le nœud $u \in V$ sera implicitement associé à l'entier u, avec $u \in \{1, \ldots, N\}$. Un graphe est dit simple s'il ne contient pas de boucles, c'est-à-dire des liens dont les deux extrémités sont le même nœud, ni de liens multiples, c'est-à-dire des liens reliant la même paire de nœuds. Par la suite, nous considérerons uniquement des graphes simples.

Une représentation classique d'un graphe consiste à dessiner les nœuds comme des points, et les liens par des traits entre les points. La Figure 2.4 affiche un exemple de graphe, défini par l'ensemble de nœuds $V = \{A, B, C, D, E, F\}$ et l'ensemble des liens
$E = \{(A, B), (A, D), (A, F), (B, D), (B, F), (C, D), (C, E), (D, E)\}$.

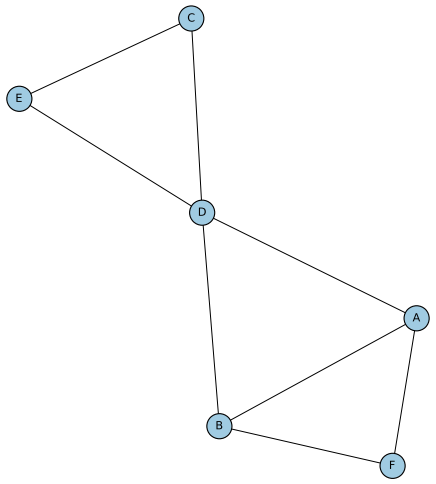


FIGURE 2.4 – Représentation d'un graphe non-pondéré et non-dirigé, défini par l'ensemble de nœuds $V = \{A, B, C, D, E, F\}$ représentés par les ronds bleus, et l'ensemble des liens $E = \{(A, B), (A, D), (A, F), (B, D), (B, F), (C, D), (C, E), (D, E)\}$, représentés par les traits entre les nœuds.

Graphe pondéré et/ou dirigé Deux variantes sont fréquentes dans l'analyse de graphe :

— Les liens du graphe peuvent être pondérés, en leur assignant une valeur représentant un degré de proximité ou au contraire, une dissimilarité : le graphe est dit pondéré ;

— Un lien peut exister dans une direction mais pas dans l'autre : le graphe dit dirigé.

Par la suite, les graphes seront considérés non-pondérés et non-dirigés, sauf mention contraire explicite.

Matrice d'adjacence La matrice d'adjacence d'un graphe, notée $\boldsymbol{A}$, permet de représenter sous la forme d'une matrice les relations entre les différents éléments de V. $\boldsymbol{A}$ est une matrice de dimensions

$n \times n$ définie par $\boldsymbol{A} = (a_{uv})_{u,v=1,\ldots,n}$ avec :

$$a_{uv} = \begin{cases} 1 & \text{si les nœuds } u \text{ et } v \text{ sont connectés} \\ 0 & \text{sinon} \end{cases} \tag{2.1}$$

La matrice d'adjacence du graphe présenté à la Figure 2.4 est ainsi donnée par :

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} \tag{2.2}$$

Lorsque G est non-dirigé, on a $a_{ij} = a_{ji}$ et donc la matrice $\boldsymbol{A}$ est symétrique. Si G est un graphe pondéré, a_{uv} est égal au poids associé au lien entre u et v.

Voisinage d'un nœud Le voisinage d'un nœud u, noté $\texttt{Adj}(u)$, est l'ensemble des nœuds adjacents au nœud u, c'est-à-dire les nœuds connectés au nœud u.

Dans l'exemple de la Figure 2.4, le voisinage du nœud D est donné par $\texttt{Adj}(D) = \{A, B, C, E\}$.

Chemin Un chemin est une séquence ordonnée de nœuds distincts, tels que deux nœuds consécutifs sont adjacents. La longueur d'un chemin est définie comme le nombre de liens dans le chemin.

Dans l'exemple dans la Figure 2.4, la séquence ordonnée $[A, D, E, C]$ forme un chemin, dont la longueur est 3.

Clique Une clique est un sous-ensemble de nœud formant un graphe complet. Le nombre de nœud donne la taille de la clique.

Dans l'exemple dans la Figure 2.4, les nœuds $\{C, E, D\}$ forment une clique de taille 3.

2.2 Mesures sur les graphes

Il existe une très grande quantité de mesures pour caractériser la structure des graphes [63]. Seules trois mesures sont utiles dans ce chapitre.

Degré Le degré d'un nœud $u \in V$, noté d_u est défini comme le nombre de nœuds connectés à u, c'est-à-dire :

$$d_u = \sum_{v=1}^{n} a_{uv} = |\texttt{Adj}(u)| \tag{2.3}$$

Dans l'exemple de la Figure 2.4, le degré du nœud D est 4.

Coefficient de clustering Le coefficient de clustering mesure à quel point les nœuds du graphe se regroupent ensemble. Il consiste à calculer le rapport entre le nombre de triangles (triplets de nœuds connectés entre eux) et le nombre de triplets de nœuds pouvant former un triangle :

$$C = \frac{3 \times \text{Nombre de triangles}}{\text{Nombre de triplets connectés}} \tag{2.4}$$

Si tous les triplets de nœuds connectés forment un triangle (par exemple dans un graphe complet), le coefficient de clustering est égal 1. Si aucun triplet connecté ne forme de triangle (par exemple dans un graphe cycle), alors le coefficient de clustering est égal à 0. Plus C est proche de 1, plus il y a de triangles et plus le voisinage des nœuds sont connectés. Le facteur 3 dans l'Équation 2.4 vient du fait qu'il est possible de considérer chaque triangle de trois manières différentes.

Dans l'exemple de la Figure 2.4, A, B et F sont des triplets connectés formant un triangle, alors que A, D, et C sont des triplets connectés ne formant pas un triangle. le coefficient de clustering est égal à 0.77.

Longueur moyenne du plus court chemin Un plus court chemin entre deux nœuds u et v est un chemin de longueur minimale entre u et v.

Dans l'exemple de la Figure 2.4, un des plus courts chemins entre C et F est $[F, B, D, E]$, et est de longueur 3.

La longueur moyenne du plus court chemin d'un graphe est la moyenne de la longueur du plus court chemin entre toutes les paires de nœuds du graphe. Dans l'exemple de la Figure 2.4, la longueur moyenne du plus court chemin est égale à 1.6.

2.3 Familles de graphes

Une présentation de quelques familles de graphes est réalisée, en se restreignant à celles utiles pour la suite pour valider notre approche, car il existe de nombreux résultats théoriques du fait de leur régularité. La Figure 2.5 propose une illustration pour quelques exemples de graphes.

Graphe chemin Un graphe chemin avec n nœuds, noté P_n, est une séquence ordonnée de n nœuds telle que chaque nœud, à l'exception du premier et du dernier, est connecté au nœud précédent et au nœud suivant.

Graphe cycle Un graphe cycle avec n nœuds, noté C_n, est un graphe chemin dont le premier et le dernier nœuds sont connectés, formant une séquence circulaire de nœuds.

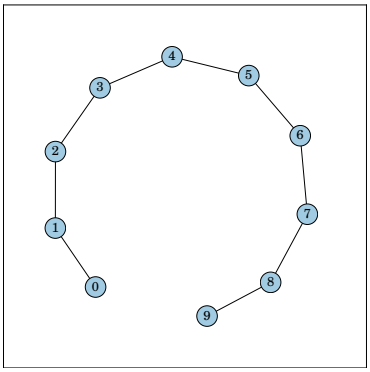
Graphe roue Un graphe roue avec n nœuds, noté W_n, est un graphe cycle dont les nœuds sont également connectés avec un unique nœud appelé *hub*.

Graphe complet Un graphe complet avec n nœuds, noté K_n, est un graphe pour lequel tous les nœuds sont connectés entre eux.

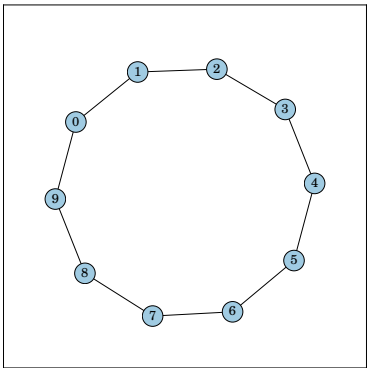
Graphe biparti complet Un graphe biparti complet est composé de deux ensembles avec respectivement n_1 et n_2 nœuds : chaque nœud du premier ensemble est connecté avec tous les nœuds du second ensemble, et il n'y a aucune connexion entre les nœuds d'un même ensemble. On le note $K_{n_1;n_2}$.

Produits cartésiens de graphes Le produit cartésien de deux graphes $G = (V_G, E_G)$, avec $\#V_G = n_G$, et $H = (V_H, E_H)$, avec $\#V_H = n_H$, noté $G \times H$, est le graphe dont l'ensemble des nœuds est $V_G \times V_H = \{(u, v) | u \in V_G, v \in V_H\}$, et dont les nœuds (u_G, u_H) et (v_G, v_H) sont connectés si et seulement si $u_G = v_G$ et $(u_H, v_H) \in E_H$ ou $u_H = v_H$ et $(u_G, v_G) \in E_G$.

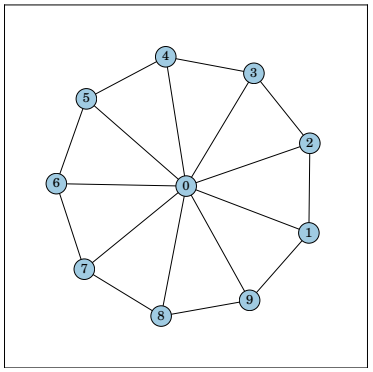
Graphe régulier en anneau Un graphe k-régulier en anneau avec n nœuds et $k \in \{2, \ldots, \lfloor \frac{n}{2} \rfloor\}$, est un graphe dans lequel chaque nœud u est connecté aux nœuds $\{u - \frac{k}{2}, u - \frac{k}{2} + 1, \ldots, u - 1, u + 1, \ldots, u + \frac{k}{2} - 1, u + \frac{k}{2}\}$. Pour $k = 2$, on retrouve le graphe cycle.



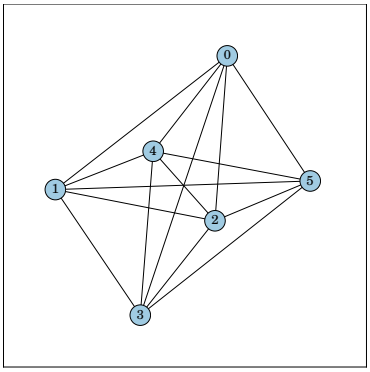
(a) Graphe chemin ($n = 10$)



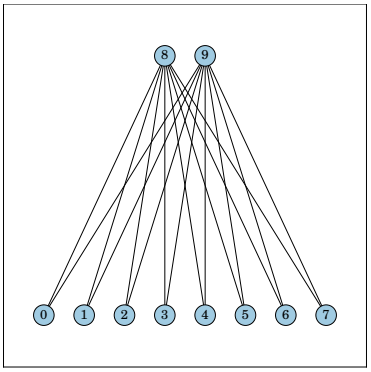
(b) Graphe cycle ($n = 10$)



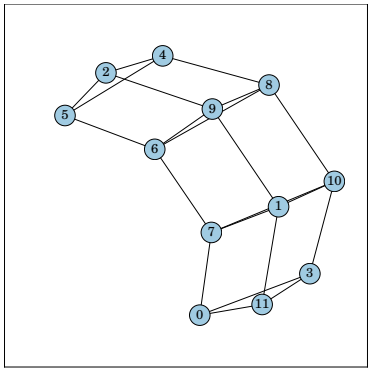
(c) Graphe roue $n = 10$



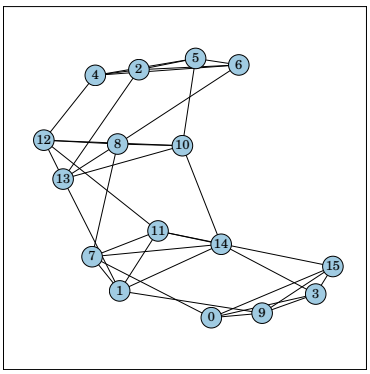
(d) Graphe complet $n = 6$



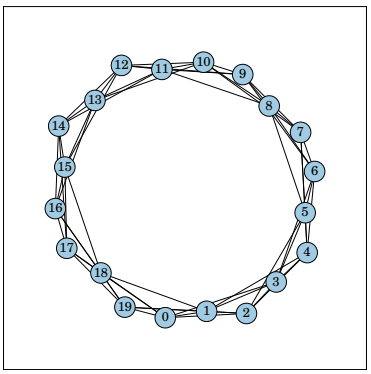
(e) Graphe biparti complet ($n_1 = 8$, $n_2 = 2$)



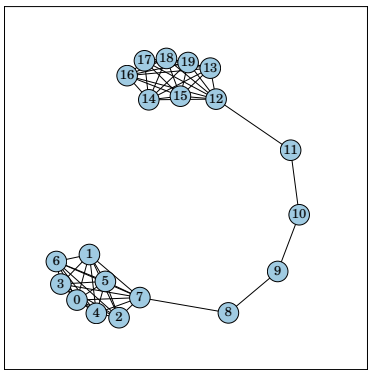
(f) Produit cartésien entre un graphe chemin et un graphe cycle ($n_G = 4$, $n_H = 3$



(g) Produit cartésien entre un graphe chemin et un graphe complet ($n_G = 4$, $n_H = 4$



(h) Graphe 6-régulier en anneau ($n = 20$)



(i) Graphe « Barbell » ($n_c = 8$, $n_p = 10$)

FIGURE 2.5 – Représentation de quelques graphes standards définis dans la Section 2.3.

Graphe « Barbell » Un graphe « Barbell » [11] est défini comme deux cliques de taille n_c, liées par un chemin avec n_p nœuds.

2.4 Réseaux complexes

Les réseaux complexes sont des systèmes physiques qui peuvent se représenter sous forme de graphe. L'étude de réseaux complexes issus de données réelles a permis de mettre en évidence plusieurs propriétés liées à leur structure. Des modèles ont alors été développés pour reproduire ces propriétés, afin d'étudier théoriquement les systèmes en question. Ces propriétés sont décrites, ainsi que les modèles permettant de les simuler.

2.4.1 Propriétés

Propriété petit monde La propriété petit-monde (ou *small-world property*) [247] est une propriété des réseaux dont la longueur moyenne du chemin le plus court est petite comparée au nombre de nœuds du réseau. On retrouve souvent cette propriété dans les réseaux sociaux.

Structure en communautés Une communauté est une notion assez floue, qui peut se définir comme un regroupement de nœuds tel que le nombre de liens entre les nœuds d'une communauté est significativement plus élevé qu'entre les nœuds appartenant à des communautés différentes. Il existe de nombreux travaux sur la détection de communautés dans un graphe, chacune adaptant précisant cette définition [97].

Cette propriété est par exemple bien connue dans les réseaux sociaux, dans lesquels les personnes ont tendance à appartenir à des groupes sociaux distincts [111].

2.4.2 Modèles de réseau complexe

Modèle d'Erdős-Rényi Le modèle d'Erdős-Rényi est un modèle de graphes aléatoires, permettant d'obtenir des graphes sans structure particulière. Ce modèle est paramétré par le nombre de nœuds n dans le graphe, et la probabilité p qu'un lien existe entre chaque paire de nœuds.

Algorithme 1 : Modèle d'Erdős-Rényi

Entrées : $n \in \mathbb{N}^*$, $p \in [0, 1]$
Sorties : G un graphe simple, non-dirigé et non-pondéré

1 **début**
2 Définir un graphe $G = (V, E)$ avec $|V| = n$ et $E = \emptyset$
3 **pour** $\forall (u, v) \in V^2$ *tel que* $u \neq v$ **faire** Ajouter (u, v) à E avec une probabilité p

Modèle de Watts-Strogatz Le modèle de Watts-Strogatz est un modèle utilisé pour générer des graphes avec une propriété petit monde. Le modèle est paramétré par le nombre de nœuds n, la régularité k du

graphe initial, et la probabilité p de reconnexion d'un lien.

Algorithme 2 : Modèle de Watts-Strogatz

Entrées : $n \in \mathbb{N}^*$, $k \in \{2, 4, \ldots, ...\}$, $p \in [0, 1]$
Sorties : G un graphe simple, non-dirigé et non-pondéré

début
  Définir un graphe $G = (V, E)$ formant un graphe k-régulier en forme d'anneau
  pour $\forall (u, v) \in V^2$ *tel que* $u \neq v$ **faire**
    si $(u, v) \in E$ **alors**
      Changer v par k avec une probabilité p, avec $k \in \{w | w \neq u, w \neq v, (u, k) \in E\}$ choisi de manière uniforme

Modèle à blocs stochastiques Le modèle à blocs stochastiques (*Stochastic block model* (SBM)) [138] est un modèle génératif de graphe avec une structures en communautés. Il peut décrire une très large variété de structures de graphe, comme des graphes hierarchiques, multi-échelles, ou des composantes déconnectées.

Un modèle à blocs stochastiques est considéré par la suite afin de construire un graphe avec C communautés. Une communauté est aléatoirement attribuée à chaque nœud, et entre toutes les paires de nœuds, un lien existe avec une probabilité p_{intra} si les deux nœuds appartiennent à la même communauté, et p_{inter} si les deux nœuds appartiennent à des communautés différentes, avec $p_{\text{inter}} \ll p_{\text{intra}}$.

Algorithme 3 : Modèle à blocs stochastiques

Entrées : $n \in \mathbb{N}^*$, $C \in \{2, 3, \ldots, N\}$, $p_{\text{intra}} \in [0, 1]$, $p_{\text{inter}} \in [0, 1]$
Sorties : G un graphe simple, non-dirigé et non-pondéré

début
  Définir un graphe $G = (V, E)$ avec $|V| = n$ et $E = \emptyset$
  Assigner aléatoirement avec une probabilité uniforme une communauté à chaque nœud
  pour $\forall (u, v) \in V^2$ *tel que* $u \neq v$ **faire**
    si u *et* v *sont dans la même communauté* **alors** Ajouter (u, v) à E avec une probabilité p_{intra}
    sinon Ajouter (u, v) à E avec une probabilité p_{inter}

Modèle mixte de Watts-Strogatz à blocs stochastiques Ce modèle, introduit dans cette thèse, est un mélange entre le modèle de Watts-Strogatz et le modèle à blocs stochastiques. Il est défini comme une généralisation du graphe « Barbell », en ce sens qu'il permet de considérer des graphes avec plusieurs communautés, qui ne sont pas nécessairement des cliques, ainsi que des structures k-régulières entre ces

communautés, à la place d'un chemin.

Algorithme 4 : Modèle mixte de Watts-Strogatz à blocs stochastiques

Entrées : $n \in \mathbb{N}^*$, $k \in \{2, 4, \ldots, ...\}$, $C \in \{2, 3, \ldots, N\}$, $p_{\text{intra}} \in [0, 1]$, $p_{\text{inter}} \in [0, 1]$
Sorties : G un graphe simple, non-dirigé et non-pondéré

début
  Définir un graphe $G = (V, E)$ formant un graphe k-régulier en forme d'anneau
  Définir aléatoirement $C + (C - 1)$ groupes, en préservant l'ordre des nœuds
  Définir les groupes $\{2, 4, \ldots, C + (C - 2)\}$ comme étant les parties régulières, et les groupes $\{1, 3, \ldots, C + (C - 1)\}$ comme étant les communautés
  pour $\forall (u, v) \in V^2$ *tel que* $u \neq v$ **faire**
    si u *et* v *sont dans la même communauté* **alors** Ajouter (u, v) à E avec une probabilité p_{intra}
    sinon Ajouter (u, v) à E avec une probabilité p_{inter}

Des représentations d'instances générés à l'aide de chaque modèle sont données dans la Section 3 du Chapitre 3.

3 Cadre général des problèmes d'étiquetage de graphe

3.1 Présentation

Les problèmes d'étiquetage de graphe concernent l'attribution d'un indice, ou étiquette, aux nœuds ou aux liens du graphe. La manière dont ces indices sont attribués est guidée par la minimisation d'une fonction objectif, définie par le problème auquel on s'intéresse. Il existe ainsi une grande variété de problèmes d'étiquetage de nœuds dans un graphe, la plupart étant reliée à une application spécifique, par exemple en analyse numérique ou en visualisation.

Chung [57] a proposé un cadre qui permet d'unir dans un même formalisme de nombreux problèmes d'étiquetage de graphe. Les problèmes d'étiquetage de graphe consistent à trouver l'application π de V dans V_H qui minimise une fonction objectif, où V_H désigne l'ensemble des nœuds d'un graphe hôte $H = (V, E_H)$. Cette fonction objectif est très souvent définie par rapport à une distance d_H entre deux sommets, correspondant à la longueur du plus court chemin entre ces deux sommets dans le graphe hôte H. Nous nous restreignons par la suite à deux fonctions objectifs souvent considérées :

1. la distance maximale d_H entre les étiquettes de deux nœuds connectés est minimisée, c'est-à-dire on cherche l'étiquetage $\hat{\pi}$ tel que :

$$\hat{\pi} = \arg\min_{\pi} \max_{\{u,v\} \in E} d_H(\pi[u], \pi[v]) \tag{2.5}$$

2. la somme des distances d_H entre toutes les paires de nœuds connectés est minimisée, c'est-à-dire on cherche l'étiquetage $\hat{\pi}$ tel que :

$$\hat{\pi} = \arg\min_{\pi} \sum_{\{u,v\} \in E} d_H(\pi[u], \pi[v]) \tag{2.6}$$

Ces problèmes ont été intensivement étudiés dans le cas où le graphe hôte est un chemin P, c'est-à-dire $E_P = \{\{i, i+1\} \mid i = 0 \ldots n-2\}$. La longueur du plus court chemin entre deux nœuds u et v dans ce graphe est donnée par :

$$d_P(\pi[u], \pi[v]) = |\pi[u] - \pi[v]| \tag{2.7}$$

Ces problèmes sont appelés respectivement *Banwidth Problem* (Équation 2.5) et *Bandwidth Sum Problem* (Équation 2.6).

Lin [152] et Jianxiu [135] ont introduit les problèmes pour lesquels le graphe hôte est un cycle C, avec $E_C = \{\{i, i+1\} \mid i = 0 \dots n-2\} \bigcup \{n-1, 0\}$. Dans ce cas, la distance entre deux nœuds u et v dans ce graphe est donné par :

$$d_C(\pi[u], \pi[v]) = \min\{|\pi[u] - \pi[v]|, n - |\pi[u] - \pi[v]|\} \tag{2.8}$$

Ces problèmes sont appelés respectivement *Cyclic Banwidth Problem* (Équation 2.5) et *Cyclic Bandwidth Sum Problem* (Équation 2.6). Ce chapitre sur concentre sur le *Cyclic Bandwidth Sum Problem* (CBSP), basé sur la minimisation du *Cyclic Bandwidth Sum* (CBS) :

$$\min_{\pi} CBS(G) = \min_{\pi} \sum_{\{u,v\} \in E} d_C(\pi[u], \pi[v]) \tag{2.9}$$

L'hypothèse retenue est que ce dernier problème est adapté à la recherche d'un étiquetage cohérent avec la structure du graphe. La minimisation de la différences des étiquettes laisse en effet espérer que deux nœuds proches dans le graphe vont avoir des étiquettes proches. De plus, la présence de structures cycliques dans les réseaux complexes, est un élément qui motive le choix du *Cyclic Bandwidth Sum Problem* comme problème relié à notre application.

3.2 État de l'art

De nombreux travaux ont été réalisés sur l'étude des problèmes d'étiquetage de graphes. À partir des travaux de Chung [57], de nombreux travaux ont été proposés pour la résolution théorique ou algorithmique de ces problèmes : Díaz [75] a proposé un tour d'horizon de plusieurs problèmes avec un aspect algorithmique, en détaillant les applications sous-jacentes à chacun des problèmes. De nombreuses heuristiques ont été proposées pour trouver des solutions approximatives à ces problèmes, parmi elles on peut citer celles concernant le *Bandwidth Problem* [64], le *Bandwidth Sum Problem* [18, 203, 204] ou le *Cyclic Bandwidth Problem* [152, 206], mais également des problèmes reliés tels que le *Antibandwidth Problem* [45] ou le *Cyclic Antibandwidth Problem* [154], où la fonction objectif n'est plus une minimisation mais une maximisation. Ces problèmes sont généralement NP-complets, comme montré par Papadimitriou pour le *Bandwidth Problem* [185] et Lin pour le *Cyclic Bandwidth Problem* [152].

Peu de résultats sont en revanche disponibles dans la littérature sur la résolution du *Cyclic Bandwidth Sum Problem*. Deux articles se concentrent sur des aspects mathématiques du problème : Jianxiu [135] a introduit le problème et proposé des résultats théoriques pour des graphes standards, tel que le graphe roue ou les graphes k-réguliers, concernant soit une valeur optimale du CBS, soit une borne supérieure pour cette valeur, et ce en fonction du nombre de nœuds et de liens dans le graphe. En 2007, Chen et al. [56] ont étendu ce travail en ajoutant quelques résultats, par exemple sur les graphes complets bipartis. Si ces résultats théoriques sur la valeur optimale du CBS pour certaines classes de graphes ne donnent pas d'indication sur la manière d'obtenir l'ordre des nœuds permettant d'atteindre cette valeur optimale, ils sont néanmoins utiles pour évaluer les performances des heuristiques.

Seulement une heuristique a été proposée par Satsangi et al. [213, 212] pour résoudre ce problème, basée sur un parcours à voisinage variable *General Variable Neighborhood Search* et appelée **gvns**. L'idée de cette méta-heuristique consiste à modifier la solution à la fois de manières globale et locale : à partir d'une solution initiale, une phase de mélange est d'abord appliquée dans laquelle les nœuds sont aléatoirement décalés, inversés, échangés dans la séquence ordonnée, sans tenir compte de la proximité entre les nœuds. Cette opération permet de parcourir l'espace des solutions afin d'échapper aux vallées. Dans un deuxième temps, une recherche locale est réalisée afin de descendre une vallée vers un minimum local, en intervertissant consécutivement des nœuds consécutifs dans la séquence, ou en échangeant des

nœuds adjacents dont la contributions à la valeur du CBS est la plus forte.L'heuristique développée dans la section suivante est comparée à l'heuristique **gvns**, ainsi qu'à la méthode **hla** [18], développée pour résoudre le *Bandwidth Sum Problem*. Cette heuristique propose une approche « Diviser pour régner » : elle repose sur une décomposition en arbre binaire équilibré, qui décrit une partition récursive de l'ensemble des nœuds V. À partir de cet arbre, les auteurs proposent un algorithme optimal pour calculer les meilleures séquences des nœuds en décidant, pour chaque nœud de l'arbre, l'ordre de ses deux enfants.

4 Heuristique pour la minimisation du *Cyclic Bandwidth Sum* d'un graphe

L'objectif de l'heuristique est d'ordonner les nœuds du graphe en fonction de sa structure. Les identifiants des nœuds sont ainsi contraints par la régularité de la structure, qui peut se présenter sous différentes formes. Par exemple, dans le cas simple d'un graphe cycle, le comportement adéquat de l'algorithme doit être le suivant : à partir d'un nœud quelconque, la numérotation se fait à partir de ce nœud puis vers un des deux voisins, et continue en suivant le cycle jusqu'à revenir au premier nœud. Dans le cas moins trivial où le graphe est organisé en plusieurs cliques, l'algorithme doit parcourir tous les nœuds d'une clique, avant de sauter à la suivante et parcourir les nœuds suivants. De manière plus générale, l'algorithme doit adapter son parcours à la structure du graphe, pour toutes les structures possibles.

Une solution naïve pour atteindre cet objectif consiste à réaliser une marche aléatoire sur le graphe, qui numérote successivement les nœuds quand ils sont atteints. Cette approche a néanmoins un inconvénient : le choix du nœud suivant dans la marche aléatoire dépend seulement du voisinage du nœud courant, et non d'un voisinage plus étendu. Cela implique que si le nœud courant est relié à un nœud lointain, la structure ne sera pas correctement conservée. De plus la marche aléatoire doit être contrôlée afin d'éviter de retourner vers des nœuds déjà visités, mais également afin d'empêcher que la marche ne s'arrête avant d'avoir visité tous les nœuds. Enfin, le caractère aléatoire du parcours rend le contrôle de la solution finale compliqué.

L'heuristique proposée dans ce chapitre permet d'éviter cet écueil en décomposant la recherche d'un ordre des nœuds adapté en deux étapes. La première étape réalise des parcours locaux contrôlés par la structure, afin d'obtenir une collection de chemin indépendants en relation avec la structure locale du graphe. La deuxième étape consiste à fusionner tous les chemins obtenus par une approche gloutonne en minimisant globalement la valeur de *Cyclic Bandwidth Sum* à partir des chemins locaux.

4.1 Étape 1 : Parcours du graphe à travers des nœuds localement similaires

La première étape consiste à chercher une collection de chemins dans le graphe, c'est-à-dire des séquences de nœuds consécutivement connectés. L'algorithme réalise un parcours en profondeur dans laquelle le nœud suivant est choisi parmi les voisins du nœud courant, de manière à préserver la connectivité du chemin. Le choix du nœud suivant est basé sur un indice de similarité entre chaque voisin et le nœud courant, qui dépend de l'intersection entre les voisinages des deux nœuds : plus les voisinages sont proches, plus la probabilité pour le nœud d'être choisi est grande. Concrètement, le parcours est réalisé de la façon suivante : à partir d'un nœud, l'algorithme saute vers le nœud voisin le plus similaire qui n'est pas encore dans un chemin, et ainsi de suite jusqu'à ce qu'il n'y ait plus de nœud accessible. L'algorithme commence alors un nouveau chemin à partir d'un nœud libre, et continue de construire des chemins jusqu'à ce que tous les nœuds soient dans un chemin. À la fin de cette étape, on dispose d'une collection de chemins formant une partition de l'ensemble des nœuds du graphe, chaque nœud appartenant à un unique chemin.

Initialisation N'importe quel nœud qui n'a pas encore été inséré dans un chemin peut être utilisé comme nœud de départ pour le parcours. Néanmoins, afin de favoriser des chemins longs, les nœuds

à la périphérie du graphe sont de meilleurs candidats : intuitivement, un chemin idéal partirait d'une extrémité du graphe puis parcourrait tous les nœuds du graphe jusqu'à une autre extrémité. Dans le cas simple d'un graphe chemin par exemple, partir au milieu du chemin mènerait à deux chemins, alors qu'il paraît évident que partir à une des extrémités permettrait d'obtenir un unique chemin. Il existe plusieurs mesures afin de déterminer la centralité d'un nœud, qui peuvent être mise en oeuvre afin de trouver des nœuds périphériques. La mesure la plus simple, à savoir le degré du nœud, est choisie pour des raisons de calcul. Le chemin avec le plus faible degré est ainsi le point de départ du chemin.

Construction d'un chemin Un chemin est obtenu en réalisant un parcours en profondeur du graphe, dans lequel le nœud suivant est choisi parmi les voisins du nœud courant selon ces deux critères :

1. le nœud n'est pas encore dans un chemin ;
2. le voisinage du nœud est celui qui est le plus similaire avec le voisinage du nœud courant.

La similarité entre les voisinages de deux nœuds est calculée à l'aide de l'indice de Jaccard [130], qui permet de comparer la similarité entre deux ensembles en calculant le ratio entre le nombre d'éléments communs entre les deux ensembles et le nombre total d'éléments dans les deux ensembles. L'indice de similarité entre les nœuds u et v, noté $J(u, v)$, est défini de la façon suivante :

$$J(u,v) = \frac{|(\texttt{Adj}(u) \cup \{u\}) \cap (\texttt{Adj}(v) \cup \{v\})|}{|(\texttt{Adj}(u) \cup \{u\}) \cup (\texttt{Adj}(v) \cup \{v\})|} \tag{2.10}$$

Cette mesure est égale à 1 si les nœuds u et v sont connectés et ont les mêmes nœuds adjacents, sinon elle est strictement inférieure à 1. Une valeur proche de zéro signifie que le nombre total de nœuds dans les deux voisinages est beaucoup plus élevé que le nombre de nœuds en commun.

Lorsque l'algorithme rencontre un nœud adjacent de degré 1, c'est-à-dire adjacent seulement au nœud courant, il est préférable que ce nœud ne soit pas sélectionné comme le nœud suivant du chemin, malgré la similarité maximale avec le nœud courant, afin de ne pas finir le parcours. Néanmoins, il est également important d'insérer ce nœud dans le chemin courant, auquel il est forcément proche. Cette entorse aux conditions (1) et (2) est ainsi réalisée dans ce cas de figure uniquement, afin de prendre en compte ce nœud tout en laissant le parcours se poursuivre.

Fin du parcours Le parcours d'un chemin se termine lorsque tous les nœuds adjacents du nœud courant ont été insérés dans un chemin. L'algorithme débute alors un nouveau chemin en utilisant les nœuds restants, jusqu'à ce qu'il ne reste plus de nœuds disponibles.

4.2 Étape 2 : Fusion gloutonne des chemins

La seconde étape de l'algorithme cherche à agréger les chemins obtenus à l'étape 1 de manière à obtenir une séquence ordonnée de nœuds. La fusion est réalisée en suivant le processus suivant : à partir d'une liste vide, les chemins sont ajoutés un à un de manière séquentielle de telle sorte qu'à chaque itération, la valeur de CBS est globalement minimale. Plus précisément, pour une séquence ordonnée partielle et un chemin donnés, la position à laquelle le chemin est inséré dans la séquence ordonnée partielle, ainsi que le sens dans lequel ce chemin est inséré (c'est-à-dire si le chemin est inversé ou non), sont choisis tels que la valeur du CBS, calculée sur les nœuds composant la séquence ordonnée partielle et le chemin, est minimisée. Les chemins sont sélectionnés tour à tour en fonction de la taille, les plus grands chemins étant d'abord sélectionnés, afin de mieux explorer l'espace des solutions possibles.

Calcul incrémentiel du CBS Le calcul du CBS, donné par l'Équation 2.9, a un coût computationnel très élevé, car il nécessite de prendre en compte chaque lien du graphe. Pour chaque insertion d'un chemin, la valeur de CBS est calculée deux fois, quand le chemin est inséré à l'endroit et à l'envers, et

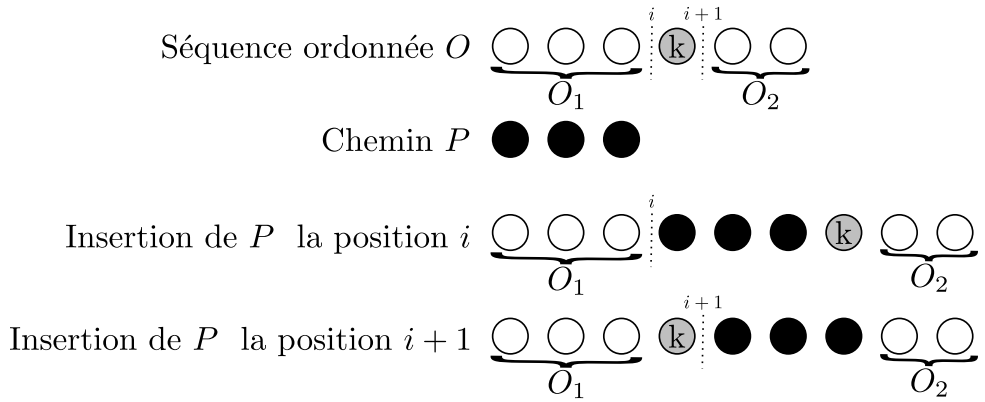


FIGURE 2.6 – Schéma de l'insertion d'un chemin dans une séquence ordonnée en fonction des positions d'insertion i et $i+1$.

ceci pour chaque position possible de la séquence ordonnée. Si la séquence ordonnée partielle contient n nœuds, il y a ainsi $n+1$ positions auxquelles le chemin peut être inséré, et le choix de la meilleure position requiert alors $2(n+1)$ calculs de CBS. Ce coût important peut être allégé en observant que d'une position à une autre, beaucoup de liens gardent la même contribution dans la valeur de CBS. Une approche est ainsi proposée, à travers des mises à jour incrémentielles de la valeur du CBS à chaque décalage d'indice : seuls les liens dont les nœuds ont vu leur étiquette modifiée sont pris en compte.

Les notations suivantes sont adoptées : à une itération fixée, la séquence ordonnée partielle est notée O et contient n_0 nœuds, et le chemin est noté P et contient n_p nœuds. On note i la position d'insertion du chemin dans la séquence ordonnée, avec $i \in \{0, \ldots, n_o\}$. La séquence ordonnée est décomposée en trois parties : la première partie est notée O_1 et contient les nœuds situés avant la position i. Le nœud juste après la position i est noté k, alors que la troisième partie constituée des nœuds après k est notée O_2. Avec ces notations, l'insertion du chemin P dans la séquence ordonnée O à la position i revient à obtenir la nouvelle séquence ordonnée composée de O_1 suivie du chemin P, du nœud k et terminée par O_2. De même, l'insertion du chemin P dans la séquence ordonnée à la position $i+1$ revient à obtenir la nouvelle séquence composée de O_1, suivie du nœud k, du chemin P et terminée par O_2. La Figure 2.6 résume dans un schéma le processus effectué.

À l'aide de cette représentation, il est clair que les étiquettes changent seulement pour les nœuds dans P et pour le nœud k, lorsque la position d'insertion passe de i à $i+1$. Soit $\pi_i[u]$ l'identifiant d'un nœud u quand P est inséré à la position i, c'est-à-dire après le nœud à la i^{e} position dans la séquence ordonnée. Les changements d'étiquette pour chaque partie de la nouvelle séquence ordonnée sont les suivants :

$$\pi_{i+1}[k] = \pi_i[k] - n_p \tag{2.11}$$
$$\forall u \in P,\ \pi_{i+1}[u] = \pi_i[u] + 1 \tag{2.12}$$
$$\forall u \in O_1,\ \pi_{i+1}[u] = \pi_i[u] \tag{2.13}$$
$$\forall u \in O_2,\ \pi_{i+1}[u] = \pi_i[u] \tag{2.14}$$

On note $CBS^{(i)}$ la valeur du CBS quand P est inséré à la position i. Le calcul de $CBS^{(i)}$ peut être décomposé en fonction des différents ensembles de nœuds définis plus haut :

$$\begin{aligned} CBS^{(i)} = {} & CBS^{(i)}(O_1, O_1) + CBS^{(i)}(O_2, O_2) + CBS^{(i)}(O_1, O_2) + CBS^{(i)}(P, P) \\ & + CBS^{(i)}(k, O_1) + CBS^{(i)}(k, O_2) + CBS^{(i)}(k, P) \\ & + CBS^{(i)}(P, O_1) + CBS^{(i)}(P, O_2) \end{aligned} \tag{2.15}$$

où $CBS^{(i)}(X, Y) = \sum_{u \in X, v \in Y, \{u,v\} \in E} d_C(\pi_i[u], \pi_i[v])$ est la valeur de CBS lorsque seulement les liens du graphe entre les ensembles X et Y sont considérés, avec $d_C(\pi[u], \pi[v])$ défini par l'Équation 2.8.

La définition de la distance d_C permet de mettre en évidence de façon triviale que les nœuds dans les ensembles aux extrémités de la séquence ordonnée ne sont pas affectés par le décalage, et qu'ainsi la valeur de d_C reste la même :

$$CBS^{(i+1)}(O_1, O_1) = CBS^{(i)}(O_1, O_1) \tag{2.16}$$
$$CBS^{(i+1)}(O_2, O_2) = CBS^{(i)}(O_2, O_2) \tag{2.17}$$
$$CBS^{(i+1)}(O_1, O_2) = CBS^{(i)}(O_1, O_2) \tag{2.18}$$
$$CBS^{(i+1)}(P, P) = CBS^{(i)}(P, P) \tag{2.19}$$

Quand le décalage a une influence sur les identifiants des nœuds, il est nécessaire de considérer non seulement les changements induits par le décalage, mais également quels termes entre $|\pi[u] - \pi[v]|$ et $n_o - |\pi[u] - \pi[v]|$ va être le minimum, l'un ou l'autre pouvant l'être aux indices i et $i+1$. Une preuve est discutée lorsque les liens impliquent k et un des nœuds dans O_1. Les autres résultats reprennent le même raisonnement.

Théorème 4.1. *(Liens entre k et les nœuds de O_1)*
Soit $u \in O_1$ et $\Delta = \pi_i[k] - \pi_i[u]$:

1. *si $\Delta \leq \frac{n}{2}$ alors $CBS^{(i+1)}(k, u) = CBS^{(i)}(k, u) - n_p$.*
2. *si $\Delta \geq \frac{n}{2} + n_p$ alors $CBS^{(i+1)}(k, u) = CBS^{(i)}(k, u) + n_p$.*
3. *si $\frac{n}{2} < \Delta < \frac{n}{2} + n_p$ alors $CBS^{(i+1)}(k, u) = CBS^{(i)}(k, u) + 2\Delta - (n_o + n_p)$*

Démonstration. Pour tout $u \in O_1$, on a $\pi_{i+1}[u] = \pi_i[u] < \pi_{i+1}[k] < \pi_i[k]$ à partir des équations 2.11 et 2.13. Ainsi, $0 < \pi_{i+1}[k] - \pi_{i+1}[u] < \Delta$, permettant la suppression de la valeur absolue dans l'Équation 2.8.

Considérons le cas où le terme minimal utilisé pour calculer $CBS_i(u, k)$ dans l'Équation (2.8) est le premier terme. Cela signifie que :

$$\Delta \leq n_o - \Delta \Leftrightarrow \Delta \leq \frac{n_o}{2} \tag{2.20}$$

Lorsque l'on considère $CBS_{i+1}(u, k)$, le premier terme de l'Équation 2.8 est retenu si :

$$\begin{aligned} &\pi_{i+1}[k] - \pi_{i+1}[u] \leq n_o - (\pi_{i+1}[k] - \pi_{i+1}[u]) \\ \Leftrightarrow &\Delta - n_p \leq n_o - (\Delta - n_p) \\ \Leftrightarrow &2(\Delta - n_p) \leq n_o \\ \Leftrightarrow &\Delta \leq \frac{n}{2} + n_p \end{aligned} \tag{2.21}$$

De manière symétrique, le deuxième terme est retenu dans l'Équation 2.8 pour $CBS^{(i)}(u, k)$ si $\Delta \geq \frac{n_o}{2}$ et pour $CBS^{(i+1)}(u, k)$ si $\Delta \geq \frac{n_o}{2} + n_p$.

Ainsi, en utilisant les équations 2.11 et 2.13, il y a trois cas possibles :

1. si $\Delta \leq \frac{n_o}{2}$, alors le premier terme est retenu pour $CBS^{(i)}(u, k)$ et $CBS^{(i+1)}(u, k)$:

$$\begin{aligned} CBS^{(i+1)}(k, u) - CBS^{(i)}(k, u) &= (\pi_{i+1}[k] - \pi_{i+1}[u]) - (\pi_i[k] - \pi_i[u]) \\ &= (\pi_i[k] - n_p - \pi_i[u]) - (\pi_i[k] - \pi_i[u]) \\ &= -n_p \end{aligned} \tag{2.22}$$

2. si $\Delta \geq \frac{n}{2} + n_p$, alors le second terme est retenu pour $CBS^{(i)}(u,k)$ et $CBS^{(i+1)}(u,k)$:

$$\begin{aligned} CBS^{(i+1)}(k,u) - CBS^{(i)}(k,u) &= (n_o - (\pi_{i+1}[k] - \pi_{i+1}[u])) - (n_o - (\pi_i[k] - \pi_i[u]) \\ &= -(\pi_i[k] - n_p - \pi_i[u]) + (\pi_i[k] - \pi_i[u]) \\ &= n_p \end{aligned} \tag{2.23}$$

3. $\frac{n}{2} < \Delta < \frac{n_o}{2} + n_p$, alors le second terme est retenu pour $CBS^{(i)}(u,k)$ et le premier terme pour $CBS^{(i+1)}(u,k)$:

$$\begin{aligned} CBS^{(i+1)}(k,u) - CBS^{(i)}(k,u) &= (\pi_{i+1}[k] - \pi_{i+1}[u]) - (n_o - (\pi_i[k] - \pi_i[u])) \\ &= (\pi_i[k] - n_p - \pi_i[u]) - n_o + (\pi_i[k] - \pi_i[u]) \\ &= 2\Delta - (n_o + n_p) \end{aligned} \tag{2.24}$$

□

À chaque itération, seule une portion réduite des liens a besoin d'être parcourue, réduisant considérablement le temps de calcul.

4.3 Commentaires

4.3.1 Influence de l'initialisation

Malgré le caractère déterministe de la définition, l'heuristique peut laisser apparaître un comportement aléatoire : pour différentes exécutions sur un même graphe, l'étiquetage obtenu peut être différent, avec des variations significatives de la valeur atteinte de CBS. Cela arrive à cause de la numérotation initiale des nœuds : lorsqu'un tri est réalisé, suivant n'importe quel critère, si plusieurs nœuds ont la même valeur pour le critère, alors le premier nœud rencontré par l'algorithme est sélectionné avant les autres. Cela arrive (1) quand les nœuds sont triés en fonction de leur degré pour sélectionner le premier nœud d'un chemin, (2) quand plusieurs chemins ont la même longueur, et (3) quand l'insertion du chemin dans la séquence ordonnée peut se faire à plusieurs positions différentes. Par exemple pour (1), si les nœuds u et v ont tous les deux le degré le plus faible, mais que u a été rencontré avant v par l'algorithme, alors u est choisi pour débuter le chemin. Si v avait été rencontré avant, le chemin aurait été commencé par v et aurait mené vers une séquence de nœuds différente.

4.3.2 Chemins locaux contre solution globale

Un inconvénient de notre approche est qu'elle est basée sur des parcours locaux dans le graphe. Ainsi, l'algorithme ne peut pas aller vers un nœud qui n'est pas voisin avec le nœud précédent. La séquence ordonnée est ainsi très proche de la structure du graphe, comme souhaité. Néanmoins, la solution optimale est parfois incohérente avec la structure du graphe, par exemple en cas de de sauts importants, ou alors elle est cohérente mais en utilisant une organisation différente, non atteignable par l'heuristique. Comme notre motivation première est de suivre la structure du réseau, la séquence ordonnée obtenue peut mener à de mauvais résultats concernant la valeur optimale du CBS.

Dans la suite de ce chapitre, l'heuristique est désignée sous le nom de **mach**.

5 Implémentation algorithmique et complexité

5.1 Implémentation algorithmique

À partir d'un graphe $G = (V, E)$ connecté, l'algorithme retourne une séquence ordonnée des n nœuds. Les fonctions suivantes sont définies sur les nœuds du graphe :

— `Degré`(u) : retourne le degré du nœud u ;
— `Adj`(u) : retourne les nœuds adjacents de u ;

L'élément `liste` est utilisé comme structure de données, et consiste en une liste d'éléments distincts, disposant des méthodes suivantes :

— `Insertion-Liste`(A, a, $index$) : insertion de l'élément a dans la liste A à la position $index$. Si $index$ n'est pas donné, l'élément a est inséré en dernière position. a est une liste, auquel cas la liste est insérée à la position $index$ et fusionne avec la liste A ;
— `Suppression-Liste`(A, a) : suppression de l'élément a dans la liste A ;
— `Longueur`(A) : retourne le nombre d'éléments dans la liste A ;
— `Inversion`(A) : retourne la liste A dans le sens inverse.

Comme discuté dans la section précédente, l'heuristique consiste en l'exécution consécutive de deux étapes. Pour des raisons de lisibilité, l'implémentation algorithmique est discutée pour chacune des deux étapes de manière séparée.

Algorithme 5 : Étape 1 : Parcours du graphe à travers des nœuds localement similaires

Entrées : $G = (V, E)$ un graphe connecté, non-pondéré et non-dirigé
Sorties : $Chemins$ une liste de chemins indépendants

```
1  début
2      S ⟵ liste ; Insertion-Liste(S, V)
3      Chemin ⟵ liste
4      tant que S n'est pas vide faire
5          u_c ⟵ arg min_{u∈S} Degré(u)
6          Suppression-Liste(S, u_c)
7          continuer ⟵ Vrai
8          C ⟵ liste
9          tant que continuer = Vrai faire
10             Insertion-Liste(C, u_c)
11             H ⟵ liste
12             pour v ∈ Adj(u_c) ∩ S faire
13                 si Degré(v) = 1 alors
14                     Insertion-Liste(C, v)
15                     Suppression-Liste(S, v)
16                 sinon Insertion-Liste(H, v)
17
18             si H n'est pas vide alors u_c ⟵ arg max_{w∈H} Similarité(u, w)
19             sinon continuer ⟵ Faux
20         Insertion-Liste(Chemins, C)
```

L'Algorithme 5 présente l'implémentation de l'étape 1 décrite à la Section 4.1. La ligne 2 initialise une liste S contenant tous les nœuds du graphe, alors que la ligne 3 initialise une liste vide $Paths$ qui contiendra les chemins trouvés. Le parcours des chemins est réalisé de la ligne 4 à la ligne 20 et est

répété jusqu'à ce que tous les nœuds soient dans un chemin, c'est-à-dire lorsque la liste S est vide. Le nœud de S avec le degré le plus faible est choisi comme le début d'un chemin (ligne 5), et devient le nœud courant u_c. Il est ensuite enlevé de la liste S (ligne 6). Une variable $continuer$ est alors initialisée (ligne 7) et permet d'arrêter le parcours lorsque le chemin ne peut plus continuer. Ce chemin, défini ligne 8 et noté C, est défini comme une liste de nœuds, et se termine lorsqu'il n'y a plus de nœud possible vers lequel continuer à partir du nœud courant. La construction du chemin se fait de la ligne 9 à 19 : tant que le chemin peut continuer, le nœud courant est ajouté au chemin C (ligne 10). Une nouvelle liste H est définie (ligne 11) afin de contenir les potentiels nœuds successeurs au nœud courant. Ces successeurs sont sélectionnés parmi les nœuds adjacents de u_c qui ne sont pas encore dans un chemin, c'est-à-dire ceux qui sont toujours présents dans la liste S (ligne 12). Pour chacun de ces successeurs, notés v, si le degré de v est égal à 1, c'est-à-dire que le nœud u est seulement adjacent avec le nœud u_c, alors v est directement ajouté au chemin C (lignes 13 à 15). Sinon, il est ajouté à la liste H (ligne 16). Une fois tous les voisins parcourus, le choix du nœud suivant se fait parmi les nœuds dans H, en sélectionnant le nœud qui maximise la similarité avec le nœud courant u_c suivant l'Équation 2.10 (ligne 18). Si la liste H est vide, il n'y a pas de successeur possible et le chemin s'arrête (ligne 19). Le chemin est alors inséré dans la liste des chemins $Chemins$ (ligne 20), et l'algorithme boucle jusqu'à ce que la liste S soit vide, c'est-à-dire que tous les nœuds aient été insérés dans un chemin.

Algorithme 6 : Étape 2 : Fusion gloutonne des chemins

Données : $G = (V, E)$ un graphe connecté, non-pondéré et non-dirigé
Entrées : $Chemins$ une liste de chemins indépendants
Sorties : $Ordre$ une séquence ordonnée de nœuds

```
1 début
2     Ordre ⟵ liste
3     tant que Chemins n'est pas vide faire
4         C_c ⟵ arg max_{C∈Chemins} Longueur(C)
5         Suppression-Liste(Paths, C_c)
6         position, reverse ⟵ CBS-Incrémentiell(Ordre, C_c)
7         si reverse = Vrai alors Insertion-Liste(Ordre, Reverse(C_0), position)
8         sinon Insertion-Liste(Ordre, C_c, position)
```

L'Algorithme 6 présente la deuxième étape de l'heuristique **mach**. Une liste $Ordre$ est d'abord initialisée (ligne 2) et consistera à la séquence ordonnée partielle qui va peu à peu fusionner avec les chemins obtenus à l'étape 1. L'algorithme commence à parcourir les chemins pour les insérer dans la liste $Ordre$ (lignes 3 à 8). Le chemin courant C_c est choisi en sélectionnant le plus long chemin (ligne 4). Le chemin courant est ensuite enlevé de la liste des chemins (ligne 5). La position à laquelle le chemin C_c est inséré dans l'ordre partiel $Ordre$ est donné par la fonction `CBS-Incrémentiel`, ainsi que l'indication si le chemin doit être inversé ou non (ligne 6). À partir de ces informations, le chemin est inséré à la position donnée, soit à l'envers (ligne 7) soit à l'endroit (ligne 8). Lors du premier passage dans la boucle, la liste $Ordre$ est vide, et une seule position d'insertion est possible. Le processus est répété jusqu'à ce qu'il n'y ait plus de chemin dans la liste $Chemins$. La liste $Ordre$ contient alors une séquence ordonnée des nœuds du graphe G.

5.2 Étude de la complexité de l'algorithme

La complexité de l'Algorithme 5 est tout d'abord examinée. La liste S initialisée ligne 2 peut être implémentée comme une file de priorité minimale avec un tas binaire minimum, contenant les nœuds avec pour clé le degré de chaque nœud. Le temps nécessaire pour construire le tas S est $O(n)$. Les

lignes 5 et 6 peuvent être réalisées en utilisant la fonction d'extraction d'un élément, ayant pour temps $O(\log n)$. La liste $Chemins$ peut également être implémentée comme une file de priorité maximale avec un tas binaire maximum, contenant les chemins avec pour clé la longueur du chemin. La ligne 14 prend dans le pire des cas un temps $O(\log n)$. La boucle à la ligne 9 est exécutée au plus n fois. Les fonctions `Insertion-Liste` et `Suppression-Liste` sont en temps constant. La liste H des nœuds qui sont adjacents à u_c et sont dans S est implémentée comme une file de priorité maximale avec un tas binaire, permettant d'insérer un nouvel élément (ligne 16) avec un temps $O(\log(|H|))$, qui est dans le pire des cas $O(\log(|\texttt{Adj}(u_c)|))$. La boucle entre les lignes 9 et 17 est exécutée $|\texttt{Adj}(u_c)|$ fois, et à chaque itération :

1. le calcul de la similarité prend un temps $\Theta(\min(|\texttt{Adj}(u_c)|, |\texttt{Adj}[v]|))$
2. l'insertion prend un temps $O(\log(|\texttt{Adj}(u_c)|))$

Ainsi, la boucle est en temps $O(|\texttt{Adj}(u_c)|^2)$. La ligne 18 peut être réalisée en temps $O(\log(|\texttt{Adj}(u_c)|))$ et la complexité totale des lignes 9 à 19 est en temps $O(|\texttt{Adj}(u_c)|^2)$. On obtient alors un temps total de $O(\sum_{u\in V}(|\texttt{Adj}(u_c)|^2))$. Das a montré dans [66] que :

$$\sum_{u\in V} |\texttt{Adj}(u_c)|^2 \leq m\left(\frac{2m}{n-1} + n - 2\right) \tag{2.25}$$

On en conclut que le coût total de l'Algorithme 5 est $O(n\log n + mn) = O(mn)$.

Une analyse similaire est réalisée pour évaluer le temps pris par l'Algorithme 6. La ligne 4 est en $O(n)$, dans le pire cas.

La fonction `CBS_Incremental` parcourt :

1. tous les liens entre les nœuds du chemin courant C et ceux de `Ordre`. Dans le pire des cas, il y a n séparations possibles en O_1 et O_2. Pour chaque séparation, on va considérer les arêtes entre `Chemin` et O_1, puis entre `Chemin` et O_2. Si on agrège sur tous les chemins, on a les m arêtes du graphe, et on peut borner par $O(mn)$.
2. tous les liens entre $Ordre$ à la position $position$ et les autres nœud de $Ordre \cup C$. Comme on teste toutes les positions $position$ dans `Ordre`, dans le pire cas on le fait pour tous les liens du graphe. Comme cela est réalisé pour tous les chemins, on a dans le pire es les arêtes du graphe. Et on fait ça pour tous les chemins, soit dans le pire cas un temps de $O(mn)$.

Les autres instructions de la boucle sont exécutées en temps constants. Ainsi, le temps total passé dans l'Algorithme 6 est $O(mn)$.

Finalement, on obtient une complexité totale pour l'heuristique de $O(mn)$.

6 Évaluation de l'heuristique sur la minimisation du *Cyclic Bandwidth Sum*

Cette section décrit les expériences réalisées pour évaluer les performances de l'heuristique **mach**, en testant la capacité d'obtenir une bonne solution pour le *Cyclic Bandwidth Sum Problem*.

6.1 Processus expérimental

L'évaluation de l'heuristique est réalisée en utilisant le processus expérimental suivant : la valeur de CBS obtenue en utilisant **mach** est comparée à une valeur de référence, choisie parmi les résultats théoriques s'ils sont disponibles, ou sinon comme la valeur du CBS obtenue en utilisant une autre méthode. Afin d'évaluer les performances de l'algorithme sans être influencé par les identifiants initiaux

des graphes, qui sont souvent corrélés avec la structure du graphe, une étape de randomisation des identifiants est réalisée avant chaque exécution des heuristiques. 30 répétitions sont réalisées pour chaque instance de graphe.

Une comparaison est réalisée entre la valeur médiane du CBS atteinte par l'heuristique **mach** sur les 30 répétitions, notée *CBS médian*, et une valeur de référence, notée *ref CBS*, qui dépend du type de graphe étudié. Cette comparaison se fait en calculant une distance relative *rd* :

$$rd = \frac{\textit{CBS médian} - \textit{ref CBS}}{\textit{ref CBS}} \tag{2.26}$$

Le signe de *rd* indique si *CBS médian* est plus grand ($rd > 0$) ou plus petit ($rd < 0$) que la valeur de référence, alors que sa valeur indique la distance entre *CBS médian* et *ref CBS*. Par exemple, $rd = 0.80$ indique que *CBS médian* est 1.80 fois plus grand que la valeur de référence, alors que $rd = -0.25$ signifie que la valeur médiane du CBS obtenue est 1.25 fois plus petite que la valeur de référence.

Afin d'étudier la variabilité des résultats, la valeur absolue des écarts à la moyenne (*median absolute deviation*) de la valeur du CBS sur les répétitions, notée *mad CBS*, est également calculée, et est normalisée par la valeur de référence *ref CBS*. Cette valeur est notée *nmad* et est donnée par :

$$nmad = \frac{\textit{mad CBS}}{\textit{ref CBS}} \tag{2.27}$$

Si *nmad* est égal à 0, cela signifie que la valeur de CBS obtenue est stable sur toutes les répétitions. Au contraire, si *nmad* est différent de 0, cette valeur indique l'intervalle de variabilité du CBS autour de la médiane *CBS médian*. Par exemple, $nmad = 0.3$ indique que la valeur du CBS varie en moyenne de 0.3 fois la valeur de référence autour de la valeur médiane.

6.2 Jeux de données

6.2.1 Graphes avec une valeur optimale du CBS connue

Graphe chemin La valeur optimale du CBS pour un chemin P_n de taille n est $\text{CBS}_{\text{opt}}(P_n) = n - 1$. La collection comprend tous les chemins jusqu'à 448 nœuds.

Graphe cycle La valeur optimale du CBS pour un cycle C_n de taille n est $\text{CBS}_{\text{opt}}(C_n) = n$. La collection comprend tous les cycles jusqu'à 448 nœuds.

Graphe roue La valeur optimale du CBS pour une roue avec n nœuds est $\text{CBS}_{\text{opt}}(W_n) = n + \lfloor \frac{1}{4}n^2 \rfloor$, prouvé dans [135]. La collection comprend toutes les roues jusqu'à 448 nœuds.

Graphe k-régulier La valeur optimale du CBS pour un cycle à la puissance k est $\text{CBS}_{\text{opt}}(C_n^k) = \frac{1}{2}nk(k+1)$ prouvé dans [135]. La collection comprend toutes les cycles jusqu'à 448 nœuds à la puissance k, avec $k \in \{2, 10\}$.

Graphe biparti complet Chen et al. [56] ont montré que la valeur optimale du CBS pour un graphe complet bipartite $K_{n_1 n_2}$ est donnée par :

$$\text{cbs}_{\text{opt}}(K_{n_1 n_2}) = \begin{cases} \frac{n_1 n_2^2 + n_1^2 n_2}{4} & \text{si } n_1 \text{ et } n_2 \text{ sont pairs} \\ \frac{n_1 n_2^2 + n_1^2 n_2 + n_1}{4} & \text{si } n_1 \text{ est pair et } n_2 \text{ est impair} \\ \frac{n_1 n_2^2 + n_1^2 n_2 + n_1 + n_2}{4} & \text{si } n_1 \text{ et } n_2 \text{ sont impairs} \\ \frac{n_1 n_2^2 + n_1^2 n_2 + n_2}{4} & \text{si } n_1 \text{ est impair et } n_2 \text{ est pair} \end{cases}$$

La collection comprend tous les graphes bipartis complets jusqu'à 448 nœuds, avec trois différent ratios entre la taille des deux ensembles : 1, 3, et 7. Seules les valeurs de n_1 et n_2 admissibles pour les différents ratios sont retenues.

6.2.2 Graphes avec une borne supérieure pour la valeur optimale du CBS

Produits cartésiens Jianxiu [135] a montré des bornes supérieures pour la valeur optimale du CBS lorsque G et H sont un graphe chemin, un graphe cycle ou un graphe complet :

$$\text{CBS}_{\text{opt}}(P_{n_G} \times P_{n_H}) \leq n_G(n_H - 1) + n_H^2(n_G - 1), \quad n_G \geq n_H \tag{2.28}$$

$$CBS_{\text{opt}}(C_{n_G} \times C_{n_H}) \leq n_G(n_H^2 + 2n_H - 2), \quad n_G \geq n_H \geq 3 \tag{2.29}$$

$$CBS_{\text{opt}}(K_{n_G} \times K_{n_H}) \leq \frac{1}{6} n_G n_H \left(n_H^2 + 3n_H \left\lfloor \frac{n_G}{2} \right\rfloor \left\lceil \frac{n_G}{2} \right\rceil - 1 \right), \quad n_G \geq n_H \tag{2.30}$$

$$CBS_{\text{opt}}(P_{n_G} \times C_{n_H}) \leq n_H(n_G^2 + n_G - 1) \tag{2.31}$$

$$CBS_{\text{opt}}(P_{n_G} \times K_{n_H}) \leq \frac{1}{2} n_G^2 n_H \left\lfloor \frac{n_H}{2} \right\rfloor \left\lceil \frac{n_H}{2} \right\rceil + n_H(n_G - 1) \tag{2.32}$$

$$CBS_{\text{opt}}(C_{n_G} \times K_{n_H}) \leq n_H \left(\frac{1}{2} n_G^2 \left\lfloor \frac{n_H}{2} \right\rfloor \left\lceil \frac{n_H}{2} \right\rceil + 2n_G - 2 \right) \tag{2.33}$$

La collection comprend les produits cartésiens des graphes décrits plus haut, avec n_G et n_H allant jusqu'à 25, en considérant les contraintes spécifiques sur n_G et n_H si nécessaire.

6.2.3 Graphes avec une valeur optimale du CBS inconnue

Graphes aléatoires La collection comprend des graphes aléatoires de type Erdős-Rényi connectés construits pour les valeurs de $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. 10 instances du modèle d'Erdös-Rényi sont générés pour chaque valeur de p, en réglant le nombre de nœuds à 100.

Harwell-Boeing collection La collection de matrices creuses Harwell-Boeing [81] consiste en un ensemble de matrices d'adjacence issues de problèmes scientifiques et d'ingénierie. 27 matrices binaires des sous-collections *bcspwr*, *dwt*, et *can*, allant de petits graphes (24 nœuds) à des graphes plus conséquents (1454 nœuds) ont été sélectionnées, représentant une large variété de structures.

6.3 Performances de l'heuristique mach

6.3.1 Comparaison avec les résultats théoriques sur la valeur optimale du CBS

L'heuristique **mach** atteint la valeur optimale du CBS donnée par les résultats théoriques pour toutes les instances des graphes chemins, graphes cycles, graphes roues, graphes k-réguliers et graphes bipartis complets.

6.3.2 Comparaison avec la borne supérieure de la valeur optimale du CBS

Les valeurs de CBS obtenues avec l'heuristique **mach** sont comparées avec les bornes supérieures théoriques données par Jianxiu [135].

La Table 2.1 affiche les résultats obtenus : chaque ligne concerne un type de produit cartésien. Les deux premières colonnes donnent les deux types de graphes, la troisième colonne donne le nombre de graphes dans la sous-collection, correspondant à tous les couples de valeurs n_G et n_H. Enfin, les quatrième et cinquième colonnes donnent la valeur de *rd* et de *nmad* moyennées sur les graphes de chaque sous-collection.

G	H	*# graphes*	*rd* moyen	*nmad* moyen
C	C	210	0.83	0.16
C	K	400	-0.83	0.00
P	K	400	-0.84	0.00
P	P	210	0.76	0.19
K	K	210	-0.20	0.00
P	C	400	0.50	0.10

TABLE 2.1 – Performance de l'heuristique **mach** sur les produits cartésiens de graphes, cycles et graphes complets. Les deux premières colonnes donnent les deux types de graphes, la troisième colonne donne le nombre de graphes dans la sous-collection, correspondant à tous les couples de valeurs n_G et n_H. Enfin, les quatrième et cinquième colonnes donnent la valeur de *rd* et de *nmad* moyennées sur les graphes de chaque sous-collection.

Ces résultats mettent en valeur la différence de comportement de l'heuristique **mach** en fonction de la topologie du graphe. Quand le graphe est organisé en une succession de sous-graphes structurés linéairement, l'heuristique **mach** est très efficace : dans le cas de produits cartésiens d'un graphe complet avec soit un graphe cycle, soit un graphe chemin, le graphe résultant est une succession de cliques organisées dans un arrangement linéaire ou cyclique. L'heuristique **mach** est alors guidée par cette structure et découvre la structure sous-jacente efficacement, permettant d'obtenir des valeurs de CBS en dessous de la borne supérieure théorique.

En revanche, lorsque la structure présente des régularités qui ne sont pas suivant un arrangement linéaire ou cyclique, l'heuristique **mach** échoue à trouver des valeurs de CBS en dessous de la borne théorique supérieure. Cela arrive par exemple pour la grille (produit cartésien de graphes chemins), le tore (produit cartésiens de graphes cycles) et le cycle (produit cartésien d'un chemin et d'un cycle) : l'heuristique **mach** a, du fait des symétries, plusieurs manières de parcourir le graphe. La structure est néanmoins parcourue correctement, mais pas de la manière qui minimise la valeur de CBS. Ces différences de performances se retrouvent également dans la variabilité du résultat obtenu : lorsque les performances sont bonnes, la variabilité est faible, et au contraire la variabilité augmente lorsque la valeur du CBS atteinte est au-dessus de la borne théorique supérieure du CBS optimal.

6.3.3 Comparaison avec l'heuristique gvns

Une comparaison avec l'heuristique **gvns** développée par Satsangi et al. [213] est réalisée pour les graphes pour lesquels il n'existe pas de résultats théoriques sur la valeur optimale du CBS, à savoir les graphes aléatoires et les graphes de la collection Harwell-Boeing. Afin d'évaluer les performances de l'heuristique **mach**, l'heuristique **gvns** est exécutée en utilisant le code fourni par les auteurs. Pour chaque instance de graphe, la valeur minimale du CBS sur 50 répétitions est retenue, comme défini dans [213]. Le résultat obtenu est la valeur de référence utilisée pour calculer la distance relative *rd*.

La Table 2.2a présente les résultats obtenus pour les graphes de la collection Harwell-Boeing, et la Table 2.2b pour la collection de graphes aléatoires. Les deux premières colonnes donnent le nom de la sous-collection et le nombre de graphes dans la sous-collection, tandis que les troisième et quatrième colonnes donnent les valeurs de *rd* et de *nmad* moyennées sur les graphes de chaque sous-collection.

Les résultats de l'heuristique **mach** sur les graphes de la collection Harwell-Boeing sont meilleurs que ceux obtenus en utilisant l'heuristique **gvns**. Ces graphes sont en effet très structurés, et l'heuristique **mach** est adaptée pour suivre cette structure. Inversement, les résultats pour les graphes aléatoires sont en moyenne moins bons que **gvns**, d'autant plus que la densité diminue. Néanmoins, les graphes aléatoires n'ont, par définition, pas de structure, ce qui explique que l'heuristique **mach** ait du mal à trouver une valeur du CBS satisfaisante.

Nom	*# graphes*	*rd* moyen	*nmad* moyen
bcspwr	6	-0.76	0.01
dwt	9	-0.74	0.00
can	12	-0.55	0.01

(a) Collection Harwell-Boeing

p	*# graphes*	*rd* moyen	*nmad* moyen
0.1	10	0.23	0.00
0.3	10	0.10	0.00
0.5	10	0.08	0.00
0.7	10	0.05	0.00
0.9	10	0.02	0.00

(b) Graphes aléatoires

TABLE 2.2 – Performance de l'heuristique **mach** sur les graphes aléatoires et les graphes de la collection Harwell-Boeing. Les deux premières colonnes donnent le nom de la sous-collection et le nombre de graphes dans la sous-collection, tandis que les troisième et quatrième colonnes donnent les valeurs de *rd* et de *nmad* moyennées sur les graphes de chaque sous-collection.

7 Applications à des réseaux complexes

Dans cette section, une évaluation des performances de l'heuristique **mach** est réalisée sur des réseaux complexes. Contrairement à la section précédente, les expériences sont guidées par notre motivation de découvrir la structure des réseaux. La cohérence entre la topologie du graphe et la séquence ordonnée des nœuds obtenue est évaluée en utilisant des indicateurs spécifiques à chaque structure complexe. L'heuristique **mach** est comparée avec l'heuristique **hla** [18], conçu pour résoudre le problème de l'arrangement linéaire (*Bandwidth Sum Problem*). Une illustration est ensuite proposée sur un grand réseau, à l'aide d'une représentation visuelle mettant en valeur la cohérence entre ordre des nœuds et structure.

7.1 Comparaison avec l'heuristique hla sur des structures complexes

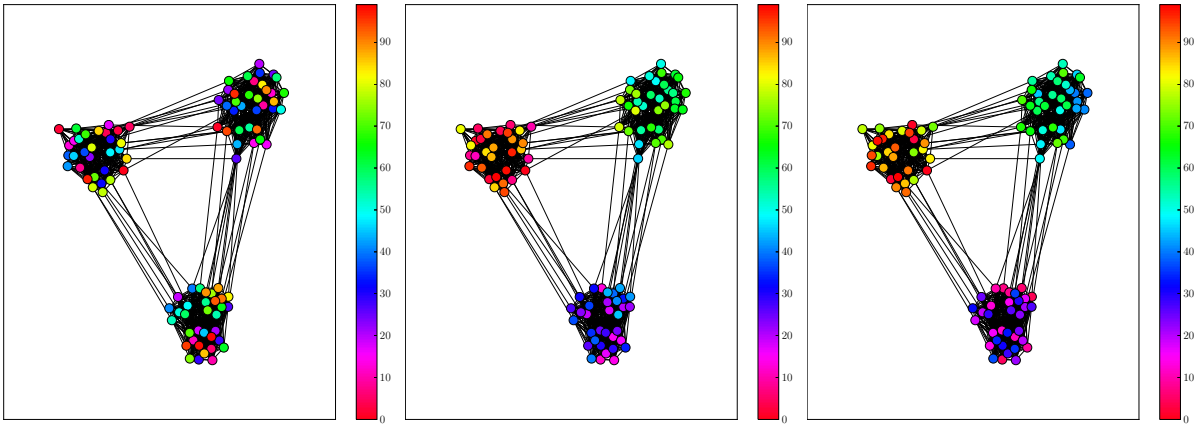
Dans cette section, une comparaison entre les heuristiques **mach** et **hla** est réalisée afin d'évaluer la capacité à retrouver la structure du graphe à travers l'étiquetage. Deux types de structures complexes sont étudiés, souvent rencontrés dans le cas d'analyse de données réelles : la structure en communautés [97] et la propriété petit-monde [247]. Pour chacune de ces deux structures, un indicateur spécifique permettant d'évaluer la cohérence entre la séquence ordonnée des nœuds et la topologie est introduite, et est dénotée CLS : une faible valeur de CLS signifie que la cohérence est élevée. En utilisant les grandeurs définies par les équations 2.26 et 2.27, les valeurs *rd* et *nmad* sont calculées, à la fois pour le CBS et le CLS. Dans les expériences suivantes, plusieurs graphes avec 200 nœuds sont générés pour différentes valeurs des paramètres du modèle utilisé. Pour toutes les combinaisons de ces paramètres, 10 instances sont générées. Pour chacune des instances, les heuristiques **mach** et **hla** sont exécutés 10 fois : comme précédemment, les identifiants initiaux sont choisis aléatoirement avant chaque répétition.

Structure en communautés Les valeurs pour les paramètres du modèle à blocs stochastiques présentée à la Section 2.4.2 sont les suivantes : C est compris entre 2 et 5, p_{intra} est à valeurs dans $\{0.7, 0.8, 0.9\}$, et p_{inter} est fixé à 0.05. Les valeurs de références pour le calcul de *rd* et *nmad* sont choisies comme étant celle pour un graphe pour lequel les nœuds sont numérotés par communauté, l'une après l'autre.

La Table 2.3 affiche les résultats obtenus : les deux premières colonnes indiquent le nombre de communautés et la valeur de p_{intra}. Les quatre colonnes suivantes donnent les valeurs de *rd* et *nmad*, à la fois pour le CBS et le CLS, moyennées sur tous les résultats obtenus par les deux heuristiques **mach** et **hla**. Les résultats montrent que les deux heuristiques sont capables de retrouver la structure en communautés, étant donné que la valeur de CLS utilisée pour mesurer la cohérence entre structure et

Collection		**mach**				**hla**			
		CBS		CLS		CBS		CLS	
C	p_{intra}	*rd moyen*	*nmad moyen*	*rd moyen*	*nmad moyen*	*rd moyen*	*nmad moyen*	*rd moyen*	*nmad moyen*
2	0.7	-0.11	0.02	0.00	0.01	-0.67	0.01	0.00	0.00
2	0.8	-0.04	0.01	0.00	0.00	-0.51	0.01	0.00	0.01
2	0.9	-0.03	0.01	0.00	0.00	-0.34	0.01	0.00	0.00
3	0.7	-0.11	0.03	0.00	0.01	-0.67	0.01	0.00	0.01
3	0.8	-0.08	0.04	0.01	0.02	-0.56	0.02	0.00	0.01
3	0.9	-0.02	0.02	0.01	0.01	-0.41	0.01	0.00	0.01
4	0.7	-0.19	0.05	0.07	0.03	-0.70	0.02	0.00	0.01
4	0.8	-0.05	0.05	0.05	0.03	-0.52	0.02	0.00	0.02
4	0.9	-0.06	0.02	0.00	0.01	-0.39	0.02	0.00	0.01
5	0.7	-0.09	0.08	0.17	0.07	-0.56	0.02	0.15	0.02
5	0.8	-0.08	0.07	0.08	0.05	-0.45	0.02	0.23	0.01
5	0.9	-0.07	0.04	0.01	0.03	-0.24	0.02	0.28	0.02

TABLE 2.3 – Comparaison entre les heuristiques **mach** et **hla** pour découvrir la structure d'un réseau avec des communautés. Les deux premières colonnes indiquent le nombre de communautés et la valeur de p_{intra}. Les quatre colonnes suivantes donnent les valeurs de *rd* et *nmad*, à la fois pour le CBS et le CLS, moyennées sur tous les résultats obtenus pour chaque valeur de C par les deux heuristiques **mach** et **hla**.



(a) Étiquetage aléatoire (b) Étiquetage obtenu en utilisant **mach** (c) Étiquetage obtenu en utilisant **hla**

FIGURE 2.7 – Représentation d'un réseau avec 100 nœuds divisés en 3 communautés, obtenu en utilisant $p_{\text{intra}} = 0.9$ et $p_{\text{inter}} = 0.05$. Les couleurs indiquent l'étiquette des nœuds comme à la Figure 2.2.

étiquetage est très proche de 0. L'heuristique **hla** semble légèrement meilleure pour minimiser la valeur de CBS, et donne également des résultats légèrement plus stables.

La Figure 2.7 affiche une représentation d'un réseau avec 100 nœuds divisés en 3 communautés, obtenu en utilisant $p_{\text{intra}} = 0.9$ et $p_{\text{inter}} = 0.05$. Les couleurs indiquent l'étiquette des nœuds comme à la Figure 2.2. Il n'y a pas de différences visuelles entre les étiquetages produits par les deux heuristiques, les deux étant capables de refléter la structure en communautés.

Réseau petit monde Les paramètres suivants sont utilisés pour le modèle de Watts-Strogatz présenté à la Section 2.4.2 dans les expériences suivantes : k comprend les valeurs paires entre 2 et 10, et p est à valeurs dans $\{0.0, 0.1, 0.2, 0.3\}$. L'indicateur CLS est calculé en prenant en compte uniquement les nœuds formant l'anneau régulier de degré k, enlevant tous les liens issus de la deuxième étape. Les valeurs de références sont calculées en prenant en compte l'étiquetage initial, suivant l'anneau.

Collection		**mach**				**hla**			
		CBS		CLS		CBS		CLS	
Degré	*p*	*rd moyen*	*nmad moyen*	*rd moyen*	*nmad moyen*	*rd moyen*	*nmad moyen*	*rd moyen*	*nmad moyen*
2	0.0	0.00	0.00	0.00	0.00	0.99	0.00	0.99	0.00
2	0.1	4.92	0.83	1.19	0.54	4.38	0.72	2.92	0.53
2	0.2	9.08	0.74	3.29	0.81	7.90	0.56	4.95	0.72
2	0.3	13.85	0.90	5.47	0.75	11.80	0.71	7.04	0.57
4	0.0	0.00	0.00	0.00	0.00	1.01	0.02	1.01	0.02
4	0.1	3.33	0.35	0.00	0.00	4.22	0.47	1.38	0.33
4	0.2	6.95	0.52	0.00	0.00	7.32	0.56	2.00	0.30
4	0.3	9.77	0.81	0.01	0.01	10.19	0.65	2.84	0.46
6	0.0	0.00	0.00	0.00	0.00	0.99	0.02	0.99	0.02
6	0.1	2.60	0.22	0.00	0.00	3.14	0.31	1.01	0.02
6	0.2	5.35	0.48	0.06	0.05	5.87	0.41	1.35	0.31
6	0.3	7.87	0.59	0.30	0.25	8.13	0.47	1.58	0.40
8	0.0	0.00	0.00	0.00	0.00	1.01	0.02	1.01	0.02
8	0.1	2.21	0.20	0.05	0.05	2.74	0.12	0.97	0.07
8	0.2	4.58	0.42	0.54	0.39	5.10	0.33	1.04	0.02
8	0.3	6.81	0.23	0.49	0.22	6.93	0.36	1.06	0.04
10	0.0	0.00	0.00	0.00	0.00	1.01	0.03	1.01	0.03
10	0.1	1.91	0.19	0.13	0.13	2.52	0.18	1.01	0.04
10	0.2	4.20	0.47	0.65	0.34	4.12	0.37	1.04	0.02
10	0.3	5.98	0.37	0.96	0.40	5.95	0.27	1.03	0.09
12	0.0	0.00	0.00	0.00	0.00	1.00	0.02	1.00	0.02
12	0.1	1.60	0.16	0.20	0.15	2.36	0.08	1.02	0.02
12	0.2	3.79	0.35	0.89	0.29	3.81	0.19	1.03	0.02
12	0.3	5.65	0.41	1.17	0.43	5.29	0.18	0.78	0.26

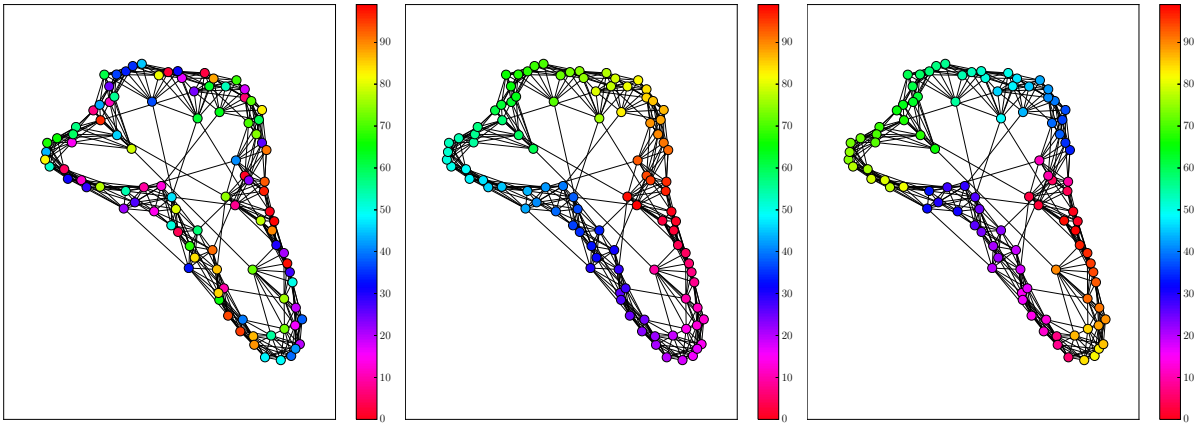
TABLE 2.4 – Comparaison entre les heuristiques **mach** et **hla** pour découvrir la structure petit monde. Les deux premières colonnes indiquent la valeur de k et de p. Les quatre colonnes suivantes donnent les valeurs de *rd* et *nmad*, à la fois pour le CBS et le CLS, moyennées sur tous les résultats obtenus par les deux heuristiques **mach** et **hla**.

La Table 2.4 donne les résultats présentés comme dans la Table 2.3. Les résultats dans cette partie tournent à l'avantage de l'heuristique **mach**, qui obtient des meilleurs résultats sur la minimisation du CLS par rapport à **hla** dans la majorité des configurations. On peut noter cependant que **hla** est meilleur pour minimiser la valeur de CBS, mettant en évidence que la solution globale du *Cyclic Bandwidth Sum Problem* n'est pas la plus cohérente avec la structure.

Comme la Figure 2.7, la Figure 2.8 montre une représentation d'un réseau petit monde avec 100 nœuds, obtenu en utilisant $k = 10$ et $p = 0.1$. Les trois illustrations mettent en valeur tout d'abord les différences en termes d'étiquetage des nœuds entre un étiquetage aléatoire et un étiquetage ordonné. Elles montrent également que **mach** est plus capable de suivre la structure du réseau : les couleurs sont plus homogènes selon l'anneau en utilisant **mach** plutôt qu'en utilisant **hla**.

Structure en communautés et propriété petit monde Des réseaux avec à la fois une structure en communauté et la propriété petit monde sont considérés en utilisant un modèle mixte de Watts-Strogatz à blocs stochastiques. Des réseaux avec 200 nœuds sont générés : le degré k est fixé à 6 et les communautés sont générées en utilisant $p_{\text{intra}} = 0.9$. Le nombre de communautés C varie de 2 à 5 et la probabilité p_{inter} est à valeurs dans $\{0.0, 0.1, 0.2, 0.3\}$. L'indicateur CLS est choisi comme l'addition des deux indicateurs définis précédemment, à savoir la valeur du CBS à l'intérieur de chaque communauté plus la valeur du CBS pour les nœuds formant l'anneau 6-régulier. Les valeurs de références sont les valeurs de CBS et CLS obtenues quand les nœuds sont étiquetés en utilisant l'ordre initial suivant l'anneau.

La Table 2.5 donne les résultats comme décrit dans la Table 2.3. Ces résultats sont proches de ceux

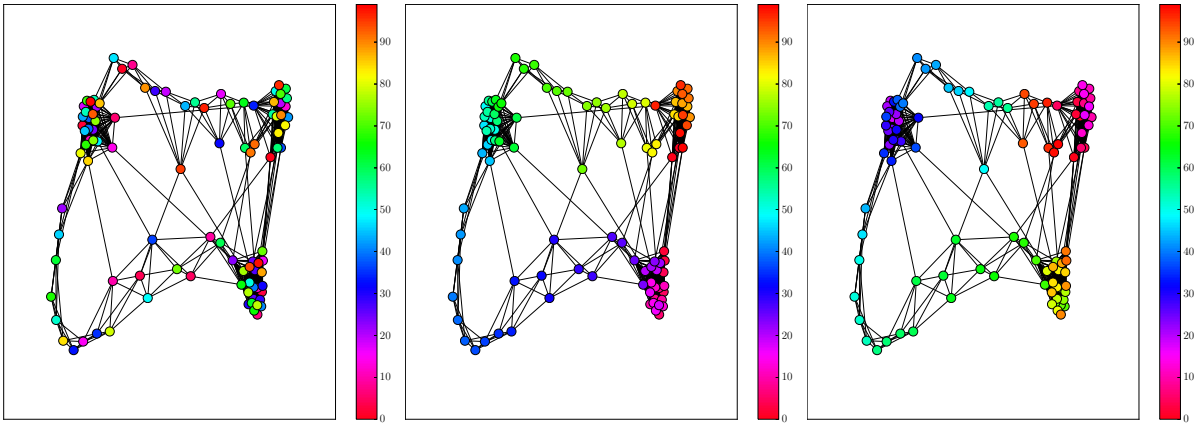


(a) Étiquetage aléatoire (b) Étiquetage obtenu en utilisant **mach** (c) Étiquetage obtenu en utilisant **hla**

FIGURE 2.8 – Représentation d'un réseau petit monde avec 100 nœuds, obtenu en utilisant $k = 10$ et $p = 0.1$. Les couleurs indiquent l'étiquette des nœuds comme à la Figure 2.2.

Collection		**mach**				**hla**			
		CBS		CLS		CBS		CLS	
C	*p*	*rd moyen*	*nmad moyen*	*rd moyen*	*nmad moyen*	*rd moyen*	*nmad moyen*	*rd moyen*	*nmad moyen*
2	0.00	0.26	0.46	0.24	0.00	0.22	0.44	0.24	0.00
2	0.10	-0.09	0.24	-0.14	0.00	-0.13	0.23	-0.14	0.00
2	0.20	0.11	0.47	0.05	0.00	0.08	0.50	0.05	0.00
2	0.30	0.18	0.28	0.10	0.00	0.14	0.29	0.10	0.00
3	0.00	0.24	0.37	0.02	0.00	0.19	0.29	0.02	0.00
3	0.10	1.10	0.64	0.59	0.00	1.06	0.44	0.60	0.00
3	0.20	0.25	0.47	-0.11	0.00	0.18	0.35	-0.09	0.00
3	0.30	0.53	0.45	0.02	0.00	0.49	0.43	0.03	0.00
4	0.00	0.61	0.53	0.09	0.00	0.61	0.62	0.11	0.00
4	0.10	0.79	0.30	0.02	0.00	0.65	0.20	0.01	0.00
4	0.20	0.50	0.26	-0.20	0.00	0.44	0.24	-0.19	0.00
4	0.30	1.06	0.35	-0.02	0.00	0.89	0.38	0.00	0.00
5	0.00	1.76	0.58	0.49	0.00	1.61	0.51	0.51	0.00
5	0.10	1.40	0.59	0.11	0.00	1.29	0.53	0.20	0.00
5	0.20	2.54	0.64	0.49	0.00	2.35	0.63	0.54	0.00
5	0.30	2.46	0.52	0.28	0.00	2.23	0.53	0.30	0.00

TABLE 2.5 – Comparaison entre les heuristiques **mach** et **hla** pour découvrir la structure petit monde avec communautés. Les deux premières colonnes indiquent le nombre de communautés et la valeur de p_{intra}. Les quatre colonnes suivantes donnent les valeurs de *rd* et *nmad*, à la fois pour le CBS et le CLS, moyennées sur tous les résultats obtenus par les deux heuristiques **mach** et **hla**.



(a) Étiquetage aléatoire (b) Étiquetage obtenu en utilisant **mach** (c) Étiquetage obtenu en utilisant **hla**

FIGURE 2.9 – Représentation d'un réseau petit monde avec 100 nœuds avec 3 communautés, obtenu en utilisant $k = 6$ et $p = 0.1$. Les couleurs indiquent l'étiquette des nœuds comme à la Figure 2.2.

obtenus précédemment : **mach** obtient de meilleurs résultats pour la minimisation du CLS par rapport à **hla**, montrant qu'il est mieux adapté pour suivre la structure de ce type de réseau. Inversement, l'heuristique **hla** est meilleure pour minimiser la valeur de CBS.

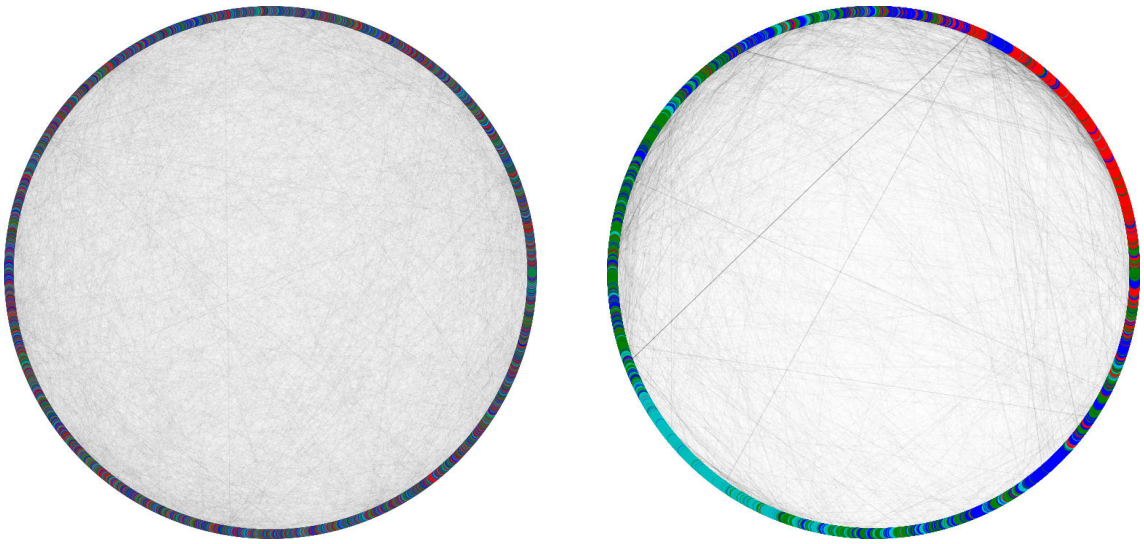
Figure 2.9 affiche une représentation d'un graphe petit monde avec 100 nœuds séparés en 3 communautés, en utilisant $k = 6$ et $p = 0.1$. Comme pour la Figure 2.8, les étiquettes des nœuds sont en cohérence avec la structure du graphe, car elles suivent à la fois la structure cyclique du nœud tout en parcourant de façon adéquate les communautés. Au contraire, l'heuristique **hla** retranscrit moins bien la structure du réseau.

Commentaires Au vu des résultats obtenus dans cette section, on se rend compte que l'heuristique **mach** est bien adaptée pour trouver un étiquetage des nœuds en cohérence avec la structure du réseau, sur des réseaux complexes possédant des propriétés petit monde ou une structure en communautés. Néanmoins, la méthode **hla**, conçue pour résoudre le *Bandwidth Sum Problem* qui est un problème très proche du *Cyclic Bandwidth Sum Problem*, semble meilleure que **mach** pour minimiser la valeur de CBS. Ce résultat n'est cependant pas incompatible avec la motivation initiale : en effet, l'objectif est de pouvoir suivre la structure du graphe, et la méthode proposée consiste à se servir du *Cyclic Bandwidth Sum Problem* comme un substitut afin de mieux définir le problème. Cette observation illustre le commentaire discuté à la fin de la Section 4 : le parcours local est favorisé au détriment du parcours global, de manière à préserver au maximum la topologie du réseau dans l'étiquetage. Les résultats obtenus dans cette section confirment le bien-fondé de cette approche.

7.2 Illustration sur un grand réseau

L'heuristique **mach** est appliquée sur un réseau issu de la collection SNAP [150]. Ce réseau est construit à partir des données bibliographiques de la base DBLP regroupant des publications informatiques [257] : deux auteurs sont connectés s'ils ont publié au moins un article ensemble. Des communautés sont définies à partir des journaux ou/et conférences dont les publications sont issues. Les quatre plus grosses communautés ont été retenues, permettant d'obtenir un réseau avec 21969 nœuds et 69956 liens. Ce réseau a une densité très faible et un coefficient de clustering élevé, qui sont des propriétés usuelles des réseaux complexes.

La Figure 2.10 affiche deux représentations circulaires du réseau DBLP : la position d'un nœud autour du cercle est donnée par son étiquette, de telle sorte que des nœuds avec des étiquettes proches



(a) Étiquetage initial

(b) Étiquetage obtenu par l'heuristique **mach**

FIGURE 2.10 – Représentation circulaire du réseau DBLP [256], restreint à 21969 nœuds appartenant à une des quatre plus grandes communautés. Deux étiquetages différents sont proposés, le premier issu des données initiales, et le deuxième obtenu avec l'heuristique **mach**. La position d'un nœud autour du cercle est donnée par son étiquette, de telle sorte que des nœuds avec des étiquettes proches sont proches sur le cercle. La couleur des nœuds indique la communauté (rouge, bleu, cyan ou vert) et est affichée en semi-transparence. Les liens entre les nœuds sont représentés en semi-transparence.

sont proches sur le cercle. La couleur des nœuds indique la communauté (rouge, bleu, cyan ou vert) et est affichée en semi-transparence. Les liens entre les nœuds sont représentés en semi-transparence. Comme le nombre de nœuds est important, les nœuds proches se chevauchent, et leurs couleurs s'additionnent : plus la couleur d'une communauté est intense, plus le nombre de nœuds appartenant à cette communauté est important. Inversement, une couleur sombre indique que les nœuds appartiennent à des communautés différentes.

La différence entre la représentation utilisant l'étiquetage initial et l'étiquetage obtenu avec l'heuristique **mach** est claire : avec l'étiquetage initial, les nœuds du cercle sont sombres, indiquant qu'aucune région du cercle n'est occupée par des nœuds appartenant à une même communauté. Au contraire, avec l'étiquetage obtenu par l'heuristique **mach**, de nombreuses régions du cercle affiche une couleur distincte et intense, indiquant que les nœuds avec une étiquette proche appartiennent à la même communauté. Les quatre couleurs, correspondant aux quatre communautés, sont distinguables, et même si certaines communautés sont divisées, la plupart des nœuds ont une étiquette proche des étiquettes des nœuds d'un même communauté. De plus, la répartition des liens à l'intérieur du disque formé par les nœuds montre également que lorsque l'étiquetage obtenu par **mach** est utilisé, les liens sont situés à la périphérie du disque, indiquant que la segmentation des nœuds en communautés est conforme à l'étiquetage.

Afin de confirmer la cohérence entre l'étiquetage et la structure en communautés du réseau DBLP mis en évidence à la Figure 2.10, la valeur de CBS entre les nœuds appartenant à des communautés différentes, ainsi qu'à l'intérieur de chaque communauté. La valeur de CBS est normalisée par le nombre de liens mis en jeu dans le calcul du CBS, afin d'avoir une valeur du CBS moyen par lien. La Figure 2.11 affiche la table de contingence correspondante, dans laquelle la couleur représente la valeur obtenue, allant du bleu (valeur faible) vers le rouge (valeur élevée). À l'intérieur de chaque communauté, la valeur de CBS est faible comparée au nombre de liens, signifiant que les nœuds ont des étiquettes proches. Au contraire, les liens reliant des nœuds appartenant à des communautés différentes ont une valeur de CBS

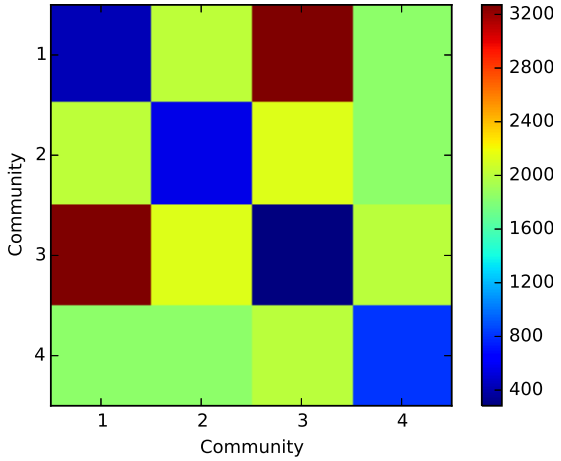


FIGURE 2.11 – Table de contingence affichant la valeur de CBS entre les nœuds appartenant à des communautés différentes, ainsi qu'à l'intérieur de chaque communauté. La valeur de CBS est normalisée par le nombre de liens mis en jeu dans le calcul du CBS, afin d'avoir une valeur du CBS moyen par lien. La couleur représente la valeur obtenue, allant du bleu (valeur faible) vers le rouge (valeur élevée).

plus élevée, signifiant que leurs étiquettes sont éloignées. Ce résultat confirme que l'heuristique **mach** est capable de retranscrire dans l'étiquetage la structure en communautés du réseau DBLP.

8 Conclusion et perspectives

La topologie d'un réseau est un élément crucial dans la compréhension du système qu'il représente. Si elle peut se décrire à l'aide de mesures sur le graphe, il n'existe pas de façon simple d'obtenir un étiquetage des nœuds cohérent avec la structure du graphe. L'heuristique proposée répond à ce problème, en proposant une approche à la fois avec des recherches locales et globales. Des résultats empiriques ont permis de valider son comportement vis à vis des objectifs visés.

Plusieurs extensions de l'algorithme peuvent être considérées, et notamment celle pour les graphes pondérés. Il n'existe pas de résultats théoriques à propos du *Cyclic Bandwidth Sum Problem* dans ce cas. Il est néanmoins pertinent de s'intéresser à cette classe de graphes, car elle se révèle être assez commune dans l'étude des réseaux complexes.

Le problème du *Cyclic Bandwidth Sum Problem* peut être étendu en prenant en compte le poids de chaque lien dans le calcul de la somme des différences des étiquettes :

$$\min_{\pi} f(\pi) \quad \text{avec} \quad f(\pi) = \sum_{\{u,v\}\in E} w_{uv} d_H(\pi(u), \pi(v)) \tag{2.34}$$

L'heuristique **mach** développée peut être facilement étendue au cas pondéré. Deux calculs doivent être modifiés :

1. le calcul de l'indice de similarité dans l'étape 1 ;
2. le calcul incrémentiel du CBS dans l'étape 2.

Le premier problème est résolu en définissant un nouvel indice de similarité entre deux nœuds u et v :

$$J_w(u,v) = \frac{N(u,v)}{D(u,v)} \tag{2.35}$$

où $N(u,v)$ représente les poids des voisins partagés par les deux nœuds et est défini par :

$$N(u,v) = 2w_{u,v} + \sum_{\substack{x \in V \\ x \sim u \\ x \sim v}} \min(w_{ux}, w_{vx}) \tag{2.36}$$

et $D(u,v)$ représente la somme totale des poids du voisinage de u et v, et est défini par :

$$D(u,v) = 2w_{u,v} + \sum_{\substack{x \in V \\ x \sim u \\ x \sim v}} \frac{w_{ux} + w_{vw}}{2} + \sum_{\substack{x \in V \\ x \sim u \\ x \nsim v}} w_{ux} + \sum_{\substack{x \in V \\ x \nsim u \\ x \sim v}} w_{vx} \tag{2.37}$$

On remarque que cette définition de l'indice de similarité dans le cas pondéré reste cohérente avec la mesure définie à l'Équation 2.10 pour le cas non-pondéré, puisque si tous les poids sont égaux 1, on obtient

$$N(u,v) = 2 + \sharp\{\text{Voisins communs entre } u \text{ et } v\} \tag{2.38}$$

et

$$D = 2 + \sharp\{\text{Voisins de } u \text{ et } v\} \tag{2.39}$$

ce qui correspond bien à l'indice défini précédemment.

L'adaptation du calcul du CBS incrémentiel est quant à elle triviale, puisqu'il est seulement nécessaire de multiplier chaque terme ajouté ou retranché par le poids du nœud correspondant.

Cette extension, qui peut être envisageable également pour les graphes dirigés, pose le problème de sa validation. Les problèmes d'étiquetage de graphe n'ont été que très peu étudiés dans le cas des graphes pondérés ou dirigés. La flexibilité de la méthode proposée permet d'envisager des solutions simples à ces problèmes.

En plus de ces extensions possibles, les pistes de poursuite de ce travail sont diverses : une comparaison avec d'autres méthodes d'étiquetage des graphes serait intéressante à mettre en œuvre, notamment pour déterminer quel problème est le plus adapté pour suivre la structure du graphe. Les résultats montrent en effet que le *Cyclic Bandwidth Sum Problem* ne reflète pas forcément la typologie du graphe. Une deuxième piste serait de proposer une étude théorique d'un étiquetage des graphes adapté aux structures de graphe rencontrées dans les réseaux complexes. Il est en effet compliqué de bien définir ce qu'est un étiquetage cohérent. Enfin, il peut être envisagé de considérer non pas des étiquetages unidimensionnels, mais multidimensionnels, adaptés à d'autres types d'application.

Le chapitre suivant introduit la transformation d'un graphe en une collection de signaux, dans laquelle l'heuristique développée est une étape importante afin de pouvoir indexer les signaux de manière adéquate.

CHAPITRE 3

Dualité entre réseaux et signaux

Résumé –

Ce chapitre est dédié à la définition d'une dualité robuste entre les réseaux et les signaux dans un but d'analyse de réseaux. La Section 1 présente brièvement le domaine de traitement du signal, et ses connexions avec les science des réseaux. La dualité entre réseaux et signaux est abordée dans les trois sections suivantes, d'abord en définissant dans la Section 2 la transformation de réseau vers signaux, en appliquant dans la Section 3 la transformation pour des modèles de réseaux présentés dans le Chapitre 2, puis en définissant dans la Section 4 une transformation inverse robuste de manière à obtenir un réseau à partir de signaux, même après perturbation des signaux. La Section 5 applique cette dualité sur une application de débruitage de graphe.

Sommaire

1 Traitement du signal et réseaux

Le traitement du signal est né des domaines de l'électronique et de l'automatique avec pour but initial de répondre aux problématiques liées à l'utilisation des signaux dans des systèmes, notamment de communication, à la fois sur des aspects de transmissions et de conditionnement [83]. Rapidement, la notion de signal a dépassé le seul cadre des signaux électriques, et le traitement du signal s'est retrouvé à l'intersection entre de nombreuses et diverses disciplines : les applications se sont étendues vers des problèmes d'acoustique [172], d'imagerie médicale [25], de mécanique [95] et plus généralement de physique [162], tandis que les aspects théoriques ont fait écho aux domaines de l'analyse, des probabilités ou de la théorie de l'information. L'analyse spectrale, un des outils clés du traitement du signal [94], est un marqueur fort de cette synergie entre ces disciplines : le concept de fréquence au cœur de cette analyse repose sur une réalité physique dans de nombreux systèmes, l'analyse harmonique développée à partir des travaux de Fourier a permis d'ancrer l'analyse spectrale dans un formalisme mathématique rigoureux, tandis que l'informatique a apporté des implémentations algorithmiques efficaces, permettant d'envisager son utilisation dans des applications industrielles. Cette interdisciplinarité a transformé l'analyse spectrale en un outil incontournable aussi bien dans les téléphones portables qu'en physique théorique.

Naturellement, le développement récent de la théorie des réseaux, et plus généralement l'apparition de données massives sous la forme de réseaux, a ouvert la voie vers une extension du traitement du signal à l'étude des réseaux. Si une connexion historique entre traitement du signal et réseaux s'est tout d'abord faite dans le domaine des télécommunications, par exemple sur des problèmes d'allocation de ressources dans des réseaux de communications [108], cette composante s'est renforcée ces dernières années et intègre désormais des problèmes divers, comme en témoigne le récent numéro du journal *IEEE Signal Processing Magazine* paru en mai 2013, et qui présente des travaux portant aussi bien sur l'estimation distribuée [52, 137, 176], l'apprentissage sur réseaux [201] que sur la modélisation bayésienne de systèmes complexes [215]. Si ces travaux répondent souvent à des problèmes concrets, deux avancées à la frontière entre réseaux et signaux visent à généraliser des concepts communs entre théorie des réseaux et traitement du signal, et méritent à ce titre une attention particulière. La première concerne le domaine du traitement du signal sur graphe, dont l'objectif est de transposer les notions et outils du traitement du signal, classiquement défini dans un espace régulier, vers le domaine des graphes. La deuxième consiste à établir des ponts entre graphes et signaux, en proposant des transformations d'un domaine vers l'autre afin de pouvoir utiliser indifféremment les deux théories de façon complémentaire. Comme le titre de ce chapitre le suggère, les travaux présentés par la suite s'intègrent dans ce dernier point. Une brève présentation de ces deux avancées est néanmoins réalisée, après avoir rappelé la transformée de Fourier dans le cas discret.

1.1 Transformée de Fourier discrète de signaux réels

Les travaux présentés dans cette thèse n'utilisent que des outils élémentaires du traitement du signal. Néanmoins, la transformée de Fourier est rappelée par la suite, étant la notation importante qui sera utilisée par la suite pour caractériser la structure d'un réseau.

Définition La transformée de Fourier discrète est une opération qui permet de représenter un signal discret réel en composantes fréquentielles, c'est-à-dire de représenter un signal comme une somme d'oscillations harmoniques. La transformée de Fourier discrète d'un signal $\boldsymbol{x}$ réel composé de n échantillons,

que l'on note $\hat{\boldsymbol{x}}$, est définie par :

$$\hat{x}[f] = \sum_{l=1}^{n} x[l] e^{-2i\pi f \frac{l-1}{n}} \tag{3.1}$$

pour les fréquences positives $f \in \{0, \ldots, \frac{n}{2}\}$. De façon analogue, la transformation de Fourier inverse discrète s'écrit :

$$\hat{x}[l] = \frac{1}{n} \sum_{f=0}^{\frac{n}{2}} \hat{x}[f] e^{2i\pi(l-1)\frac{f}{n}} \tag{3.2}$$

pour $l \in \{1, \ldots, n\}$.

La transformée de Fourier discrète peut également se représenter sous forme matricielle :

$$\hat{\boldsymbol{x}} = \boldsymbol{F}\boldsymbol{x} \tag{3.3}$$

avec $\boldsymbol{F}$, appelée matrice de Fourier dont les termes sont donnés par $f_{lk} = e^{-2i\pi f \frac{l-1}{n}}$. De la même manière, la transformation inverse s'écrit :

$$\boldsymbol{x} = \frac{1}{n} \boldsymbol{F}^* \hat{\boldsymbol{x}} \tag{3.4}$$

avec $\boldsymbol{F}^*$ la matrice adjointe de $\boldsymbol{F}$.

La partie réelle du signal $e^{-2i\pi fl}$ pour $l \in \{1, \ldots, n\}$ est un cosinus d'amplitude 1 et de fréquence f et la partie imaginaire un sinus de mêmes amplitude et fréquence. La décomposition d'un signal $\boldsymbol{x}$ sur cette base permet ainsi de représenter ce signal comme une combinaison linéaire de cosinus et de sinus. Le vecteur $\hat{\boldsymbol{x}}$ est composé de termes complexes $\hat{x}_f$, pour chaque fréquence $f \in \{0, \ldots, \frac{n}{2}\}$. Trois grandeurs vont être particulièrement intéressantes par la suite, calculées à partir de $\hat{\boldsymbol{x}}$:

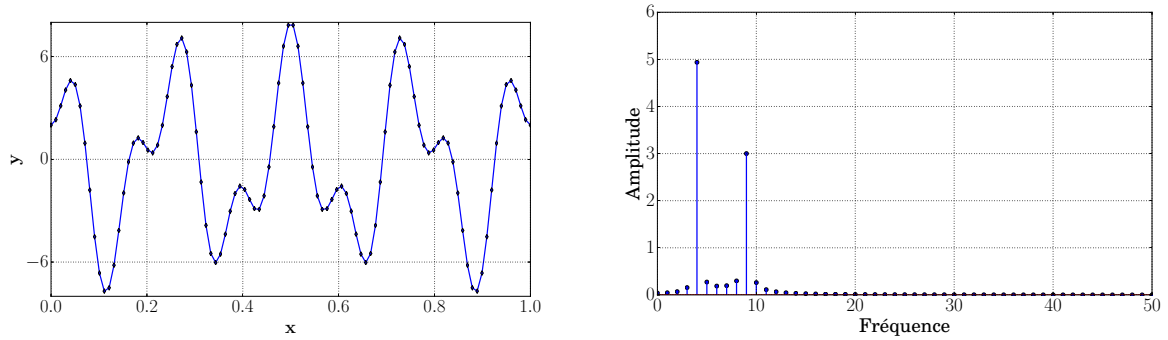
— l'amplitude du spectre $\boldsymbol{m}$, donnée par le module de chaque terme de $\hat{\boldsymbol{x}}$: $m_f = |\hat{x}_f|$;
— l'énergie du spectre $\boldsymbol{z}$, donnée par le module au carré de chaque terme de $\hat{\boldsymbol{x}}$: $z_f = |\hat{x}_f|^2$;
— la phase du spectre $\boldsymbol{\phi}$, donnée par l'argument de chaque terme de $\hat{\boldsymbol{x}}$: $\phi_f = \arg \hat{x}_f$.

L'amplitude et l'énergie donnent des indications sur l'importance d'une fréquence dans le signal par rapport aux autres, alors que la phase permet de caractériser le déphasage entre les oscillations harmoniques. Du fait de l'échantillonnage, des artefacts autour des pics d'amplitude peuvent apparaître dans le spectre [199].

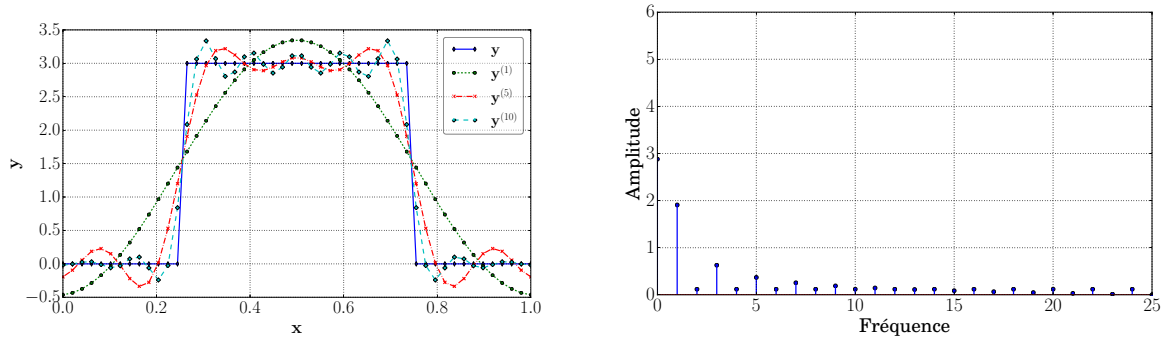
Illustrations La Figure 3.1 propose une illustration de la transformée de Fourier sur deux exemples. La figure de gauche représente un signal $\mathbf{y}$ dans le domaine classique, alors que la figure de droite représente les amplitudes obtenues à partir de $\hat{\mathbf{y}}$, la transformée de Fourier discrète de $\mathbf{y}$.

Le premier exemple illustré sur la Figure 3.1a est un signal avec 200 échantillons, défini comme la somme de deux oscillations harmoniques de fréquences respectives $f_1 = 4$ et $f_2 = 9$ et d'amplitudes respectives $a_1 = 5$ et $a_2 = 3$. Le spectre $\hat{\mathbf{y}}$ courbe présente deux pics d'amplitude, qui correspondent aux deux fréquences f_1 et f_2. Cette représentation permet ainsi de retrouver les oscillations harmoniques, à partir d'un signal pour lequel la visualisation ne permettait pas d'obtenir une décomposition simple.

Le deuxième exemple présenté à la Figure 3.1b montre la transformation de Fourier d'un signal $\mathbf{y}$ constant par morceau avec 200 échantillons. Le signal $y^{(f)}$ est également défini comme étant la transformation inverse de $\hat{\mathbf{y}}$, en ne conservant que les f premières fréquences. La représentation en fréquence de $\mathbf{y}$ montre qu'un signal constant par morceau est constitué d'une somme d'oscillations harmoniques, dont l'amplitude décroît lorsque la fréquence augmente : à partir d'une sinusoïde basse fréquence d'amplitude élevée qui approxime grossièrement le plateau, des sinusoïdes de fréquence plus élevée mais avec une amplitude plus faible sont petit à petit ajoutées, et viennent affiner les parties constantes.



(a) Représentation d'un signal avec 200 échantillons, défini comme la somme de deux oscillations de fréquences $f_1 = 4$ et $f_2 = 9$ et d'amplitudes $a_1 = 5$ et $a_2 = 3$.



(b) Représentation d'un signal $\mathbf{y}$ constant par morceau avec 200 échantillons. Le signal $\mathbf{y}^{(f)}$ représente la transformation inverse de $\hat{\mathbf{y}}$ en ne gardant que les f premières fréquences.

FIGURE 3.1 – Illustration de la transformée de Fourier sur deux exemples. La figure de gauche représente le signal $\mathbf{y}$ dans le domaine classique, alors que la figure de droite représente les amplitudes dans le domaine spectral.

1.2 Traitement du signal sur graphe

Parmi les connexions entre traitement du signal et graphe discutées en introduction, les travaux du traitement du signal sur graphe sont particulièrement intéressants car ils cherchent à généraliser les résultats du traitement du signal classique pour un signal défini sur un graphe. À la différence d'un signal défini sur une topologie régulière, par exemple un signal échantillonné dans le temps ou défini sur une grille, un signal défini sur un graphe consiste à assigner à chacun des nœuds du graphe un scalaire, et à utiliser la topologie du graphe comme support. Il n'y a pas encore de consensus établi sur la bonne façon de procéder à cette généralisation, et notamment sur la matrice adéquate pour représenter la structure du graphe. Les deux approches les plus abouties consistent à définir la matrice de Fourier du graphe comme la matrice des vecteurs propres soit de la matrice d'adjacence [210], soit de la matrice laplacienne [221], définie comme la différence entre la matrice des degrés et la matrice d'adjacence : $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$.

Cette dernière matrice, qui joue un rôle central dans la théorie spectrale des graphes [58], présente la particularité de pouvoir définir la notion de fréquence sur un graphe directement à partir de ses valeurs propres, par analogie avec le cas classique [222]. De nombreux travaux sur la généralisation des concepts du traitement du signal classique ont ainsi éclos à partir de cette définition, que ce soit les opérations usuelles telles que la convolution, la translation, la modulation ou le filtrage, mais également des outils plus complexes comme la définition d'un principe d'incertitude [8], les ondelettes [117], la décomposition en modes empiriques (EMD) [234] ou la stationnarité [109]. Ce passage vers les graphes ne se fait néanmoins pas sans difficulté, et de nombreux points font encore l'objet de vives discussions au sein de la communauté du traitement du signal sur graphe.

Si elles traitent préférentiellement les signaux définis sur le graphe, ces méthodes peuvent néanmoins être utiles pour décrire le graphe lui-même. Ainsi, une méthode de détection de communautés a récemment été développée, basée sur les ondelettes sur graphes [233]. Ces travaux utilisent la propriété multi-résolution des ondelettes pour détecter les communautés de nœuds dans le graphe à plusieurs échelles, par exemple dans le cas de communautés imbriquées les unes dans les autres. La définition rigoureuse d'un point de vue mathématique des ondelettes permet le développement d'outils guidant la recherche de communautés, par exemple afin de sélectionner les échelles pertinentes ou mesurer la stabilité d'une communauté. Cette approche originale souligne les apports du traitement du signal à l'analyse de réseaux complexes, au-delà de l'analyse de signaux.

1.3 Des réseaux vers les signaux, et inversement

Une deuxième approche consiste à changer la représentation des données afin de les étudier sous un angle différent. La transformation de signaux en réseaux a été jusque-là l'approche qui a été la plus étudiée, afin de tirer profit des outils de la théorie des réseaux. Le domaine de l'analyse non-linéaire a fait office de précurseur en la matière : plusieurs méthodes ont été proposées pour l'analyse de séries temporelles non-linéaires [46, 181], la caractérisation de la dynamique d'un système [262, 220, 78] ou encore l'identification de structures invariantes [77]. Ces approches ont notamment permis de caractériser des signaux réels, comme les rythmes cardiaques [46] ou les ondes sismiques [9], à travers les outils de la théorie des réseaux comme la détection de communautés.

L'utilisation du traitement du signal pour l'analyse de réseaux a été beaucoup moins explorée. Les méthodes développées jusque-là sont pour la plupart basées sur des marcheurs aléatoires sur le graphe : Weng et al. [248] construisent des séries temporelles à partir du graphe en récupérant les degrés des nœuds visités par le marcheur aléatoire au fur et à mesure de la progression de celui-ci dans le graphe. Plusieurs signaux sont ensuite considérés pour un même graphe, correspondant à plusieurs marches aléatoires différentes, et des corrélations entre signaux issus de différents graphes sont établies et mises en relation avec la propriété d'invariance d'échelle des graphes. À partir d'une approche relativement proche, Campanharo et al. [46] ont également proposé une méthode pour transformer un graphe à base de marches aléatoires, définissant une transformation inverse pour retrouver des séries temporelles transformées en réseau. La valeur assignée au nœud n'est plus le degré, mais une valeur du signal original obtenue après échantillonnage de l'intervalle des valeurs du signal. Girault et al. [110] ont étendu cette approche dans le cas où le graphe est l'objet d'intérêt, en cherchant les valeurs associées aux nœuds les plus appropriées, de façon à ce que le signal obtenu soit lisse, en utilisant un algorithme d'apprentissage semi-supervisé.

Une autre approche a été proposée d'abord par Haraguchi et al. [119] puis par Shimada et al. [219], basée sur le positionnement multidimensionnel [34] : les nœuds du graphe sont représentés comme un ensemble de points dans un espace euclidien, dans lequel les relations entre les nœuds sont représentées par les distances entre les points. Cette méthode a plusieurs avantages qui sont la motivation des travaux présentés dans ce chapitre : tout d'abord, cette transformation est, contrairement aux marcheurs aléatoires, complètement déterministe, c'est-à-dire qu'à un réseau correspond une unique représentation dans le domaine des signaux. Un deuxième point intéressant est que la représentation est équivalente, puisqu'il est possible de retrouver le graphe complet simplement à partir de la collection de signaux obtenue. Ainsi, les signaux contiennent l'information complète sur la structure. L'extension de cette méthode est l'objet des prochaines sections de ce chapitre.

2 Transformation de graphes en signaux

2.1 Positionnement multidimensionnel classique (CMDS)

Principe Le positionnement multidimensionnel (*MultiDimensional Scaling* (MDS)) est un ensemble de techniques utilisées pour représenter des similarités (ou des dissimilarités) entre des paires d'objets comme des distances entre des points dans un espace multidimensionnel. Les origines du MDS se trouvent dans le domaine de la psychométrie, qui étudie l'ensemble des techniques de mesures pratiquées en psychologie, afin d'étudier les jugements d'un groupe de personnes sur les similarités entre des objets. Pour cette étude, Torgerson [230] a proposé en 1952 une première version du MDS, qui est devenue au fil des ans une technique classique très utilisée dans un grand nombre de domaines, notamment pour la réduction de la dimensionnalité de données afin de faciliter leur visualisation.

Les modèles MDS consistent à proposer une application entre des données représentant des similarités (parfois appelées proximités) ou des dissimilarités et une configuration de points dans un espace euclidien. La configuration de points est construite de telle sorte qu'elle reflète les relations entre les objets. Par la suite, nous considérerons les données initiales comme des dissimilarités, c'est-à-dire que plus la valeur est grande entre deux objets, plus ces objets sont considérés comme distants.

Dans un cadre général, les dissimilarités se représentent sous la forme d'une matrice $\boldsymbol{\Delta} = (\delta_{ij})_{i,j=\{1,\dots,n\}}$, telle que pour $i, j = \{1, \dots, n\}$, δ_{ij} donne la dissimilarité entre les objets i et j. Ces dissimilarités sont représentées comme des distances euclidiennes à l'aide d'une fonction de représentation f. Enfin, une configuration de points est représentée par une matrice $\boldsymbol{X}$ de dimension $n \times q$, telle que chaque ligne correspond à un des n objets, et chaque colonne donne les coordonnées de chaque objet dans un espace euclidien de dimension q. La matrice des distances euclidiennes entre chaque paire de points est notée $\boldsymbol{D}(\boldsymbol{X})$. En utilisant ces notations, le MDS consiste à trouver, pour une matrice $\boldsymbol{\Delta}$ donnée, une configuration de points $\boldsymbol{X}$ telle que la distance euclidienne entre chaque paire de points (i, j) soit égale à $f(\delta_{ij})$. Dans le cas où il n'y a pas de solution possible, l'objectif est de trouver la solution la plus proche telle que les distances satisfont au mieux les dissimilarités. L'intérêt pour la réduction de dimensionnalité est de choisir q faible par rapport à n, typiquement égal à 2 ou à 3 pour des problèmes de visualisation.

Solution dans le cas général Il n'existe pas dans le cas général de solutions analytiques pour la résolution de ce problème. Un algorithme d'optimisation a néanmoins été proposé afin d'obtenir une solution approximative, qui consiste à minimiser l'erreur quadratique entre $\boldsymbol{\Delta}$ et $f(\boldsymbol{D}(\boldsymbol{X}))$. Les détails de son implémentation sont données dans [34].

Positionnement multidimensionnel classique Il existe une très grande diversité de fonctions de représentation, adaptées aux différents types de données qui existent pour représenter les dissimilarités [34]. Dans la suite de ce travail, nous nous plaçons dans le cas dit classique, pour lequel la fonction de représentation f est donnée par $f(\delta_{ij}) = \delta_{ij}$, c'est-à-dire que la matrice des dissimilarités $\boldsymbol{\Delta}$ est supposée être une matrice de distances euclidiennes. Le positionnement multidimensionnel classique (CMDS) possède l'avantage qu'il est possible d'obtenir la solution $\boldsymbol{X}$ à partir de la matrice $\boldsymbol{\Delta}$ analytiquement, sans passer par un algorithme d'optimisation.

Le processus se décompose en plusieurs étapes : à partir d'une matrice de distance $\boldsymbol{\Delta} = (\delta_{ij})_{i,j=1,\dots,n}$, un double centrage de la matrice $\boldsymbol{\Delta}$ élevée au carré est d'abord réalisé afin d'obtenir une matrice $\boldsymbol{B}$:

$$\boldsymbol{B} = -\frac{1}{2}\boldsymbol{J}\boldsymbol{\Delta}^{(2)}\boldsymbol{J} \tag{3.5}$$

avec $\boldsymbol{\Delta}^{(2)} = \boldsymbol{\Delta} \circ \boldsymbol{\Delta}$ où $\circ$ désigne la multiplication matricielle terme à terme, et $\boldsymbol{J} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ où $\boldsymbol{I}_n$ désigne la matrice identité de taille n et $\mathbf{1}_n\mathbf{1}_n^T$ est une matrice de dimension $n \times n$ remplie de 1. La

solution du CMDS est alors obtenue en diagonalisant la matrice $\boldsymbol{B}$:

$$\boldsymbol{X} = \boldsymbol{Q}_+ \boldsymbol{\Lambda}_+^{\frac{1}{2}} \tag{3.6}$$

avec $\boldsymbol{\Lambda}_+$ une matrice diagonale dont les termes sont les valeurs propres de la matrice $\boldsymbol{B}$ triées dans l'ordre décroissant et $\boldsymbol{Q}_+$ est la matrice des vecteurs propres correspondants. Il est important de noter que la solution n'est pas unique, puisque toute rotation, réflexion ou translation des points dans l'espace euclidien préservent les distances. Il est possible de montrer que 0 toujours une valeur propre de $\boldsymbol{B}$, par conséquent la dimension q est au maximum égale à $n - 1$.

2.2 Méthode de transformation d'un graphe en une collection de signaux

En utilisant le positionnement multidimensionnel classique, Shimada et al. [219] ont proposé une méthode pour transformer un graphe à n nœuds en signaux composés de n points et indexés par les nœuds du graphe. La méthode consiste à définir une matrice de distances entre les nœuds du graphe $\boldsymbol{\Delta}$, qui transcrit la présence ou l'absence d'un lien. Ils proposent ainsi la définition suivante :

$$\delta_{ij} = \begin{cases} 0 & \text{si } i = j \\ 1 & \text{si } a_{ij} = 1 \text{ et } i \neq j \\ w > 1 & \text{si } a_{ij} = 0 \text{ et } i \neq j \end{cases} \tag{3.7}$$

ou sous forme matricielle :

$$\boldsymbol{\Delta} = \boldsymbol{A} + w(\mathbf{1}_n \mathbf{1}_n^T - \boldsymbol{I}_n - \boldsymbol{A}) \tag{3.8}$$

avec w un poids arbitraire strictement supérieur à 1.

Une fois le CMDS appliqué, la configuration de points $\boldsymbol{X}$ obtenue revient à représenter chaque nœud du graphe comme un point dans un espace euclidien de dimension $n - 1$. Les signaux résultants que l'on va considérer par la suite seront les colonnes de la matrice $\boldsymbol{X}$, également appelées sans distinction composantes. Le j^{e} signal (ou j^{e} composante) est noté $\boldsymbol{X}^{(j)}$.

Cette définition se concentre sur la présence (dénotée par une distance égale à 1) et l'absence (dénotée par une distance égale à w) d'un lien entre deux nœuds. Ainsi, la distance entre deux nœuds dans le graphe, souvent définie comme la longueur du plus court chemin entre les deux nœuds, n'a pas d'impact direct sur la matrice $\boldsymbol{\Delta}$: deux nœuds déconnectés vont avoir une distance égale à w, qu'ils soient proches ou éloignés (au sens de la longueur du plus court chemin) dans le graphe. Néanmoins, l'information sur la proximité dans le graphe reste présente, et est rendue explicite lorsque la dimension de l'espace euclidien diminue : comme il n'est pas possible de trouver une configuration de points $\boldsymbol{X}$ telle que les distances soient bien préservées, seules les distances entre les nœuds distants dans le graphe vont être conservées, car elles représentent au mieux la structure globale du graphe. L'intérêt va ainsi être de pouvoir regarder composante par composante, comment les distances sont approximées, et en déduire des informations sur la structure du réseau.

2.3 Choix du paramètre w

Pour une matrice de distance $\boldsymbol{\Delta}$ donnée, il n'existe pas nécessairement une configuration de points $\boldsymbol{X}$ telle que les distances entre les points soient égales aux distances définies dans la matrice $\boldsymbol{\Delta}$, et ceci même si la dimensionnalité est maximale. Dans le cas où $\boldsymbol{\Delta}$ est obtenue en utilisant l'équation 3.8, la structure du graphe ainsi que la valeur du paramètre w vont avoir une influence sur l'existence d'une solution exacte. Cette influence a été étudiée dans [119] et [219] de manière empirique : dans [119], les auteurs comparent pour différentes valeurs de w une mesure de qualité Q sur plusieurs graphes obtenus à partir du modèle de Watts-Strogatz, pour différents niveau de probabilité, ainsi que pour deux réseaux issus

de données réelles. Ils en concluent que la valeur w doit être choisie entre 1 (exclu) et 1.01. La même approche est poursuivie dans [219], mais restreinte cette fois au modèle de Watts-Strogatz, également pour différentes valeurs de probabilité. Ils en concluent que pour $n = 400$, w doit être compris entre 1 (exclu) et 1.14, et que la borne supérieure de w dépend de la valeur de n et doit être le plus proche possible de 1, sans néanmoins fournir d'argument pour soutenir cette affirmation.

Une borne supérieure pour la valeur de w peut être de façon moins empirique, en se basant sur l'étude des valeurs propres de la matrice $\boldsymbol{\Delta}$: comme mentionné dans le travail de Gower [114], la matrice $\boldsymbol{\Delta}$ est exactement retrouvée à partir de $\boldsymbol{X}$ si et seulement si $\boldsymbol{B}$ est définie positive, c'est-à-dire :

$$\langle \boldsymbol{z}, \boldsymbol{B}\boldsymbol{z} \rangle \geq 0 \text{ pour tout vecteur } \boldsymbol{z} \in \mathbb{R}^n \tag{3.9}$$

ou de manière équivalente, si et seulement si $\boldsymbol{\Delta}^{(2)}$ est définie conditionnellement négative :

$$\langle \boldsymbol{z}, \boldsymbol{\Delta}^{(2)}\boldsymbol{z} \rangle \leq 0 \quad \forall \boldsymbol{z} \in \mathbb{R}^n \text{ tel que } \sum_{i=1}^{n} z_i = 0 \tag{3.10}$$

À partir de la définition de $\boldsymbol{\Delta}$ dans l'équation 3.7, on a :

$$\begin{aligned} \langle \boldsymbol{z}, \boldsymbol{\Delta}^{(2)}\boldsymbol{z} \rangle &= \langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle + w^2(\langle \boldsymbol{z}, \mathbf{1}_n\mathbf{1}_n^T\boldsymbol{z} \rangle - \langle \boldsymbol{z}, \boldsymbol{I}_n\boldsymbol{z} \rangle - \langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle) \\ &= \langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle - w^2(\langle \boldsymbol{z}, \boldsymbol{z} \rangle + \langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle) \end{aligned} \tag{3.11}$$

Deux cas peuvent être distingués :

1. Si $\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle > -\langle \boldsymbol{z}, \boldsymbol{z} \rangle$, alors

$$w^2 \geq \frac{\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle}{\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle + \langle \boldsymbol{z}, \boldsymbol{z} \rangle} \tag{3.12}$$

   car $w > 1$.

2. Si $\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle < -\langle \boldsymbol{z}, \boldsymbol{z} \rangle$, alors

$$w^2 \leq \frac{\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle}{\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle + \langle \boldsymbol{z}, \boldsymbol{z} \rangle} \tag{3.13}$$

La borne supérieure dépend de la matrice d'adjacence $\boldsymbol{A}$, c'est-à-dire, de la structure du graphe. On se place dans le cas où $\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle$ est minimal. Pour des raisons pratiques, on considère que le nombre de nœuds est pair. $\boldsymbol{A}$ est défini comme la matrice d'adjacence du graphe avec n nœuds, avec $a_{ij} = 1$ si et seulement si i et j appartiennent au même sous-ensemble parmi $\{1, \ldots, \frac{n}{2}\}$ et $\{\frac{n}{2} + 1, \ldots, n\}$. $\boldsymbol{A}$ est ainsi une matrice avec 4 blocs, avec les entrées des blocs inférieur gauche et supérieur droit égales à 1. Quant au vecteur z, il est égal à -1 pour la première moitié du vecteur, et à 1 pour la deuxième moitié : $z = [-1, -1, \ldots, 1, 1]$. $\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} \rangle$ est ainsi égal à $-\frac{n^2}{2}$ et $\langle \boldsymbol{z}, \boldsymbol{z} \rangle = n$. On obtient alors une approximation pour la borne supérieure de w :

$$w \leq \sqrt{\frac{n}{n-1}} \tag{3.14}$$

Ce résultat est en accord avec les résultats partiels décrits dans [119] et [219] : w doit être le plus proche possible de 1. En pratique, le réglage de w est plus lâche étant donné que la borne supérieure tient pour des graphes avec une structure très particulière. Le choix de $w = \sqrt{\frac{n}{n-1}}$ est néanmoins conservé afin de s'assurer que $\boldsymbol{\Delta}$ reste une matrice euclidienne.

2.4 Comparaison avec d'autres techniques

L'originalité de la méthode proposée par Shimada et al. [219] tient dans la matrice de distance utilisée. Des matrices de distances alternatives ont cependant été proposées : une mesure naturelle de la distance entre les nœuds est la longueur du plus court chemin entre ces deux nœuds. La méthode *Isomap*, développée par Tenenbaum et al. [227], applique le positionnement multidimensionnel sur la matrice des distances calculées à l'aide de la longueur des plus courts chemins, de façon à projeter le graphe dans un espace euclidien. Cette phase n'intervient cependant que dans un deuxième temps, puisque la méthode a été développée pour étudier des distributions de points définies sur des variétés particulières, et non pour l'analyse de graphes.

La méthode *Laplacian eigenmaps* proposée par Belkin et al. [23] est quant à elle une alternative au CMDS pour transformer un graphe en points dans un espace euclidien. Cette méthode se base sur la matrice laplacienne du graphe $\boldsymbol{L}$, définie par $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, avec $\boldsymbol{D}$ la matrice dont les termes diagonaux contiennent les degrés des nœuds. La collection de signaux est obtenue à partir des vecteurs propres du laplacien.

Si les méthodes présentent des similitudes, il existe néanmoins des différences dans les représentations obtenues [207]. La comparaison de ces différentes méthodes pourrait ainsi apporter une meilleure compréhension des relations entre graphes et signaux, mais ne sera pas traité par la suite.

2.5 Analyse spectrale

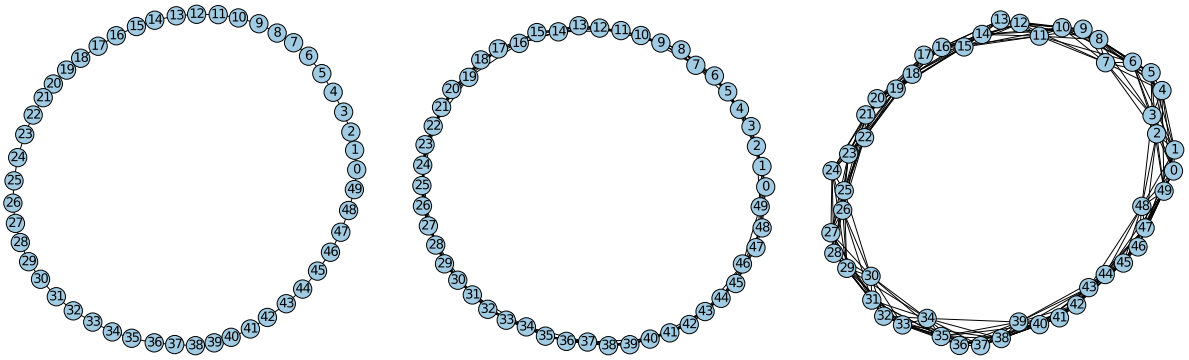
Afin de caractériser la collection de signaux obtenue, une analyse spectrale est réalisée en utilisant les méthodes standards issues du traitement du signal. Soit une collection $\boldsymbol{X}$ composée de C composantes, indexées par les n nœuds du réseau. Le spectre $\boldsymbol{S}$ donne les coefficients de Fourier complexes, obtenus en appliquant la transformée de Fourier discrète sur chacune des C composantes. En notant $\boldsymbol{S}^{(j)}$ la transformée de Fourier discrète de la j^{e} composante $\boldsymbol{X}^{(j)}$, on obtient :

$$\boldsymbol{S}^{(j)} = \boldsymbol{F}\boldsymbol{X}^{(j)} \tag{3.15}$$

avec $\boldsymbol{F}$ la matrice de Fourier. Chaque terme $\boldsymbol{S}^{(j)}$ donne le coefficient associé à chaque fréquence $f \in \{0, \ldots, \frac{n}{2}\}$. À partir de $\boldsymbol{S}$, la matrice des amplitudes $\boldsymbol{M}$, des énergies $\boldsymbol{Z}$ et des phases $\boldsymbol{\Phi}$ sont calculées suivant les définitions de la Section 1.1.

Comme les signaux sont indexés par les nœuds du graphe, les fréquences s'apparentent à des fréquences définies sur le graphe, même si cette notion n'est pas aussi bien définie que dans les travaux de traitement du signal sur graphe [222]. La comparaison a néanmoins du sens, par exemple dans le cas d'un signal unique : si le signal est majoritairement composé de basse fréquence, cela signifie que la structure du graphe est régulière le long des nœuds du graphe, pris dans l'ordre de l'indexation. Au contraire, un signal avec de fortes variations témoigne de l'irrégularité de la structure du graphe. En caractérisant pour chaque composante la régularité de la structure du graphe, on en déduit des informations sur la structure globale du graphe. L'indexation joue néanmoins un rôle crucial dans l'obtention de signaux réguliers, et pour garantir que l'analyse spectrale va être capable de capturer la régularité du graphe, il faut s'assurer qu'elle soit cohérente avec la structure du graphe. Les travaux présentés dans le Chapitre 2 viennent ainsi s'intégrer dans la transformation, en permettant d'obtenir un étiquetage des nœuds, et donc une indexation des signaux, cohérents avec la topologie du graphe.

L'analyse spectrale est exploitée par la suite afin d'établir des connexions entre des motifs fréquentiels, visibles dans un plan composante-fréquence, et la topologie du graphe correspond à la collection de signaux. Cette approche est rendue explicite à travers les illustrations de la Section 3.



(a) $k = 2$ (b) $k = 6$ (c) $k = 10$

FIGURE 3.2 – Visualisation d'un graphe k-régulier en anneau, pour différentes valeurs de k. Plus la valeur de k est grande, plus la structure en anneau est bien définie.

3 Résultats sur des modèles de graphes

3.1 Graphe k-régulier en anneau

Le premier type de graphe étudié est le graphe k-régulier en anneau, qui est un graphe classique du fait des propriétés liées à sa régularité. La Figure 3.2 affiche trois représentations de ce type de graphe, pour trois valeurs de k différentes. Plus la valeur de k est grande, plus la structure en anneau est bien définie.

Comme discuté par Shimada et al. [219], l'utilisation de la théorie des matrices circulantes [115] permet de façon immédiate de connaître les valeurs propres et vecteurs propres de la matrice de distance $\boldsymbol{\Delta}$. Cependant, ils ne proposent pas de formule explicite pour calculer les signaux obtenus en fonction de n et k, ce qui est réalisé par la suite.

D'après la théorie des matrices circulantes, toute matrice circulante $\boldsymbol{C}$ de dimension $n \times n$ a ses valeurs propres $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]$ données $\forall q \in \{0, \ldots, n-1\}$ par :

$$\lambda_q = \sum_{j=0}^{n-1} c_j \zeta^{kj} \tag{3.16}$$

où $\boldsymbol{c}$ est le vecteur circulant de $\boldsymbol{C}$ et $\zeta = e^{\frac{2i\pi}{n}}$ est la racine n^{e} de l'unité. Quant aux vecteurs propres, le vecteur propre $\boldsymbol{v}_q$ associé à la valeur propre $q, \forall q \in \{0, n-1\}$ est donné par :

$$\boldsymbol{v}_q = \sqrt{n}[1, \zeta^q, \zeta^{2q}, \ldots, \zeta^{(n-1)q}] \tag{3.17}$$

correspondant aux colonnes de la matrice de Fourier, notées $\boldsymbol{F}^{(q)}$. Ces valeurs et vecteurs propres sont complexes, à savoir $\bar{\lambda}_q = \lambda_{n-q}$ et $\bar{\boldsymbol{v}}_q = \boldsymbol{v}_{n-q}$ pour $q \neq 0$.

Nous nous plaçons dans le cas de matrices symétriques. Les valeurs propres sont réelles et doubles : $\lambda_q = \lambda_{n-q}$ pour $q > 0$; les vecteurs propres correspondants sont les parties réelles et imaginaires de $\boldsymbol{v}_q$, normalisées par $\sqrt{2}$ de façon à obtenir une matrice orthonormale :

$$\boldsymbol{v}_q = \sqrt{2}\Re(\boldsymbol{F}^{(q)}) = \sqrt{2}\cos(\frac{2\pi q}{n}) \tag{3.18}$$

et

$$\boldsymbol{v}_{n-q} = \sqrt{2}\Im(\boldsymbol{F}^{(q)}) = \sqrt{2}\sin(\frac{2\pi q}{n}) \tag{3.19}$$

correspondant à des sinusoïdes. Si n est pair, $\lambda_{\frac{n}{2}}$ est valeur propre unique, et le vecteur propre correspondant n'est pas normalisé par $\sqrt{2}$.

Ces résultats sont maintenant appliqués dans le cas de la transformation de graphe en signaux : le vecteur circulant $\boldsymbol{\delta}$ de la matrice $\boldsymbol{\Delta}$ a trois valeurs possibles :

$$\delta_i = \begin{cases} \frac{\alpha}{2n} & \text{si } i = 0 \\ -\frac{\alpha}{2}(1 - \frac{\alpha}{n}) & \text{si } i \in \{1, \dots, \frac{k}{2}\} \cup \{n - \frac{k}{2}, \dots, n-1\} \\ -\frac{\alpha}{2}(w^2 - \frac{\alpha}{n}) & \text{si } i \in \{\frac{k}{2}+1, \dots, n - \frac{k}{2} - 1\} \end{cases} \tag{3.20}$$

Le vecteur circulant de la matrice $\boldsymbol{B}$, dont les entrées sont définies par $b_i = -\frac{1}{2}[\delta_i^2 - \frac{\alpha}{n}]$, avec $\alpha = k + (n - 1 - k)w^2$, a également trois valeurs possibles :

$$b_i = \begin{cases} \frac{\alpha}{2n} & \text{si } i = 0 \\ -\frac{\alpha}{2}(1 - \frac{\alpha}{n}) & \text{si } i \in \{1, \dots, \frac{k}{2}\} \cup \{n - \frac{k}{2}, \dots, n-1\} \\ -\frac{\alpha}{2}(w^2 - \frac{\alpha}{n}) & \text{si } i \in \{\frac{k}{2}+1, \dots, n - \frac{k}{2} - 1\} \end{cases} \tag{3.21}$$

Le calcul des valeurs propres de $\boldsymbol{B}$ donne :

$$\lambda_q = \frac{\alpha}{2n} \sum_{j=0}^{n-1} \zeta^{jq} - \frac{1}{2} \left(\sum_{j=1}^{\frac{k}{2}} \zeta^{jq} + \sum_{j=n-\frac{k}{2}}^{n-1} \zeta^{jq} + w^2 \sum_{j=\frac{k}{2}+1}^{n-\frac{k}{2}-1} \zeta^{jq} \right) \tag{3.22}$$

À partir de ces valeurs propres, les signaux obtenus sont similaires à ceux obtenus en utilisant les méthodes standards de diagonalisation de matrices, à des rotations, translations et symétries près.

Lorsque les valeurs propres sont ordonnées suivant la valeur de q, les vecteurs propres sont considérés dans un ordre décroissant des fréquences. Les composantes sont cependant triées en fonction de l'énergie des valeurs propres λ_k lorsque le positionnement multidimensionnel est appliqué, ce qui correspond à trier en fonction de la valeur de q seulement dans le cas où $k = 2$: les signaux résultants sont ainsi les oscillations harmoniques dont la fréquence augmente lorsque l'énergie des composantes décroît. Lorsque k est plus élevé, les composantes ne sont plus triées en fonction de la fréquence. Étant donné que chaque signal consiste en une oscillation harmonique, et que les signaux sont triés en fonction des valeurs propres, le motif fréquentiel correspondant à un graphe k-régulier consiste, pour chaque signal, en un seul pic d'amplitude, de valeur égale à la racine carrée de la valeur propre, et associée à la fréquence correspondante.

Afin d'expliciter l'influence du paramètre k, la Figure 3.3 affiche les valeurs propres strictement positives associées à chaque composante après transformation d'un graphe k-régulier avec 200 nœuds, pour $k \in \{2, 6, 10, 14, 18\}$. Pour des raisons de lisibilité, seule une valeur propre par paire est affichée. Lorsque $k = 2$, les valeurs propres sont strictement décroissantes par rapport aux fréquences, et le tri se fait dans l'ordre des fréquences. Lorsque k est plus grand que 2, des oscillations apparaissent, d'abord douces, puis de plus en plus fortes lorsque k augmente. Ces oscillations impactent la collection de signaux obtenue : si les signaux restent des oscillations harmoniques, leur position dans la collection varie.

La Figure 3.4 affiche la collection de signaux obtenue après transformation d'un graphe k-régulier en anneau avec 200 nœuds, pour différentes valeurs de k. La figure de gauche affiche les composantes 1, 2, 50 et 100, alors que la figure de droite affiche les amplitudes pour chaque fréquence de chaque composante. Ces résultats sont conformes à l'étude théorique proposée plus haut : d'une part, les signaux obtenus sont des paires d'oscillations harmoniques avec une même fréquence et une différence de phase de $\frac{\pi}{2}$, comme le montrent les composantes 1 et 2. D'autre part, pour les premières composantes, le paramètre k n'a pas d'influence sur la fréquence, à la différence des composantes de plus basse énergie,

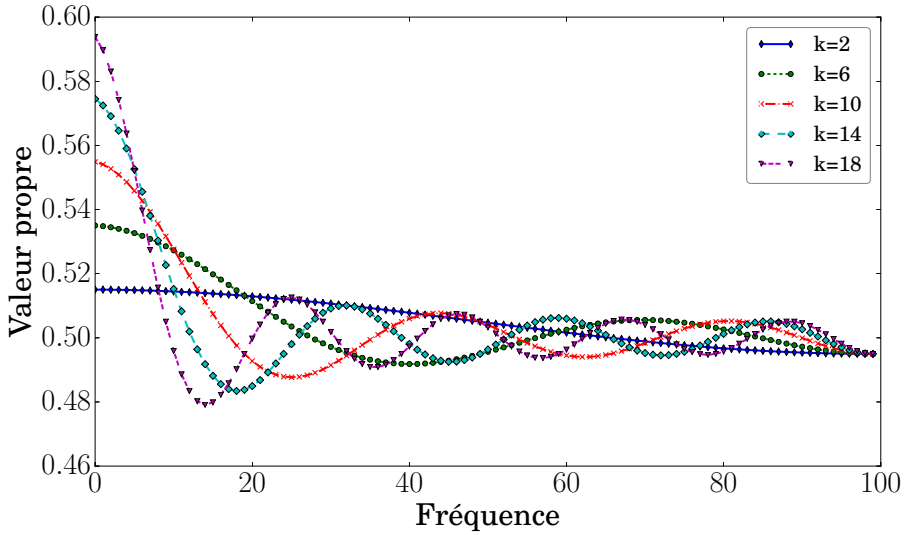


FIGURE 3.3 – Valeurs propres strictement positives associées à chaque composante après transformation d'un graphe k-régulier avec 200 nœuds, pour $k \in \{2, 6, 10, 14, 18\}$. Pour des raisons de lisibilité, seule une valeur propre par paire est affichée.

pour lesquelle le tri des valeurs propres par ordre décroissant modifie l'ordre dans lequel les oscillations harmoniques sont considérées.

Les motifs fréquentiels correspondent bien à ceux attendus, et en note les similitudes entre la répartition des pics d'amplitude dans le plan composante-fréquence et les distribution des valeurs propres en fonction de la fréquence affichées à la Figure 3.3.

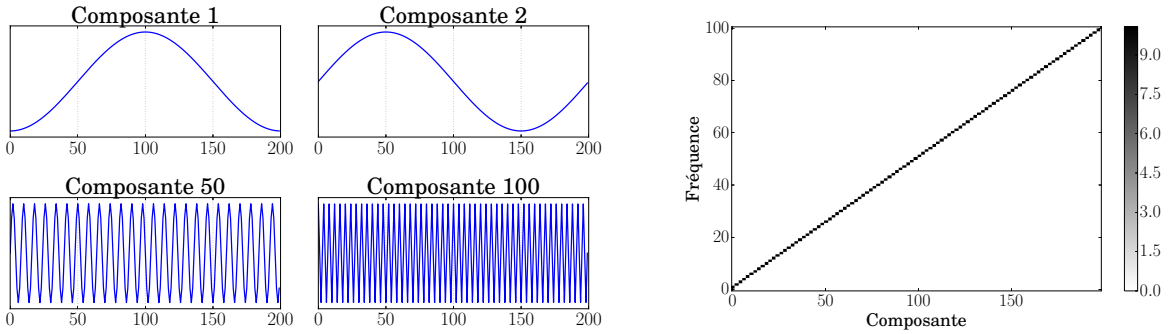
3.2 Modèle d'Erdös-Rényi

Le deuxième modèle étudié est un modèle aléatoire de type Erdös-Rényi, décrit dans la Section 2.4.2 du Chapitre 2. La Figure 3.5 illustre pour différentes valeurs de p la structure aléatoire de ces graphes, dans lesquels le paramètre p influe sur la densité de liens [1].

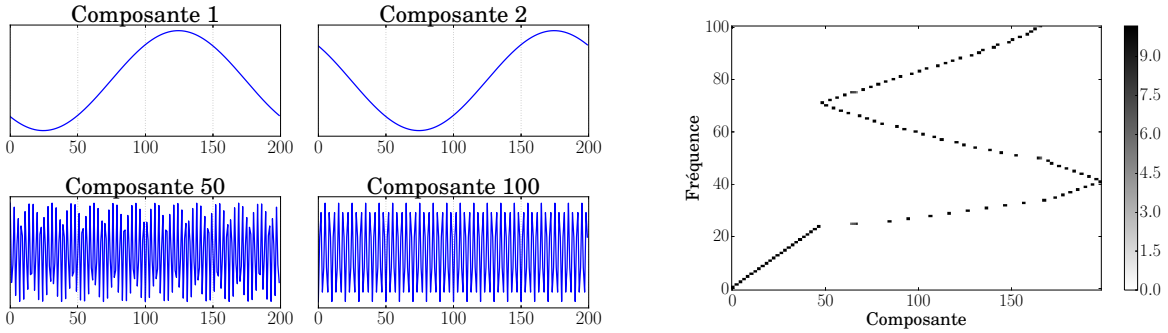
Il est compliqué de connaître précisément la distribution des valeurs propres et des vecteurs propres de la matrice $\boldsymbol{B}$ sans une étude rigoureuse basée sur la théorie des matrices aléatoires. Néanmoins, les résultats obtenus sur des matrices similaires [69] suggèrent d'une part que les vecteurs propres sont des vecteurs aléatoires dans $\mathbb{R}^n$, qui peuvent naturellement se conjecturer comme étant distribués comme des vecteurs gaussiens indépendants et identiquement distribués. D'autre part, les distributions des valeurs propres de la matrice d'adjacence [238] ainsi que des résultats empiriques montrent que la valeur propre minimale λ_{min} est égale à 0, et que les autres valeurs propres suivent une loi du demi-cercle dans l'intervalle $\left[-\sqrt{\frac{p(1-p)}{n}}, \sqrt{\frac{p(1-p)}{n}}\right]$ translatée de 0.5, donnant ainsi des intuitions sur les énergies des signaux. Le motif fréquentiel attendu est ainsi celui d'un bruit blanc, c'est-à-dire une répartition de l'énergie sur toutes les fréquences, avec néanmoins une énergie légèrement plus concentrée vers les premières composantes, du fait du tri des signaux, ainsi que vers les basses fréquences, à cause de l'indexation des nœuds qui n'est pas aléatoire.

La Figure 3.6 illustre ce résultat en affichant l'histogramme des valeurs propres strictement positives associées à chaque composante après transformation d'un graphe obtenu à partir d'un modèle de type Erdös-Rényi, pour différentes valeurs de n et p. La courbe verte représente la densité de la loi du demi-cercle associée.

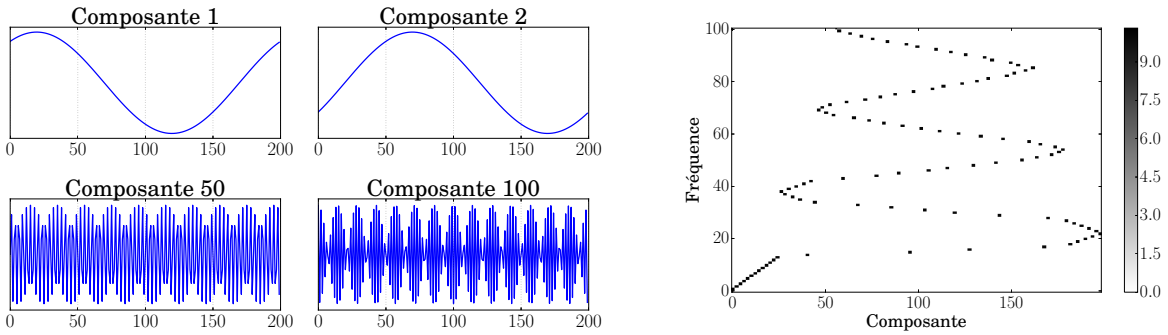
1. Seuls les graphes connectés seront considérés par la suite.



(a) $k = 2$

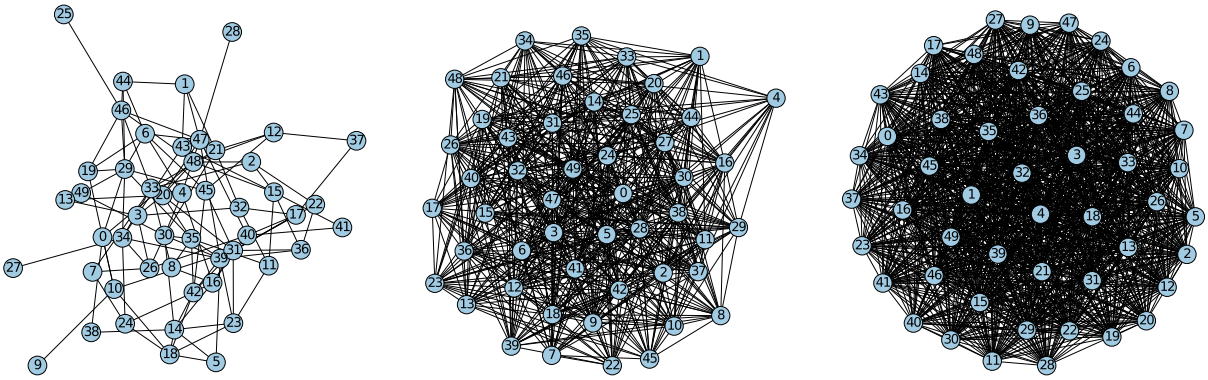


(b) $k = 6$



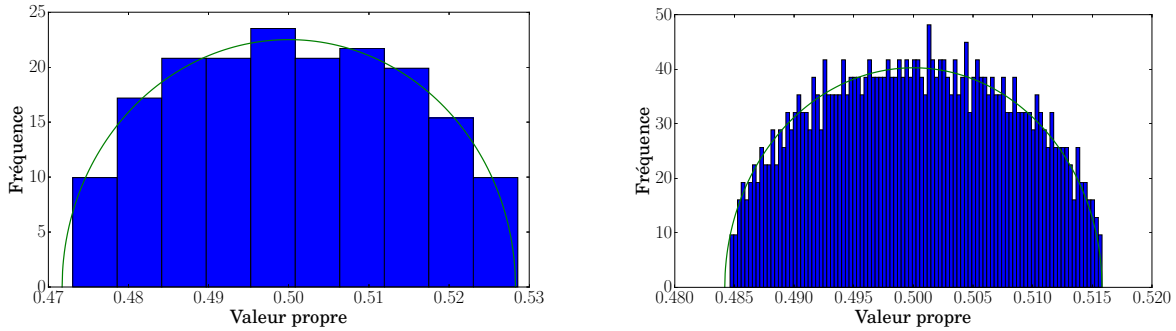
(c) $k = 12$

FIGURE 3.4 – Collection de signaux obtenue après transformation d'un graphe k-régulier en anneau avec 200 nœuds, pour différentes valeurs de k. La figure de gauche affiche les composantes 1, 2, 50 et 100, alors que la figure de droite affiche les amplitudes pour chaque fréquence de chaque composante.



(a) $k = 2$ (b) $k = 6$ (c) $k = 10$

FIGURE 3.5 – Visualisation d'un graphe aléatoire de type Erdös-Rényi, pour différentes valeurs de p. Plus la valeur de p est grande, plus la densité de liens est importante.



(a) $n = 200$ et $p = 0.8$ (b) $n = 1000$ et $p = 0.5$

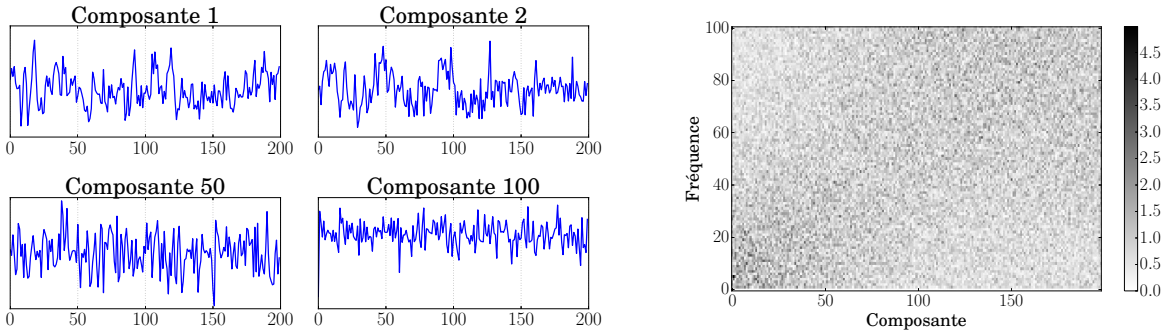
FIGURE 3.6 – Histogramme des valeurs propres strictement positives associées à chaque composante après transformation d'un graphe obtenu à partir d'un modèle de type Erdös-Rényi, pour différentes valeurs de n et p. La courbe verte représente la densité d'une loi du demi-cercle définie dans l'intervalle $[-\sqrt{\frac{p(1-p)}{n}}, \sqrt{\frac{p(1-p)}{n}}]$ et translatée de 0.5.

La Figure 3.7 affiche la collection de signaux obtenue après transformation de plusieurs graphes aléatoires de type Erdös-Rényi avec 200 nœuds pour différentes valeurs de p, de manière similaire à la Figure 3.4. Les signaux obtenus ressemblent à des vecteurs aléatoires, tandis que que l'énergie est répartie dans le plan composante-fréquence, même si l'on note une intensité légèrement supérieure dans le coin inférieur gauche, correspondant aux basses fréquences pour les premiers signaux. Cependant, la visualisation du motif fréquentiel ne laisse pas apparaître de structure particulière, ce qui traduit l'absence de structure également dans le graphe. L'influence de p est difficilement perceptible sur dans les motifs fréquentiels, si ce n'est sur la valeur de l'amplitude maximale, qui semble plus élevée lorsque p est faible.

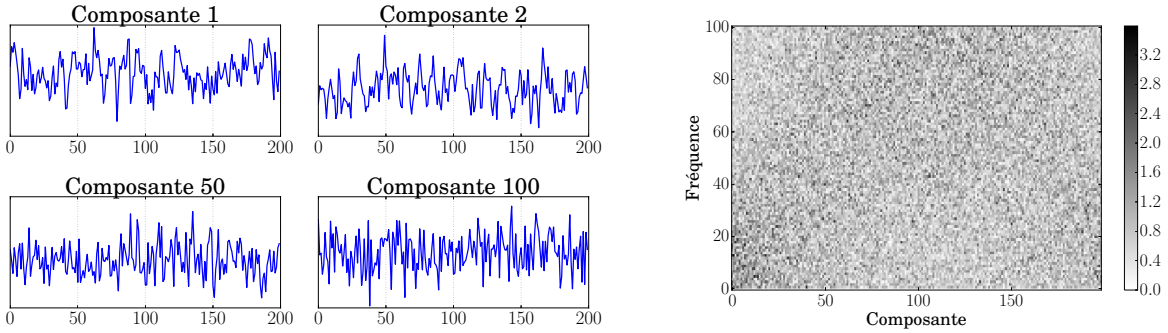
3.3 Modèle de Watts-Strogatz

Le modèle de Watts-Strogatz se construit à partir d'un graphe k-régulier en anneau, dans lequel des liens sont aléatoirement ajoutés pour créer des raccourcis dans le graphe. La Figure 3.8 illustre pour différentes valeurs de k et de p la structure en anneau de ces graphes, plus ou moins perturbée suivant la valeur de p.

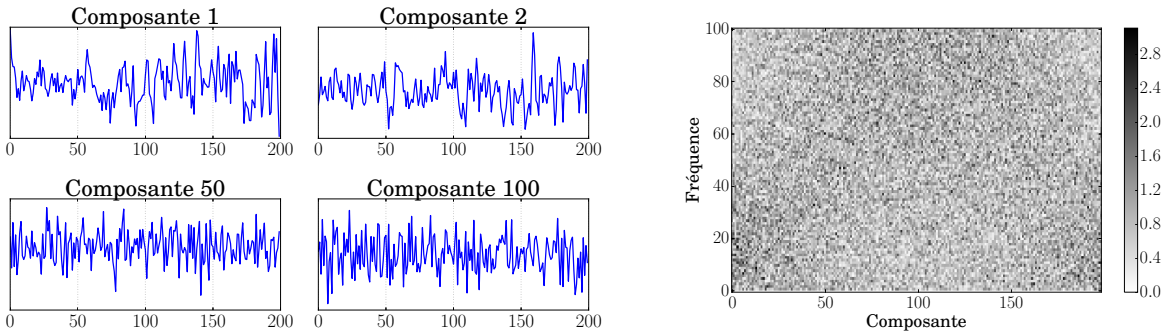
Shimada et al. [219] détaillent, en utilisant la théorie de la perturbation [139], une approximation au



(a) $p = 0.1$

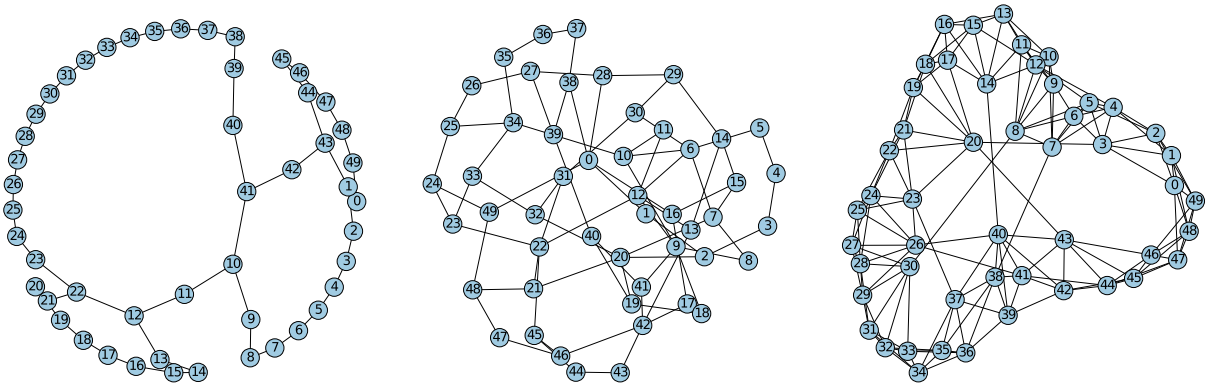


(b) $p = 0.5$



(c) $p = 0.9$

FIGURE 3.7 – Collection de signaux obtenue après transformation d'un graphe aléatoire de type Erdös-Rényi avec 200 nœuds, pour différentes valeurs de p. La figure de gauche affiche les composantes 1, 2, 50 et 100, alors que la figure de droite affiche les amplitudes pour chaque fréquence de chaque composante.



(a) $k = 2, p = 0.1$ (b) $k = 2, p = 0.5$ (c) $k = 6, p = 0.1$

FIGURE 3.8 – Visualisation d'un graphe aléatoire de type Watts-Strogatz, pour différentes valeurs de k et de p. La structure en anneau des graphes k-réguliers est perturbée à différents niveaux suivant la valeur de p.

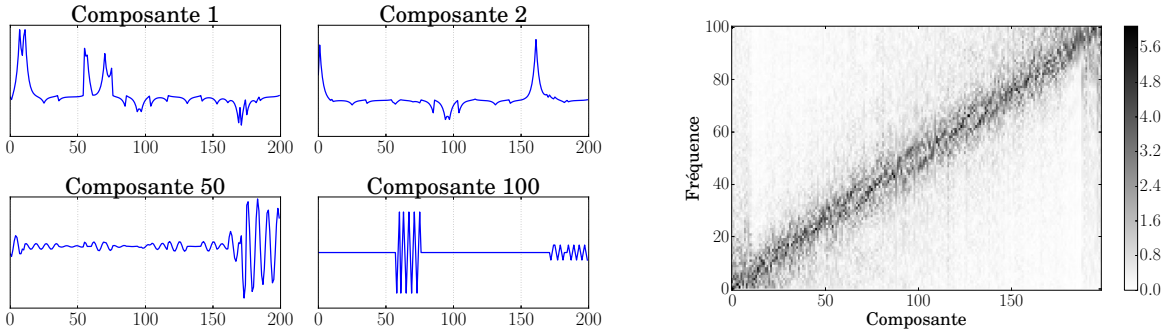
second ordre des valeurs propres et vecteurs propres de la matrice $\boldsymbol{B}$. À partir de ceux obtenus pour le graphe k-régulier, l'influence de la perturbation liée à l'ajout et à la suppression de liens est étudiée. Les simulations à partir de ce modèle perturbatif montre que plus p est élevée, plus le graphe a une structure aléatoire.

La Figure 3.9 affiche la collection de signaux obtenue après transformation de graphes générés à partir d'un modèle de Watts-Strogatz avec 200 nœuds pour différentes valeurs de k et p, de manière similaire à la Figure 3.4. La perturbation de la structure se retrouve facilement dans la représentation du motif fréquentiel : les formes caractéristiques des graphes k-réguliers, discutées et visibles à la Figure 3.4, sont toujours distinguables. En dehors de ces motifs, aucune structure particulière ne se dégage, ce qui rappelle le motif spectral obtenu pour les graphes aléatoires de type Erdös-Rényi. Sans surprise, la valeur de p a une grande influence sur la netteté de ce motif : lorsque p est faible, la forme reste assez distincte et ressort assez facilement du bruit. Au contraire, lorsque la valeur de p est assez importante, le motif du graphe k-régulier est beaucoup plus diffus et donc beaucoup moins distinguable. Cette observation est en cohérence à la fois avec la visualisation des graphes de la Figure 3.8, où la détérioration de la structure en anneau était visible, mais également avec les résultats sur le modèle de Watts-Strogatz [247, 219], qui confirme la transition d'un graphe structuré en anneau vers un graphe non structuré lorsque p augmente.

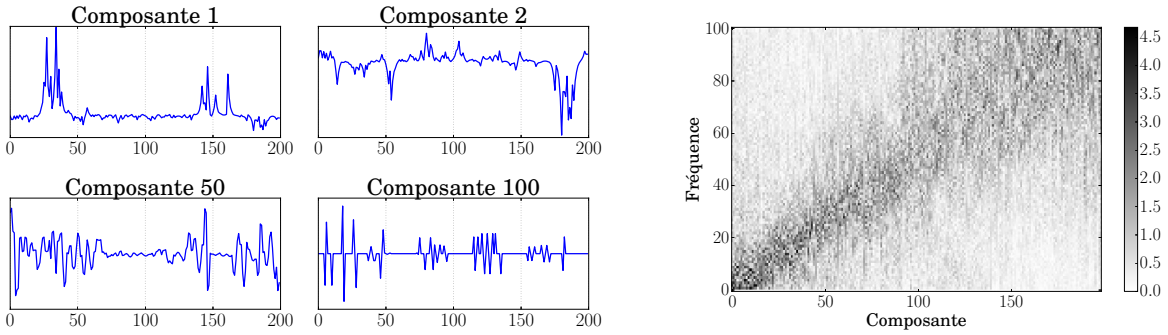
3.4 Modèle à blocs stochastiques

Le modèle à blocs stochastiques permet d'obtenir des réseaux avec des communautés. La Figure 3.10 affiche trois graphes générés à partir d'un modèle à blocs stochastiques, pour différentes valeurs du nombre de communautés C, de probabilité d'avoir un lien entre des nœuds de communautés différentes p_{inter} et de probabilité d'avoir un lien entre des nœuds d'une même communauté p_{intra}. La structure en communautés est clairement visible.

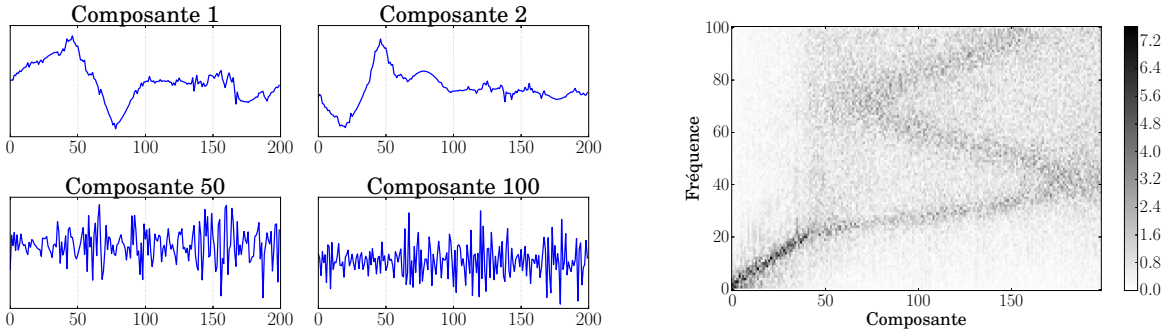
La détection de communautés est un sujet fertile en recherche et à ce titre, les propriétés spectrales de la matrice d'adjacence ont été étudiées avec attention. Sans rentrer dans les détails, il apparaît clairement que le nombre de communautés se retrouve dans les valeurs propres [54], à savoir que le nombre de valeurs propres significativement plus élevées que les autres est directement relié au nombre de communautés. De plus, les vecteurs propres associés à ces valeurs propres ont une forme qui se rapproche d'un vecteur constant par morceau, pour lequel chaque morceau correspond à une communauté [164]. L'intuition que l'on peut ainsi avoir est que la définition des communautés, c'est-à-dire le fait que les communautés soient bien distinctes, va influer sur la régularité de ces vecteurs propres, ainsi que sur l'in-



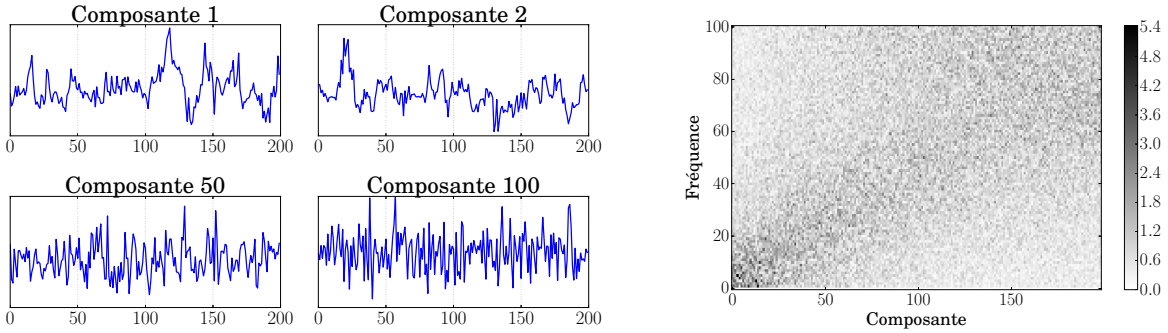
(a) $k = 2, p = 0.1$



(b) $k = 2, p = 0.5$

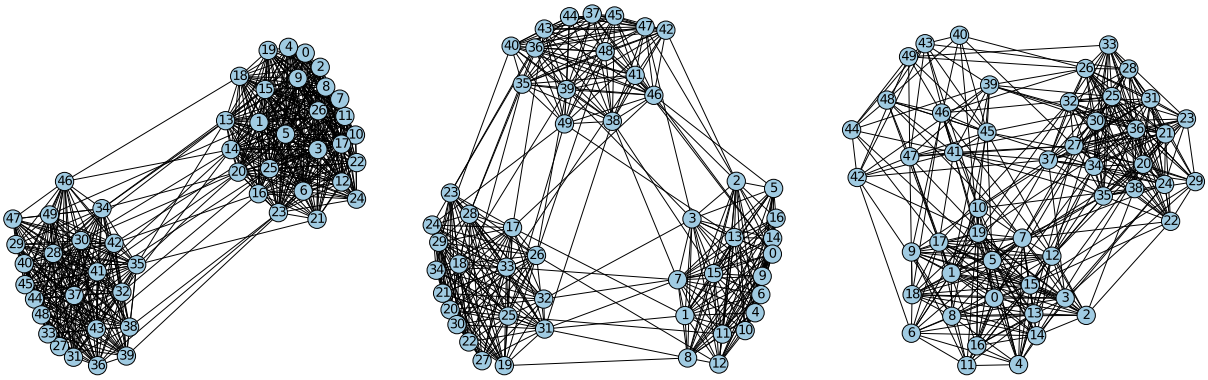


(c) $k = 6, p = 0.1$



(d) $k = 6, p = 0.5$

FIGURE 3.9 – Collection de signaux obtenue après transformation de graphes générés à partir d'un modèle de Watts-Strogatz avec 200 nœuds, pour différentes valeurs de p. La figure de gauche affiche les composantes 1, 2, 50 et 100, alors que la figure de droite affiche les amplitudes pour chaque fréquence de chaque composante.



(a) $C = 2$, $p_{\text{inter}} = 0.05$ et $p_{\text{intra}} = 0.9$ (b) $C = 3$, $p_{\text{inter}} = 0.05$ et $p_{\text{intra}} = 0.9$ (c) $C = 3$, $p_{\text{inter}} = 0.1$ et $p_{\text{intra}} = 0.7$

FIGURE 3.10 – Visualisation de graphes générés à partir d'un modèle à blocs stochastiques, pour différentes valeurs du nombre de communautés C, de probabilité d'avoir un lien entre des nœuds de communautés différentes p_{inter} et de probabilité d'avoir un lien entre des nœuds d'une même communauté p_{intra}. La structure en communautés est clairement visible, ainsi que l'influence des probabilités.

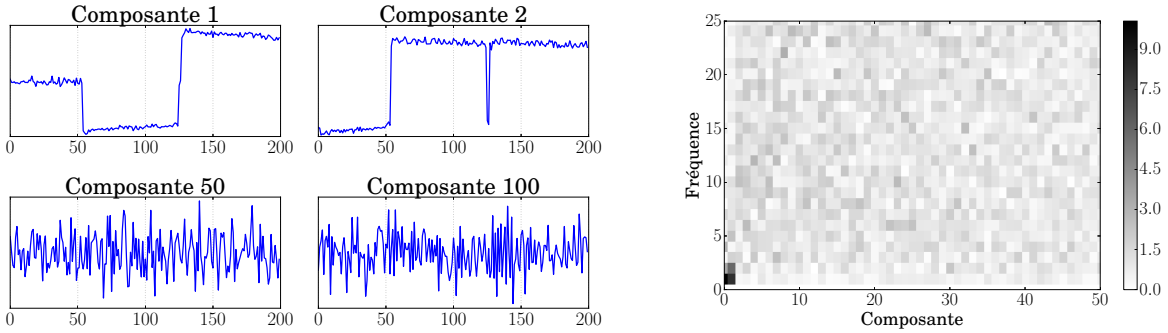
tensité des valeurs propres, et donc sur les signaux eux-mêmes. Comme il a été discuté dans le deuxième exemple présenté à la Section 1.1, un signal constant par morceau a un spectre dont l'énergie est principalement concentrée sur les basses fréquences. De plus, à l'intérieur des communautés, la structure aléatoire renvoie au modèle d'Erdös-Rényi. Le motif fréquentiel attendu est ainsi composé d'une forte énergie sur les basses fréquences pour les premières composantes, puis d'un motif non structuré ailleurs.

La Figure 3.11 affiche la collection de signaux obtenue après transformation de plusieurs graphes générés à partir d'un modèle à blocs stochastiques pour différentes valeurs de C, p_{inter} et p_{intra}. La figure de gauche affiche les composantes 1, 2, 50 et 100, alors que la figure de droite affiche les amplitudes pour chaque fréquence de chaque composante. Pour des raisons de visualisation, seules les 50 premières composantes et les 25 premières fréquences sont retenues, au-delà aucun motif particulier ne ressort. Comme décrit plus haut, le nombre de composantes constantes par morceau dépend du nombre de communautés, de même que le nombre de morceaux. Ainsi, dans le premier exemple avec 3 communautés, il y a trois plateaux visibles dans le signal, et uniquement les deux premières composantes sont composées de plateaux. Le reste des composantes, par exemple les composantes 100 et 200 affichées, présente une structure aléatoire, qui se retrouve dans le motif fréquentiel : les $C - 1$ premières composantes ont une énergie élevée et concentrée sur les basses fréquences, tandis que le reste du motif ne présente pas de structure particulière. À noter l'influence des probabilités, qui ont la même influence sur la structure du graphe que sur le motif fréquentiel, c'est-à-dire que les premières composantes sont moins énergétiques.

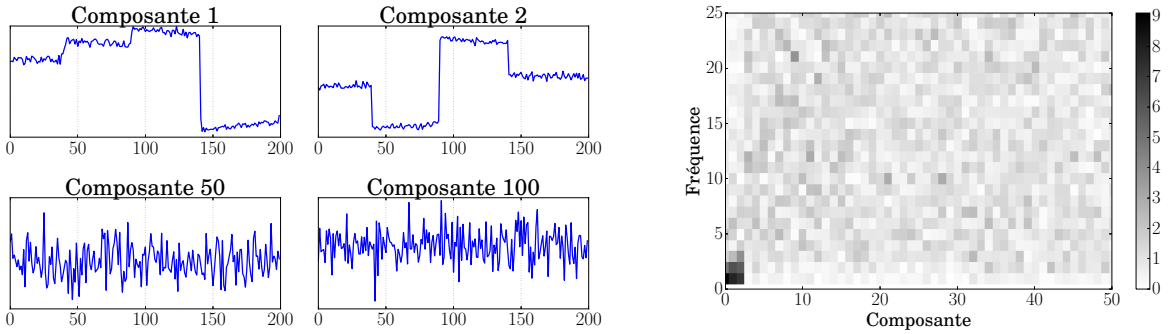
3.5 Modèle mixte de Watts-Strogatz à blocs stochastiques

Le dernier modèle étudié est un modèle mixte de Watts-Strogatz à blocs stochastiques, qui permet d'étudier des réseaux dans lesquels plusieurs structures co-existent, en l'occurrence une structure en anneau et en communautés. La Figure 3.12 affiche trois graphes générés à partir de ce modèle, pour différentes valeurs du nombre de communautés C et du degré k. Les probabilités sont fixées à 0.1 pour p_{inter} et 0.9 pour p_{intra}.

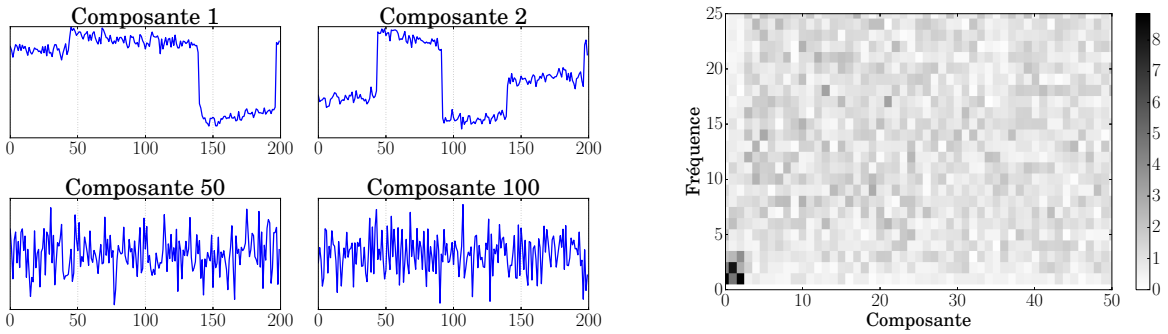
De la même manière que pour les résultats trouvés précédemment, la forme attendue des signaux résultants peut se deviner assez facilement, en supposant que la combinaison de plusieurs structures dans le graphe se traduit par la combinaison des motifs fréquentiels dans le plan composante-fréquence. Il est ainsi attendu de retrouver d'une part des signaux constants par morceau, représentant les communautés, des signaux proches d'une oscillation harmonique, représentant la structure en anneau, et des signaux



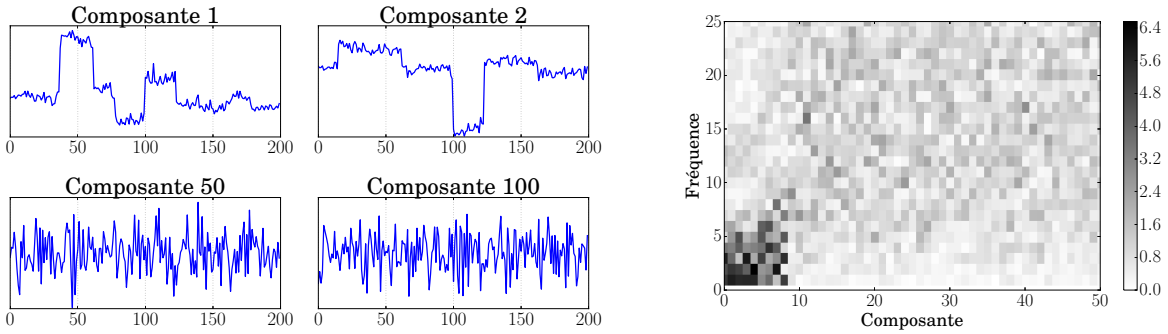
(a) $C = 3$, $p_{\text{inter}} = 0.05$ et $p_{\text{intra}} = 0.9$



(b) $C = 4$, $p_{\text{inter}} = 0.05$ et $p_{\text{intra}} = 0.9$

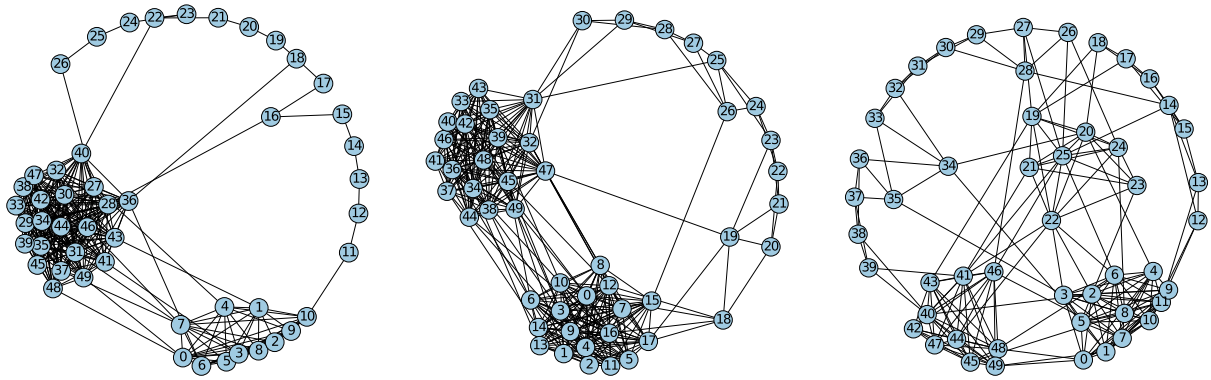


(c) $C = 4$, $p_{\text{inter}} = 0.1$ et $p_{\text{intra}} = 0.7$



(d) $C = 10$, $p_{\text{inter}} = 0.05$ et $p_{\text{intra}} = 0.9$

FIGURE 3.11 – Collection de signaux obtenue après transformation de graphes générés à partir d'un modèle à blocs stochastiques pour différentes valeurs du nombre de communautés C, de probabilité d'avoir un lien entre des nœuds de communautés différentes p_{inter} et de probabilité d'avoir un lien entre des nœuds d'une même communauté p_{intra}. La figure de gauche affiche les composantes 1, 2, 50 et 100, alors que la figure de droite affiche les amplitudes pour chaque fréquence de chaque composante. Pour des raisons de visualisation, seules les 50 premières composantes et les 25 premières fréquences sont retenues, au-delà aucun motif particulier ne ressort.



(a) $k = 2$ et $C = 2$ (b) $k = 2$ et $C = 2$ (c) $k = 4$ et $C = 3$

FIGURE 3.12 – Visualisation de graphes générés à partir d'un modèle mixte de Watts-Strogatz à blocs stochastiques, pour différentes valeurs du nombre de communautés C et du degré k. Les probabilités sont fixées de la façon suivante pour tous les graphes : $p_{\text{inter}} = 0.1$ et $p_{\text{intra}} = 0.9$. Les deux types de structures, en anneau et en communautés, sont visibles.

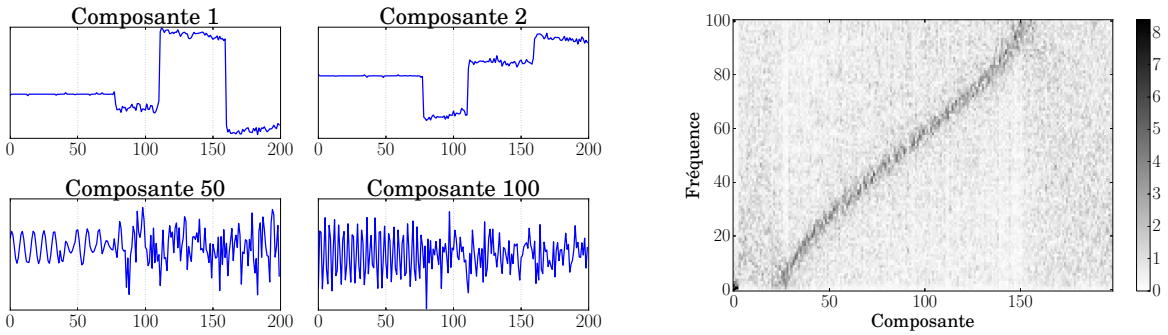
bruités représentant la structure aléatoire à l'intérieur des communautés.

La Figure 3.13 affiche la collection de signaux obtenue après transformation de plusieurs graphes générés à partir d'un modèle mixte de Watts-Strogatz à blocs stochastiques pour différentes valeurs du nombre de communautés C et du degré k. Les probabilités sont fixées pour tous les graphes à 0.1 pour p_{inter} et 0.9 pour p_{intra}. Les trois motifs fréquentiels évoqués se retrouvent effectivement, permettant de valider l'hypothèse de combinaison des motifs fréquentiels lorsque plusieurs structures apparaissent dans le réseau. Les premières composantes traduisent la structure en communautés du graphe, avec un point intéressant qui est que dans une première approximation, la structure régulière entre chaque communauté est considéré comme une communauté unique. Une fois la structure en communautés décrite, la structure en anneau apparaît, laissant apparaître des motifs similaires à ceux obtenus pour le modèle de Watts-Strogatz, dans une forme néanmoins qui semble légèrement déformée du fait de la présence de communautés au milieu de la structure en anneau.

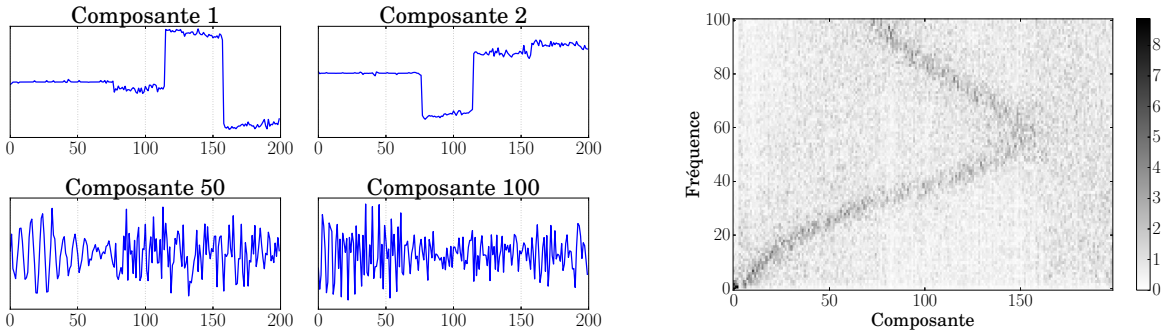
3.6 Discussions

Tout au long de cette section, différents modèles de graphes ont été étudiés de manière empirique mais également pour certains modèles de manière théorique. Des connexions claires entre la topologie du graphe et les motifs fréquentiels associés dans le plan composante-fréquence ont été établies, après transformation d'un graphe en une collection de signaux et analyse spectrale de ces signaux. De plus, des similitudes ont été mises en évidence entre les mélanges de structures de graphes et le mélange de motifs fréquentiels, permettant de mettre en lumière la dualité entre ces deux représentations.

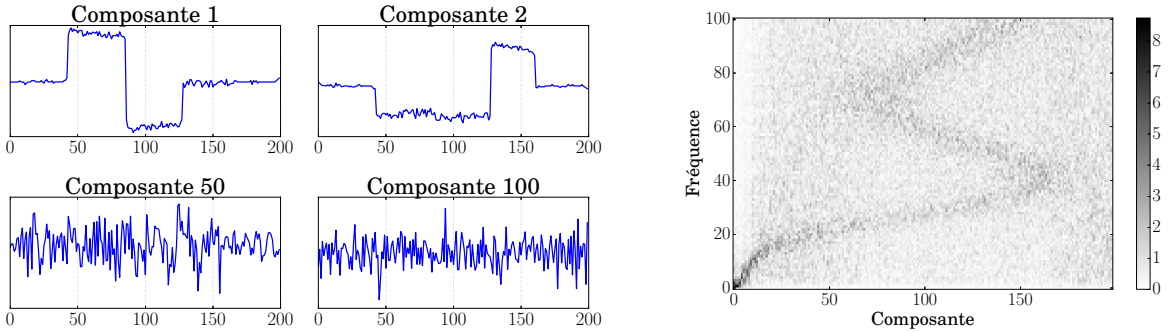
La Section 5 exploite cette dualité afin de réaliser des opérations sur le graphe via la collection de signaux. Une opération simple de débruitage d'un graphe est discutée : alors que le débruitage d'un graphe peut se révéler compliquer à mettre en œuvre [17], cette opération est réalisée à travers le débruitage des signaux correspondants, en utilisant les méthodes éprouvées de traitement du signal pour le débruitage de signaux. Néanmoins, avant de montrer cette application, il est nécessaire de pouvoir transformer une collection de signaux en graphes en définissant une transformation inverse robuste aux changements dans la collection de signaux, ce qui est réalisé dans la section suivante.



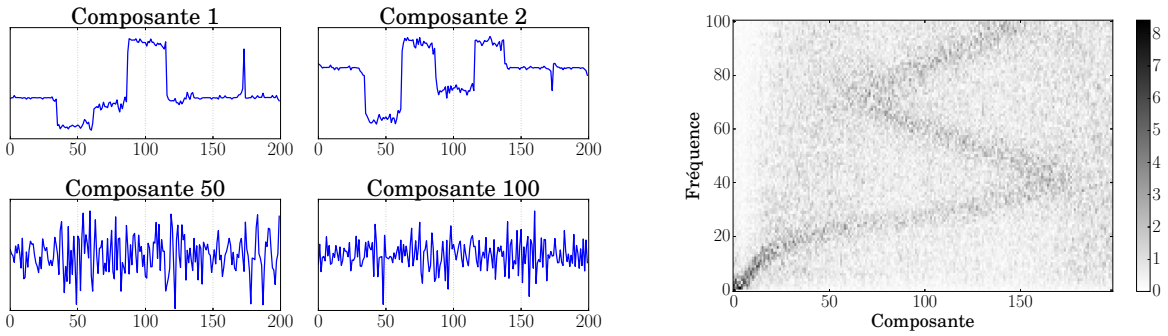
(a) $k = 2$ et $C = 3$



(b) $k = 4$ et $C = 3$



(c) $k = 6$ et $C = 3$



(d) $k = 6$ et $C = 4$

FIGURE 3.13 – Collection de signaux obtenue après transformation de graphes générés à partir d'un modèle mixte de Watts-Strogatz à blocs stochastiques pour différentes valeurs du nombre de communautés C et du degré k. Les probabilités sont fixées pour tous les graphes à 0.1 pour p_{inter} et 0.9 pour p_{intra}. La figure de gauche affiche les composantes 1, 2, 50 et 100, alors que la figure de droite affiche les amplitudes pour chaque fréquence de chaque composante.

4 Transformation inverse de signaux en graphes

4.1 Difficultés liées à la transformation inverse

Le passage de signaux vers des graphes a fait l'objet de plusieurs études, comme décrit dans la Section 1.3. Ces méthodes, dont l'objectif est l'étude de séries temporelles par la théorie des réseaux, ne sont pas adaptées pour définir une transformation inverse à l'approche proposée dans la Section 2, pour la simple raison que les signaux sont une représentation d'un graphe. En considérant les signaux comme de simples séries temporelles, cette représentation est ignorée, et la topologie du graphe reconstruit n'a aucun rapport avec celle du graphe initial. La logique qui guide le développement de la transformation inverse est la suivante : si la collection de signaux est directement issue de la transformation d'un graphe via la méthode proposée dans la Section 2, alors le graphe obtenu par transformation inverse doit être identique au graphe transformé [2]. Lorsque la collection de signaux est perturbée, alors le graphe reconstruit doit refléter après transformation inverse la perturbation dans des proportions équivalentes.

Le point de départ de la transformation inverse est le calcul des distances entre chaque paire de nœuds, que ce soit dans le cas non-dégradé ou le cas dégradé. On considère une transformation d'un graphe G en une collection de signaux $\boldsymbol{X}$. On note $\bar{\boldsymbol{X}}$ la collection de signaux possiblement dégradée obtenue à partir de $\boldsymbol{X}$. La matrice des distances $\bar{\boldsymbol{D}}(\bar{\boldsymbol{X}}) = \{\bar{d}(\bar{\boldsymbol{X}})_{uv}\}_{u,v\in\{1,\dots,n\}}$ s'obtient en calculant la distance euclidienne entre chaque paire de nœuds de la façon suivante :

$$\bar{d}(\boldsymbol{X})_{uv} = \sqrt{\sum_{c=1}^{C} (\bar{x}_{uc} - \bar{x}_{vc})^2} \tag{3.23}$$

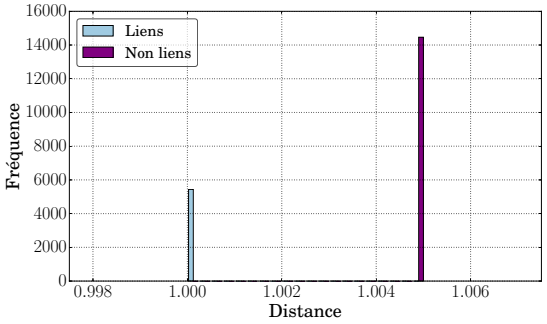
À partir de ces distances, l'objectif est de pouvoir retrouver une matrice de distance entre les nœuds $\bar{\boldsymbol{\Delta}}$ et donc un graphe $\bar{G} = (\bar{V}, \bar{E})$. La présence ou non d'un lien entre deux nœuds dépend de la distance entre les deux nœuds correspondants dans l'espace euclidien : si leur distance est grande, le lien n'existe pas et inversement, si la distance est petite, le lien existe. La difficulté va résider dans l'évaluation de la distance, afin de caractériser quand cette distance est grande, et quand cette distance est petite.

Lorsque la collection de signaux est directement issue de la transformation d'un graphe, c'est à dire que $\bar{\boldsymbol{X}} = \boldsymbol{X}$, cette caractérisation est immédiate : le positionnement multidimensionnel (MDS) construit justement la matrice $\boldsymbol{X}$ de telle sorte que $\boldsymbol{D}(\boldsymbol{X}) = \boldsymbol{\Delta}$. Les distances ont alors deux valeurs possibles, 1 et w, et les distances égales à 1 se retrouvent entre des paires de points connectés alors que celles égales à w concernent les paires de points non connectés.

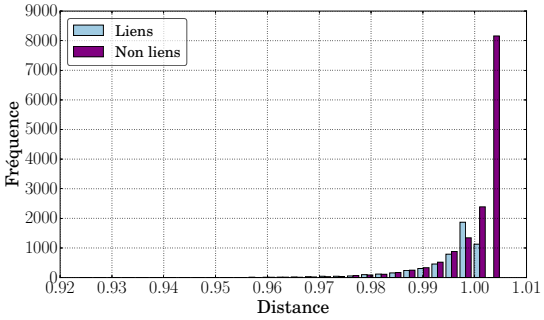
Lorsque la collection de signaux $\boldsymbol{X}$ est dégradée, c'est-à-dire que $\bar{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{P}$ avec $\boldsymbol{P}$ une matrice de perturbation, les distances ne peuvent plus être discriminées simplement : d'une part, les distances ne sont plus concentrées sur deux valeurs, et d'autre part, les distances représentant des liens et celles représentant des non-liens se mélangent. En conséquence, il n'est plus possible d'utiliser un simple seuillage entre les valeurs 1 et w pour déterminer la matrice d'adjacence.

La Figure 3.14 illustre ces deux cas sur un réseau de 200 nœuds avec 4 communautés, généré avec un modèle à blocs stochastiques : l'histogramme des distances euclidiennes entre chaque paire de points $\{d(\boldsymbol{X})_{uv}\}_{u,v\in\{1,\dots,n\}}$ est affiché, avec en bleu clair les distances correspondant à des paires de nœuds connectés dans le graphe initial, et en violet les distances aux paires de nœuds déconnectés. Lorsque la collection n'est pas dégradée (Figure 3.14a), les valeurs se répartissent entre deux valeurs 1 et w, chacune de ces deux valeurs représentant l'absence ou la présence d'un lien. Au contraire, lorsque la collection de signaux est dégradée (Figure 3.14b), ici en enlevant le signal avec la plus faible énergie, les distances se répartissent sur l'intervalle $[0.92, 1.01]$. De plus, la discrimination entre les distances représentant des liens et celles représentant l'absence de lien est beaucoup moins visible. Dans cet exemple, la seule discrimination possible est de noter les distances supérieures à 1 comme des non-liens ; pour les autres

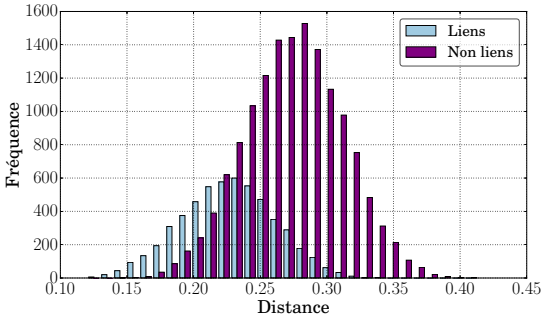
2. On supposera par la suite que la valeur de w est choisie de façon adéquate, comme discuté dans la Section 2.3.



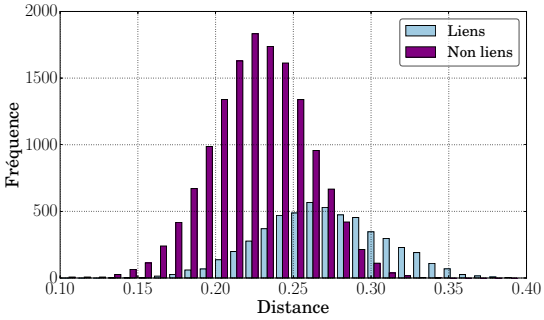
(a) Collection de signaux non dégradée



(b) Collection de signaux dégradée en enlevant la composante de plus basse énergie



(c) Collection de signaux dégradée en gardant les 50 premières composantes



(d) Collection de signaux dégradée en gardant les 50 dernières composantes

FIGURE 3.14 – Histogramme des distances euclidiennes entre chaque paire de points $\{d(\boldsymbol{X})_{ij}\}_{i,j\in\{1,\dots,n\}}$: les rectangles bleu clair représentent les paires de points correspondant à des paires de nœuds connectés dans le graphe initial, tandis que les rectangles violets correspondent aux paires de nœuds déconnectés. Les signaux sont obtenus après transformation d'un graphe de 200 nœuds avec 4 communautés, généré à l'aide d'un modèle à blocs stochastiques.

distances, l'incertitude demeure, sachant qu'en pratique le type de relation que les distances représentent n'est pas connu.

La section suivante présente une transformation inverse robuste aux modifications pouvant affecter la collection de signaux, avec un objectif de refléter les perturbations des signaux dans la structure du graphe reconstruit. La transformation proposée améliore la discrimination entre les distances représentant des liens et des non-liens, et intègre une procédure de seuillage guidée par les distances obtenues, sans connaissance a priori sur le graphe original.

4.2 Transformation inverse robuste

L'amélioration proposée se base sur le postulat que l'énergie des signaux a une importance dans la discrimination des distances. En effet, comme discuté dans la Section 2, les signaux avec une forte énergie ont une influence importante sur la description de la structure globale du réseau : si la distance entre deux nœuds u et v est élevée dans un signal à haute énergie, cela signifie que les deux nœuds sont probablement éloignés dans le graphe, au sens de la longueur du plus court chemin entre les deux nœuds. Au contraire, si la distance est faible, alors les nœuds sont probablement connectés, ou sinon dans un voisinage proche. Les Figures 3.14c et 3.14d illustrent cette hypothèse sur le graphe avec communautés présenté ci-dessus, en considérant les cas pour lesquels la collection de signaux est uniquement composée respectivement des 50 premiers signaux (donc les plus énergétiques) et des 50 derniers signaux (donc les

moins énergétiques). Dans le premier cas, les distances représentant des liens sont bien celles qui sont les plus faibles et malgré le recouvrement entre les deux types de distances, un seuillage conserverait une partie non négligeable des liens du graphe original. Au contraire, lorsque seules les composantes les moins énergétiques sont conservées, un seuillage mènerait à considérer une majorité de liens non présents dans le graphe initial, et donc à obtenir une structure sensiblement différente de la structure originale. La structure du graphe est présente dans les signaux les plus énergétiques et négliger l'importance de l'énergie des composantes prive ainsi la reconstruction du graphe d'une information importante sur la hiérarchie des composantes représentant la structure du réseau. Pour pallier ce problème, la solution consiste à calculer les distances en prenant en compte l'énergie des composantes dans le calcul.

4.2.1 Distances pondérées par l'énergie

L'intégration des énergies dans la transformation inverse s'effectue en considérant une nouvelle définition de la distance entre les points, incluant une pondération par les énergies des contributions de chaque composante. L'énergie de la composante c, notée z_c, est calculée de manière classique [3] de la façon suivante :

$$\bar{z}_c = \sum_{i=1}^{n} \bar{x}_{ic}^2 \tag{3.24}$$

permettant de définir la distance pondérée $\bar{d}^{(p)}(\bar{\boldsymbol{X}})_{uv}$ entre deux nœuds u et v, en se basant sur le fait que dans le cas pondéré, on a un poids de 1 associé à chaque composante, et que donc la somme totale des poids vaut C :

$$\bar{d}^{(p)}(\bar{\boldsymbol{X}})_{uv} = \sqrt{\frac{C}{\sum_{c=1}^{C} \bar{z}_c^{\alpha}} \sum_{c=1}^{C} \bar{z}_c^{\alpha} (\bar{x}_{uc} - \bar{x}_{vc})^2} \tag{3.25}$$

avec $\alpha \geq 0$. Le paramètre α contrôle l'importance de la pondération par les énergies : si α est élevé, les signaux à haute énergie ont plus d'importance dans le calcul des distances par rapport aux signaux à faible énergie. Lorsque $\alpha = 0$, on retrouve le cas non pondéré. En pratique, le choix de α se fait de façon empirique. Une fois les distances obtenues, un seuillage est réalisé de façon à obtenir une matrice d'adjacence binaire $\bar{\boldsymbol{A}}$. Ce seuillage est discuté dans la sous-section suivante.

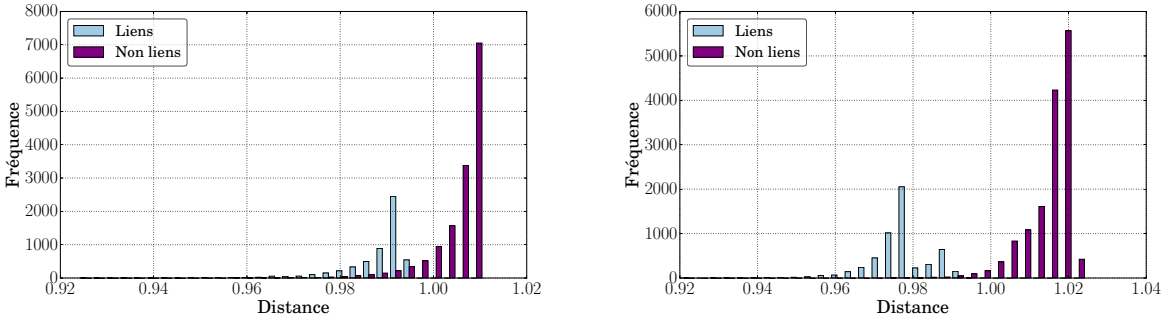
La Figure 3.15 propose une illustration des améliorations permises par l'ajout des énergies dans le calcul des distances, sur le même graphe que dans la Figure 3.14. La figure de gauche affiche les distances pondérées calculées avec $\alpha = 2$, alors que la figure de droite utilise $\alpha = 5$. La Figure 3.15a présente le cas où la collection de signaux est dégradée en enlevant le signal de plus faible énergie, alors que pour la Figure 3.15b, la collection de signaux est dégradée en ne gardant que les 50 premières composantes.

Par rapport aux Figures 3.14b et 3.14c, qui montre l'histogramme des distances lorsque la pondération par les énergies n'est pas mise en place, on remarque d'une part une meilleure séparation des distances, en particulier pour le premier exemple. De plus, la valeur de α a une influence notable sur la séparabilité des distances : plus α est élevé, plus la discrimination entre les deux types de distances est claire.

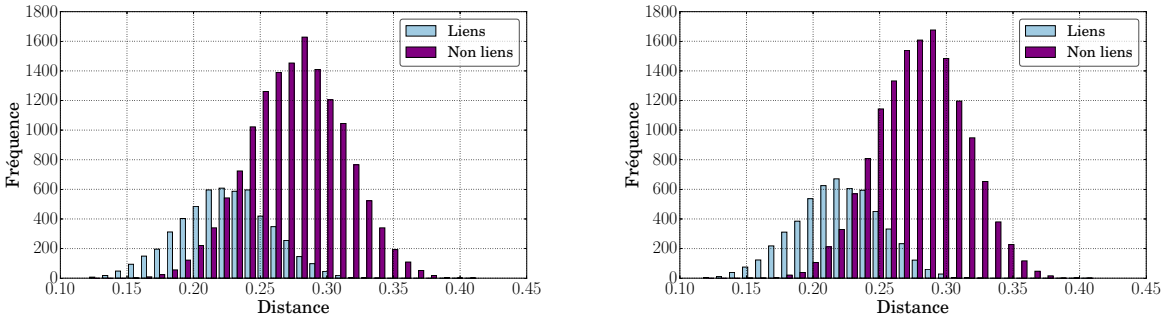
4.2.2 Seuillage des distances

La sélection du seuil est, comme cela a pu être remarqué plus haut, une étape cruciale dans la différentiation entre les distances représentant des liens et les distances représentant des non-liens. Plusieurs approches co-existent : la première consiste à utiliser l'information sur le graphe original, en préservant

3. Par la relation de Parseval, cette définition donne le même résultat que le calcul de l'énergie du spectre, d'où la même notation



(a) Collection de signaux dégradée en enlevant le signal de plus faible énergie



(b) Collection de signaux dégradée en ne gardant que les 50 premières composantes

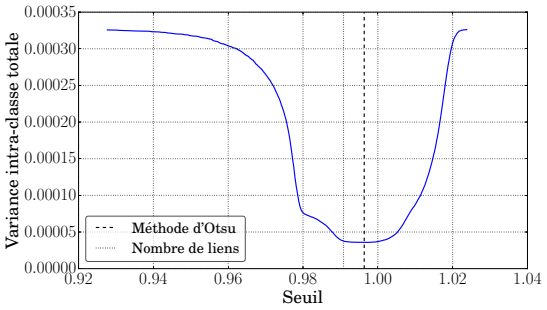
FIGURE 3.15 – Histogramme des distances euclidiennes calculées en prenant en compte l'énergie des signaux. La figure de gauche affiche les distances pondérées calculées avec $\alpha = 2$, alors que la figure de droite utilise $\alpha = 5$.

le nombre de liens m du graphe. Comme déjà mentionné dans [219], la sélection du seuil se fait en sélectionnant la valeur telle que le nombre de paires de nœuds dont la distance est inférieure à ce seuil est égal à m. Cette approche peut néanmoins être limitée, car elle suppose que le nombre de liens dans le graphe reconstruit est identique au nombre de liens dans le graphe de départ. Or, la modification de la collection de signaux, et donc la modification de la structure du graphe qu'il représente, peuvent faire varier sensiblement le nombre de liens.

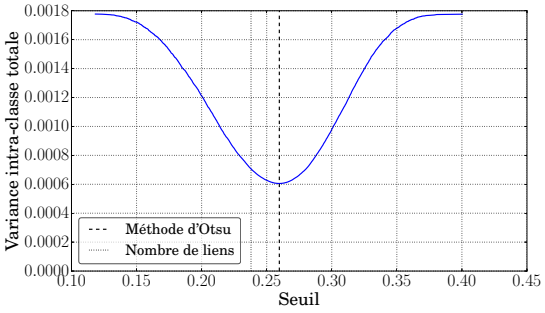
La méthode qui est proposée par la suite pour réaliser le seuillage des distances est inspirée des techniques de binarisation en traitement d'image, qui cherche à obtenir, à partir d'une image avec des pixels pouvant prendre plusieurs couleur, une image en noir et blanc. Lorsque l'image est composée de niveaux de gris, c'est-à-dire lorsque les pixels sont décrits uniquement par une intensité de la couleur noire, un algorithme standard de binarisation a été proposé par Otsu [184] : pour tous les seuils possibles, les variances des intensités dans la classe des pixels coloriés en blancs et dans la classe des pixels coloriés en noirs sont calculées, et le seuil choisi est celui qui minimise la somme de ces variances.

Dans le contexte de discrimination des distances, la distribution considérée n'est plus celle des intensités des pixels, mais des distances entre chaque paire de points. Contrairement au cas des images, où le nombre de seuils est égal au nombre de niveau d'intensité possible (par exemple 256 dans une image 8 bits) et ce quelle que soit la taille de l'image, ici le nombre de seuils possibles est égal au maximum au nombre de paires de nœuds, c'est à dire $\frac{n(n-1)}{2}$. Deux solutions s'offrent alors, la première consiste à choisir les seuils parmi toutes les distances possibles, et donc de faire $\frac{n(n-1)}{2}$ calculs. Une deuxième solution moins coûteuse consiste à discrétiser l'espace des distances pour ne garder qu'un nombre réduit de seuils possibles. La première solution est retenue par la suite, étant donnée la taille relativement faible des graphes que l'on traite.

L'algorithme 7 propose la procédure de seuillage adaptée de la méthode d'Otsu pour discriminer les



(a) Collection de signaux dégradée en enlevant la composante de plus basse énergie



(b) Collection de signaux dégradée en gardant les 50 premiers signaux

FIGURE 3.16 – Évolution de la variance totale intra-classe v_t, en fonction du seuil choisi, pour le seuillage des distances pondérées calculées avec $\alpha = 5$. Les lignes verticales représentent les seuils choisis par chacune des deux méthodes de seuillage.

distances représentant des liens et les distances représentant l'absence de lien. Le choix du nombre de liens est ainsi guidé par les distances issues de la collection de signaux, et ne dépend plus directement de la structure originale.

Algorithme 7 : Procédure de seuillage des distances adaptée de la méthode d'Otsu

Entrées : $\boldsymbol{d}$ une distribution de distances
Sorties : $\bar{\tau}$ un seuil

début
  pour *Pour chaque distance dans* $\boldsymbol{d}$*, définissant le seuil* τ **faire**
    Calculer les ensembles $\boldsymbol{d}_B$ et $\boldsymbol{d}_F$ comme les valeurs de $\boldsymbol{d}$ respectivement plus petites et plus grandes que le seuil
    Calculer w_B et w_F comme les proportions d'éléments respectivement dans $\boldsymbol{d}_B$ et $\boldsymbol{d}_F$
    Calculer les variances v_B et v_F de respectivement $\boldsymbol{d}_B$ et $\boldsymbol{d}_F$
    Stocker pour le seuil τ la variance totale $v_\tau = w_B v_B + w_F v_F$
  Sélectionner $\bar{\tau}$ tel que v_τ est minimal

La Figure 3.16 montre l'évolution de la variance totale intra-classe v_t, en fonction du seuil choisi, pour le seuillage des distances pondérées calculées avec $\alpha = 5$. Les lignes verticales représentent les seuils choisis par chacune des deux méthodes de seuillage. Les distributions de distances sont celles présentées dans les Figures 3.15a et 3.15b. En comparant les histogrammes et les seuils obtenus, on remarque que les deux méthodes donnent des résultats proches mais pas identiques, menant ainsi à des reconstructions différentes : la méthode d'Otsu a ainsi tendance à considérer plus de distances comme étant des liens par rapport à un seuillage basé sur le nombre de liens dans le graphe original.

4.3 Évaluation des performances

4.3.1 Protocole expérimental

Afin d'évaluer l'impact de ces modifications sur la reconstruction du graphe à partir d'une collection de signaux, dégradée ou non-dégradée, les performances de la transformation inverse robuste proposée sont évaluées. Cette évaluation est difficile à réaliser de manière précise : si la collection de signaux est directement obtenue en utilisant une transformation de graphe vers signaux introduite à la Section 2, alors la transformée inverse est triviale et ne nécessite pas la mise en place d'une méthode sophistiquée. Inversement, si la collection de signaux n'est pas la représentation directe d'un graphe, ou si les signaux

sont perturbés, alors il n'est pas possible de connaître l'effet réel de cette perturbation sur la structure du graphe, et donc d'évaluer exactement la reconstruction du graphe.

Afin de pouvoir néanmoins évaluer les performances des améliorations proposées, le graphe reconstruit $\bar{G}$ obtenu à partir d'une collection de signaux perturbée est comparé au le graphe original G à l'aide d'un indice de similarité, noté $Q(G, \bar{G})$, basé sur l'indice de Jaccard [130] déjà évoqué dans le Chapitre 2. La qualité de la reconstruction est mesurée en comparant les ensembles des liens de chaque graphe, en calculant le ratio entre le nombre de liens en commun entre G et $\bar{G}$ sur le nombre total de liens :

$$Q = \frac{|E \cap \bar{E}|}{|E \cup \bar{E}|} \tag{3.26}$$

Cet indice est ainsi compris entre 0 si aucun lien n'est en commun, et 1 si l'ensemble des liens est identique pour les deux graphes. Il est important de noter qu'avec cet indicateur, obtenir 0 ne veut pas forcément signifier que la structure est complètement différente. Dans le cas par exemple d'un graphe avec deux communautés, le graphe $\bar{G}$ tel que $Q = 0$ présente une structure où les nœuds de la communauté ne sont pas reliés entre eux, définissant des anticommunautés [259].

Deux types de perturbation de la matrice $\boldsymbol{X}$ sont considérés par la suite :

1. suppression de signaux : à partir de la collection complète, les signaux sont peu à peu enlevés, en commençant par les moins énergétiques ;
2. ajout de bruit : un bruit gaussien est ajouté, de moyenne 0 et de variance σ, avec σ pris dans un ensemble contenant s valeurs comprises entre 0 % et 10 % de la valeur maximale des signaux.

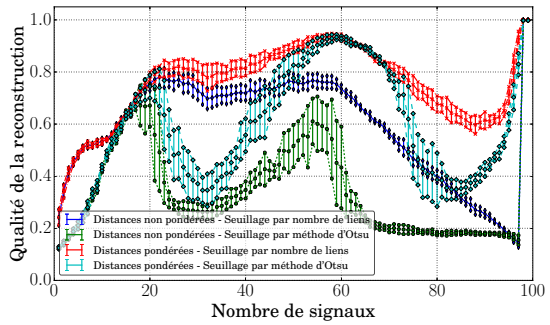
Quatre configurations sont étudiées, suivant si les distances sont calculées en prenant en compte l'énergie des composantes ou non, puis suivant la méthode de seuillage utilisée. Lorsque les distances sont calculées en utilisant l'énergie des composantes, le paramètre α est arbitrairement fixé à 5. Plusieurs structures de graphe sont étudiées, parmi celle décrites à la Section 3. Pour chaque structure, 5 instances sont générées. Le nombre de nœuds est fixé à 100 pour la première perturbation, et à 200 pour la deuxième, en raison de la différence de coût algorithmique entre les deux expériences.

4.3.2 Résultats et discussions

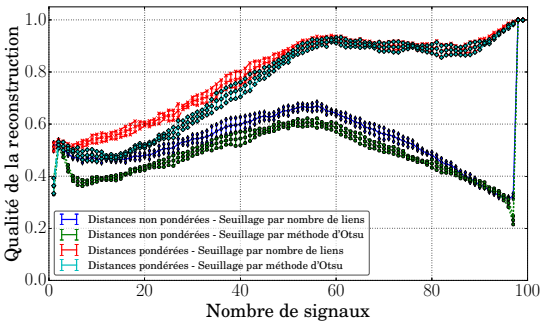
Pour chacune des deux expériences, les résultats sont présentés sous la forme suivante : pour chaque structure, la moyenne de la qualité de la reconstruction obtenue est affichée, ainsi que l'écart type autour de la moyenne. Chaque figure est composée de 4 courbes, correspondant à chacune des configurations de l'expérience.

La Figure 3.17 affiche les résultats lorsque la perturbation consiste à ne garder qu'un nombre réduit de signaux. La structure du réseau a une influence importante sur la forme des courbes. Notamment, les graphes avec une structure en anneau laisse apparaître des irrégularités dans la reconstruction : la qualité de reconstruction ne dépend pas directement du nombre de composantes retenues comme l'on pourrait s'y attendre. Cela peut s'expliquer en partie par le fait qu'il est possible, avec un nombre réduit de composantes, de retrouver la structure régulière, alors qu'au contraire, ajouter des composantes va amener à considérer les liens en dehors de cette structure régulière, qui peuvent perturber la reconstruction.

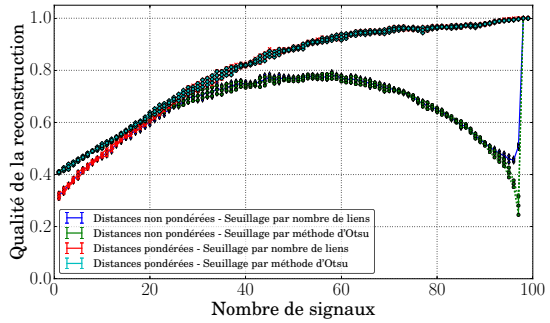
Les résultats de la transformation inverse permettent tout d'abord de voir que sans surprise, lorsque la collection de signaux n'est pas dégradée, la qualité de reconstruction est égale à 1, et ce pour toutes les configurations. En revanche, une légère perturbation de $\boldsymbol{X}$, lorsque le signal de plus faible énergie est enlevé, modifie sensiblement la structure du graphe lorsque les distances ne sont pas pondérées. Inversement, la prise en compte de l'énergie dans le calcul des distances permet d'obtenir une structure de $\hat{G}$ proche de celle de G. Le cas idéal est représenté lorsque la structure est aléatoire, pour laquelle plus la perturbation augmente, plus la structure du graphe reconstruit s'éloigne de celle du graphe original. Globalement, l'utilisation de la pondération par les énergies des composantes dans le calcul des distances



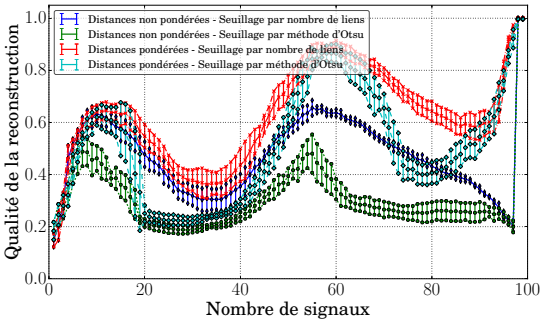
(a) Modèle de Watts-Strogatz avec $k = 6$, $p = 0.1$



(b) Modèle à blocs stochastiques avec $C = 4$, $p_{\text{intra}} = 0.7$ et $p_{\text{inter}} = 0.1$



(c) Modèle d'Erdös-Rényi avec $p = 0.4$



(d) Modèle mixte de Watts-Strogatz à blocs stochastiques avec $C = 4$, $k = 6$, $p_{\text{intra}} = 0.7$ et $p_{\text{inter}} = 0.1$

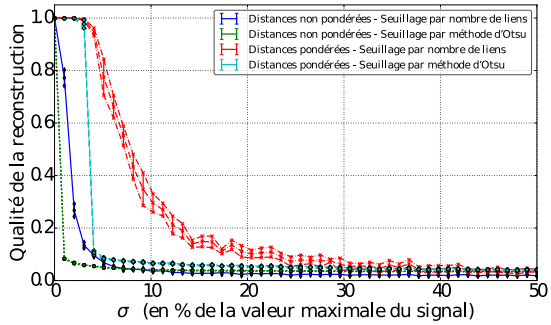
FIGURE 3.17 – Évaluation des performances de la transformation inverse robuste. La collection de signaux est perturbée en ne gardant qu'un nombre restreint de signaux. Pour chaque structure, la moyenne de la qualité de la reconstruction obtenue est affichée, ainsi que l'écart type autour de la moyenne. Chaque figure est composée de 4 courbes, correspondant à chacune des configurations de l'expérience.

améliore significativement la reconstruction, en répercutant de manière plus linéaire les perturbations sur les signaux dans la structure du graphe.

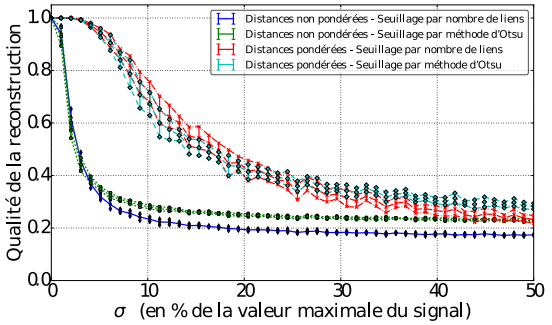
La comparaison entre les deux méthodes de seuillage est compliquée, étant donné qu'il existe dans la mesure un biais selon lequel prendre exactement le même nombre de liens dans le graphe reconstruit que dans le graphe original va amener à de meilleurs résultats dans la comparaison des deux graphes. Même en ne prenant pas en compte ce biais, on remarque que les deux courbes sont assez proches l'une de l'autre, mettant en évidence que la méthode d'Otsu permet de seuiller les distances de manière satisfaisante.

La Figure 3.18 affiche les résultats lorsque la collection de signaux est perturbée en ajoutant un bruit gaussien, de moyenne 0 et de variance σ, avec σ exprimée en pourcentage de la valeur maximale de $\boldsymbol{X}$. Les résultats sont ici beaucoup plus nets, et mettent bien en évidence que plus la variance est grande, c'est-à-dire plus la perturbation est importante, plus l'impact sur la structure du graphe est important. Cette expérience vient confirmer les résultats obtenus précédemment, à savoir que la prise en compte des énergies dans le calcul des distances donne un comportement de la transformation inverse beaucoup plus robuste. De même, les deux méthodes de seuillage obtiennent des résultats assez proches.

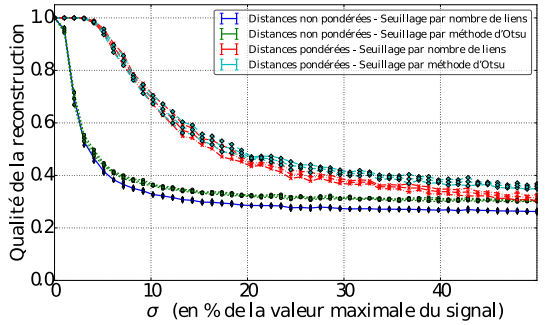
En conclusion de cette section, la prise en compte de l'énergie des composantes a permis de définir une transformation inverse qui permet, à partir d'une collection de signaux perturbée, de répercuter cette perturbation dans la structure du graphe. S'il n'est pas possible d'évaluer de quelle façon celle-ci se répercute, les résultats semblent néanmoins indiquer que la structure est légèrement affectée lorsque la perturbation est faible, et au contraire s'éloigne beaucoup plus de la structure du graphe original



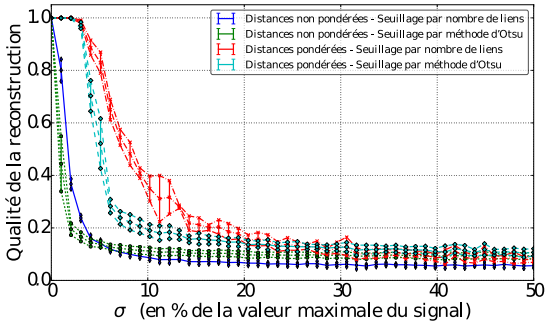
(a) Modèle de Watts-Strogatz avec $k = 6$, $p = 0.1$



(b) Modèle à blocs stochastiques avec $C = 4$, $p_{\text{intra}} = 0.7$ et $p_{\text{inter}} = 0.1$



(c) Modèle d'Erdös-Rényi avec $p = 0.4$



(d) Modèle mixte de Watts-Strogatz à blocs stochastiques avec $C = 4$, $k = 6$, $p_{\text{intra}} = 0.7$ et $p_{\text{inter}} = 0.1$

FIGURE 3.18 – Évaluation des performances de la transformation inverse robuste. La collection de signaux est perturbée en ajoutant un bruit gaussien, de moyenne 0 et de variance σ, avec σ exprimée en pourcentage de la valeur maximale des signaux. Pour chaque structure, la moyenne de la qualité de la reconstruction obtenue est affichée, ainsi que l'écart type autour de la moyenne. Chaque figure est composée de 4 courbes, correspondant à chacune des configurations de l'expérience.

lorsque la perturbation est importante. Ces résultats sont conformes à nos attentes, et appellent à l'étude de l'influence de perturbations de signaux sur la structure du graphe, ce qui est réalisé à travers une application pour le débruitage de graphe.

5 Traitement sur le graphe par les outils de traitement du signal

Les sections précédentes ont permis de définir une méthode de transformation d'un graphe en une collection de signaux, ainsi que la transformation inverse qui permet, à partir d'une collection de signaux, de retrouver le graphe. Cette transformation inverse a la particularité qu'elle est capable de prendre en compte les modifications de la collection de signaux, en les répercutant, autant que possible, sur la structure du graphe. Cette section présente maintenant comment il est possible de faire du traitement de graphe à l'aide des outils de traitement du signal.

La Figure 3.19 donne une vue d'ensemble du cadre d'étude pour réaliser le filtrage d'un graphe à travers les signaux correspondants, et permet de récapituler les différents objets et opérations rencontrés jusque-là. La procédure est la suivante : on souhaite étudier un réseau représenté sous la forme d'un graphe G. Les nœuds du graphe sont d'abord étiquetés de façon à obtenir une séquence de nœuds cohérente avec la structure du graphe. Pour cela, l'algorithme développé dans le Chapitre 2 est appliqué, et donne un graphe G^r correspondant au graphe G avec les nouvelles étiquettes pour les nœuds. À partir de

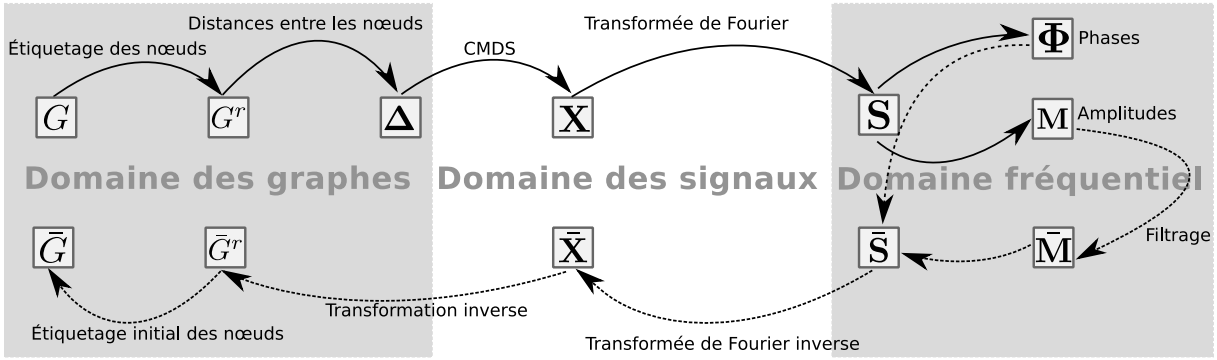


FIGURE 3.19 – Diagramme décrivant le cadre d'étude pour réaliser le filtrage d'un graphe à travers les signaux correspondants. Les carrés décrivent des objets, tandis que les flèches représentent les différentes opérations.

ce graphe, une matrice de distance $\boldsymbol{\Delta}$ est calculée à l'aide de l'équation 3.8. Une collection de signaux $\boldsymbol{X}$ est ensuite obtenue en appliquant le positionnement multidimensionnel (CMDS) sur la matrice $\boldsymbol{\Delta}$. Le graphe change ainsi de représentation et devient une collection de signaux. Afin d'étudier ces signaux, une analyse spectrale est réalisée sur les colonnes de la matrice $\boldsymbol{X}$, permettant d'obtenir les spectres $\boldsymbol{S}$ des signaux. Une méthode de filtrage est ensuite appliquée sur ces spectres, donnant une matrice $\bar{\boldsymbol{S}}$ qui, après transformation de Fourier inverse, redonne une collection de signaux $\bar{\boldsymbol{X}}$ qui n'est plus cette fois une représentation exacte d'un graphe. Grâce à l'utilisation de la transformation inverse robuste définie dans la Section 4, un nouveau graphe $\bar{G}^r$ est obtenu, puis le graphe $\bar{G}$ en redonnant à chaque nœud son étiquette initiale.

Avant de mettre en œuvre ce processus sur des exemples, une méthode de débruitage de signaux est présentée.

5.1 Filtrage de Wiener

Le filtrage de Wiener [249] est une technique développée par Norbert Wiener pour le débruitage de signaux, dont l'objectif initial fut la prédiction, à partir de la trajectoire d'un avion ennemi, de son point de passage le plus probable de façon à positionner efficacement les défenses antiaériennes [159].

Le problème est le suivant : à partir d'un signal $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{b}$, supposé égal à la somme du signal d'intérêt $\boldsymbol{s}$ et de bruit $\boldsymbol{b}$, l'objectif est d'estimer au mieux la valeur de $\boldsymbol{s}$. On suppose que $\boldsymbol{s}$ et $\boldsymbol{b}$ sont des processus stationnaires et de moyenne égale à 0. On cherche ainsi un filtre linéaire stationnaire qui donne la meilleure approximation de $\boldsymbol{s}$, que l'on note $\hat{\boldsymbol{s}}$. Le filtre de Wiener consiste à minimiser l'erreur quadratique moyenne $e = \mathbb{E}[(\boldsymbol{s} - \hat{\boldsymbol{s}})^2]$ où $\mathbb{E}(.)$ désigne l'espérance mathématique.

La minimisation de e, dans lequel intervient le terme inconnu $\boldsymbol{s}$, se fait en décomposant le critère e comme une somme de la fonction d'autocorrélation du signal $\boldsymbol{s}$, de la fonction d'autocorrélation de $\boldsymbol{x}$ et de la fonction d'intercorrélation entre $\boldsymbol{s}$ et $\boldsymbol{x}$. En utilisant les hypothèses sur le signal, et en passant dans le domaine spectral par transformation de Fourier, il est ainsi possible d'obtenir l'estimation $\hat{\boldsymbol{s}}$. Les détails techniques peuvent être trouvés dans [199].

5.2 Débruitage de graphe par filtrage des signaux

En utilisant la procédure donnée dans la Figure 3.19, une application pour le débruitage de graphe est proposée. La collection de signaux $\boldsymbol{X}$, obtenue à partir du graphe G que l'on souhaite débruiter, est filtrée à l'aide d'un filtre de Wiener, puis reconstruite en graphe en utilisant la transformation inverse définie dans la Section 4, en utilisant $\alpha = 5$ pour le calcul des distances pondérées et la méthode d'Otsu comme méthode de seuillage.

Le débruitage est appliqué sur trois réseaux, le premier obtenu à partir d'un modèle de Watts-Strogatz de degré 6 avec une probabilité réglée à 0.8, le deuxième à partir d'un modèle à blocs stochastiques avec 4 communautés et $p_{\text{intra}} = 0.7$ et $p_{\text{inter}} = 0.2$, et le troisième un modèle mixte de Watts-Strogatz à blocs stochastiques de degré 6, avec 4 communautés et $p_{\text{intra}} = 0.7$ et $p_{\text{inter}} = 0.1$. Le bruit est délibérément réglé à des niveaux assez élevés de manière à d'étudier le débruitage dans des conditions difficiles.

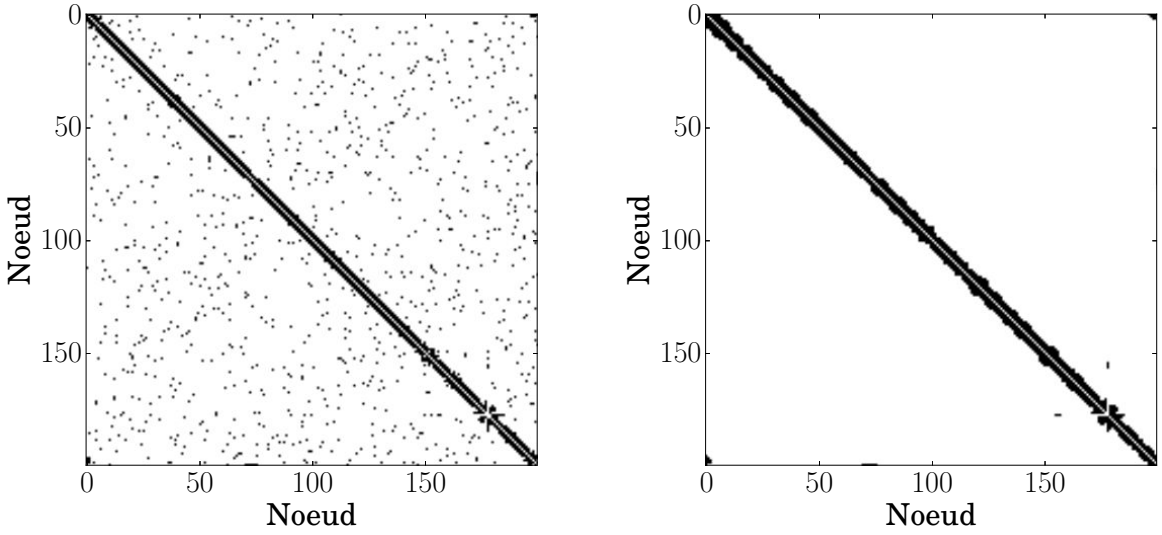
La Figure 3.20 affiche pour les trois types de graphes considérés les matrices d'adjacence des réseaux de manière à faire ressortir visuellement la structure. La figure de gauche affiche la matrice d'adjacence avant débruitage, alors que la figure de droite affiche la matrice d'adjacence après le débruitage. Il apparaît clairement que le débruitage des signaux permet de débruiter le graphe. Dans le cas du modèle de Watts-Strogatz, seule la diagonale, représentant le graphe 6-régulier en anneau, est visible après débruitage. De même pour le graphe avec les communautés, le graphe obtenu après débruitage devient un regroupement de 4 cliques, correspondant à chaque communauté, et révélant ainsi de manière claire la structure.

Le troisième cas est intéressant, car il permet de mettre en évidence une limite du filtrage de Wiener dans ce cas. Le résultat obtenu est un réseau dans lequel les quatre communautés ressortent distinctement comme dans le cas précédent, mais une cinquième communauté apparaît, constituée des nœuds de la partie régulière. Le débruitage supprime ainsi la structure secondaire du graphe en ne conservant qu'un découpage en communautés. Ce résultat fait écho à ceux discutés dans la Section 3.5.

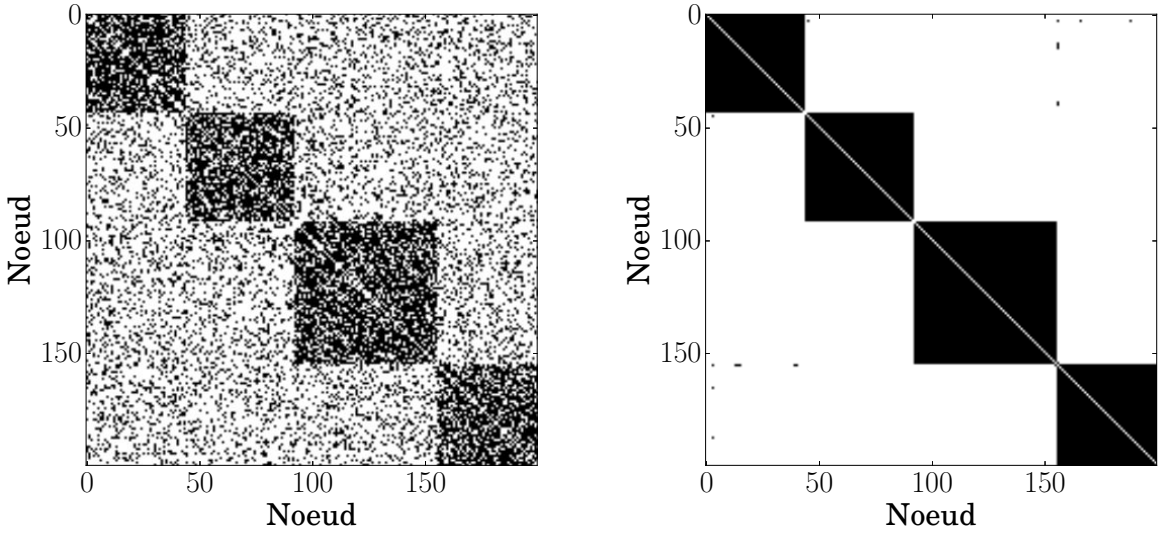
6 Conclusion et perspectives

Dans ce chapitre, une méthode innovante d'analyse de graphe est proposée, à l'aide des outils de traitement du signal. À partir de la méthode de transformation d'un graphe vers une collection de signaux proposée par Shimada et al. [219], une extension est développée de manière à étudier la collection de signaux afin d'en tirer de la connaissance sur le graphe correspondant. Une première contribution a consisté à proposer des résultats empiriques et théoriques sur des modèles de graphe à la structure bien déterminée, de manière à relier des motifs fréquentiels à la topologie du graphe. Une deuxième contribution a permis de définir une transformation inverse robuste d'une collection de signaux vers un graphe, en faisant en sorte que les modifications des signaux se répercutent de manière appropriée sur la structure du graphe. Ces deux extensions ont permis de mettre en place un débruitage du graphe à travers le débruitage de son image dans le domaine des signaux, en utilisant un simple filtrage de Wiener.

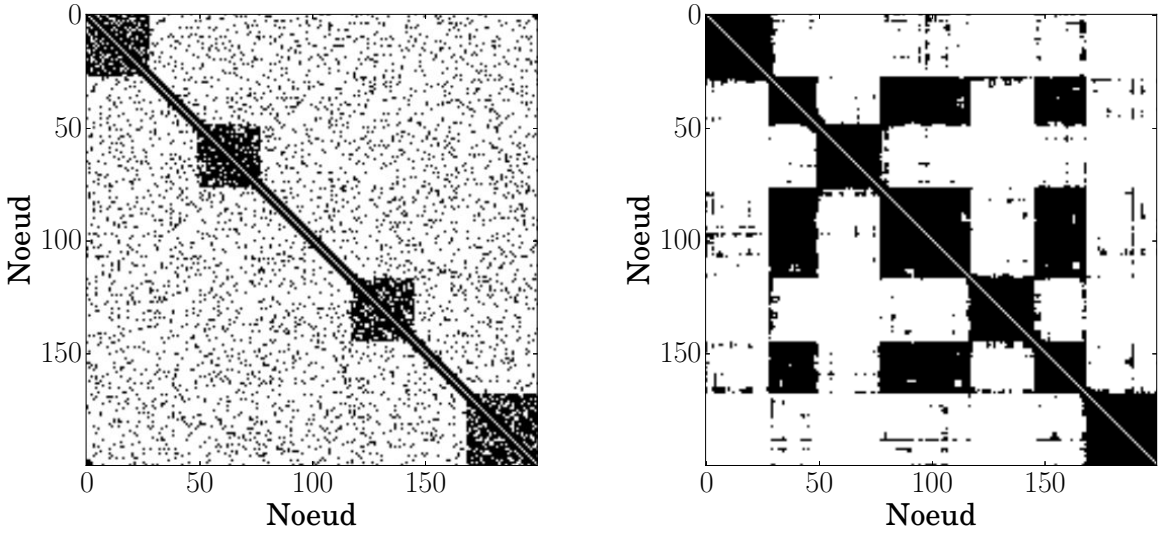
Ces travaux, s'ils permettent déjà d'obtenir des résultats prometteurs sur des modèles de graphe simple, posent néanmoins de nombreuses questions. Sur le plan théorique, il est vraisemblablement possible d'obtenir, comme pour le graphe k-régulier ou le modèle de Watts-Strogatz, des formulations analytiques des signaux obtenus après transformation de modèles à blocs stochastiques. Cette analyse pourrait également s'étendre vers d'autre types de structures, notamment des modèles à invariance d'échelle que l'on retrouve fréquemment dans les systèmes complexes. L'intérêt de disposer de tels résultats serait qu'il serait possible d'établir des modèles paramétriques pour les motifs fréquentiels correspondants, et d'envisager, à partir d'un motif fréquentiel donné, de chercher le modèle de graphe ou le mélange de modèles de graphe le plus proche. Enfin, l'application pour le débruitage de graphe amène évidemment à se demander quelles peuvent être les conséquences sur la structure du graphe de l'utilisation d'autres méthodes classiques de traitement du signal, que ce soit des méthodes de débruitage, comme du filtrage adaptatif, ou des décompositions des signaux par exemple. Les résultats obtenus, même s'ils ne sont limités qu'à des structures bien précises, permettent en tout cas d'apprécier le potentiel de la dualité entre réseaux et signaux.



(a) Modèle de Watts-Strogatz $k = 6$, $p = 0.1$



(b) Modèle à blocs stochastiques $C = 4$



(c) Modèle mixte de Watts-Strogatz à blocs stochastiques $C = 4$

FIGURE 3.20 – Représentation pour les trois types de graphes considérées des matrices d'adjacence des réseaux de manière à faire ressortir visuellement la structure. La figure de gauche affiche la matrice d'adjacence avant débruitage, alors que la figure de droite affiche la matrice d'adjacence après le débruitage.

CHAPITRE 4

Décomposition de réseaux temporels

Résumé –

Ce chapitre propose une extension de la dualité entre réseaux et signaux, décrite dans le Chapitre 2, aux réseaux temporels. La Section 1 décrit les réseaux temporels, et justifie l'utilisation d'outils de traitement du signal pour leur analyse. La Section 2 discute de cette extension, et la valide à travers deux exemples de réseaux temporels. La Section 3 introduit une méthode de décomposition des motifs fréquentiels obtenus à chaque pas de temps, et montre comment cette méthode permet de représenter un réseau temporel comme plusieurs sous-réseaux temporels, chacun avec une structure déterminée. Enfin, la Section 4 propose une application sur le réseau temporel obtenu à partir des données du système Vélo'v, introduit dans le Chapitre 1.

Sommaire

1 Les réseaux temporels

1.1 Temporalité dans les réseaux

Dans la nature, le temps est un paramètre fondamental dans la réalisation des processus complexes. Il est en effet difficile d'imaginer décrire un système, comprendre ses mécanismes et prédire son comportement sans prendre en compte la dynamique temporelle de ce système, et ceci quelle que soit sa nature. L'importance du temps interroge les chercheurs de toutes les disciplines [73], et s'il est complexe de le caractériser en tant que tel, sa prise en compte est primordiale dans des domaines aussi variés que l'épidémiologie, la biologie, l'économie, les télécommunications ou la musique. Les systèmes se représentant sous la forme de réseaux, c'est-à-dire comme un ensemble de relations entre des entités, n'échappent pas à cet effet du temps, et la compréhension de leur fonctionnement nécessite de considérer la dynamique de ces relations. La théorie des réseaux, dont les outils sont pour la plupart adaptés à des représentations statiques, nécessitent ainsi une extension pour considérer cette dynamique.

Dynamique des réseaux La prise en compte du temps se révèle essentielle dans l'étude des réseaux, comme en témoignent de nombreuses applications, par exemple pour étudier la vitesse de propagation d'une épidémie dans un réseau de contacts humains [226], comprendre les interactions moléculaires dans les cellules [191], étudier les interconnexions entre différents réseaux de transport [103] ou encore étudier les flux dans les réseaux de communications [127]. Ces systèmes font appel à des mécanismes complexes, pour lesquels une vision statique ou agrégée n'est pas suffisante. La théorie des graphes, qui sert de base à la théorie des réseaux, n'est pourtant pas adaptée à la prise en compte du temps : les objets mathématiques introduits, ainsi que les résultats sur ces objets, sont fondamentalement statiques, et n'intègrent pas de formalisme adapté à la temporalité. Cet écueil a tout d'abord été évité en considérant non pas une dynamique de la structure du réseau, mais des processus évoluant dessus [19]. L'idée est ainsi de caractériser en quoi la structure du réseau, décrite par la théorie des graphes, a une influence sur la diffusion d'un phénomène, que ce soit la diffusion d'une maladie dans une population [60], la propagation de rumeurs au sein d'un système social [52], ou des paquets de données dans Internet [80]. Les outils pour étudier ces aspects viennent principalement de la physique statistique, comme l'utilisation d'une équation maîtresse décrivant l'évolution temporelle du système pour chaque nœud du réseau, ou d'un modèle d'Ising sur le réseau. Les résultats mettent en évidence, à partir d'interactions locales entre les nœuds du réseau, l'émergence de phénomènes collectifs macroscopiques.

La dynamique des réseaux eux-mêmes, c'est-à-dire lorsque les connexions entre les nœuds du réseau changent au cours du temps, n'a été abordée que plus récemment. Jusque-là, l'approche retenue dans la très grande majorité des travaux a consisté à supprimer l'aspect temporel en se ramenant à un graphe statique, par exemple par agrégation des liens sur un intervalle de temps. Si elle permet déjà d'obtenir des éléments de compréhension, cette approche présente des limites, car elle ne permet pas notamment de retranscrire la séquence dynamique d'apparition et de disparition des liens : deux liens, actifs l'un après l'autre mais jamais au même moment, vont se retrouver actifs simultanément dans une représentation agrégée. De plus, les résultats obtenus sur une version agrégée du graphe peuvent ne pas se retrouver à chaque pas de temps : Braha et al. [41] font par exemple état, dans un réseau social, de personnes avec un fort degré lorsque le réseau est agrégé, mais qui ne sont à un temps donné connectées qu'avec un nombre réduit de personnes. Depuis la mise à disposition récente de données massives sous la forme de réseau temporel, les recherches sur l'étude de l'évolution de leur structure ont connu un essor important. Ces recherches se sont effectuées sous différents noms, principalement en fonction du domaine de recherche : graphes dynamiques (*dynamic graphs*) [120], graphes temporels (*temporal graphs*) [144] et donc réseaux temporels (*temporal networks*) [140], ainsi que d'autres noms difficilement traduisibles en français comme *time-evolving, time-varying or evolving graphs* [87, 96]. Ces multiples dénominations traduisent à la fois l'émergence de ce domaine qui n'en est encore qu'à ses balbutiements et dans lequel il n'existe pas encore de consensus sur la bonne approche, mais également le dynamisme de cette nouvelle

communauté, qui regroupe des équipes de recherche de divers horizons. Les différents états de l'art sur ce domaine [124, 47] illustrent également ce dynamisme.

Caractérisation de la structure des réseaux temporels De manière similaire à l'analyse des réseaux statiques, la détection de communautés reste un sujet privilégié parmi les thèmes de recherches dans la caractérisation des réseaux temporels. Alors qu'il existe déjà d'innombrables méthodes sur les réseaux statiques reposant sur de multiples définitions [97], l'ajout du temps comme un nouveau degré de liberté complique d'autant plus la recherche de communautés au sein d'un réseau temporel. Un enjeu majeur porte ainsi sur l'évolution des communautés au cours du temps. Plusieurs phénomènes peuvent ainsi survenir : il peut y avoir croissance ou au contraire contraction de la communauté, fusion de plusieurs communautés ou division d'une communauté, naissance ou disparition d'une communauté. Ces phénomènes, qui n'ont évidemment de sens qu'avec l'ajout de la temporalité, compliquent sensiblement la détection, car celle-ci suppose que les communautés sont stables au cours du temps.

Comme l'explique Cazabet [50] dans l'état de l'art qu'il dresse sur la détection de communautés dynamiques, il existe plusieurs approches. En se concentrant sur une représentation du réseau temporel comme une succession d'instantanés de graphes, c'est-à-dire qu'à chaque pas de temps correspond un graphe statique, une première approche a consisté à utiliser les méthodes classiques de détection de communautés temps par temps, et à apparier les communautés obtenues entre elles de manière à pouvoir retracer l'évolution au cours du temps de la communauté [126, 223, 245]. Aynaud [13] montre néanmoins que l'application directe de la détection de communautés à chaque pas de temps afin d'en tirer des conclusions sur des communautés temporelles n'est pas appropriée, car impliquant trop d'instabilités : les communautés évoluent sans rapport les une avec les autres sans considérer les communautés obtenues à des temps différents. Les évolutions des communautés ne sont ainsi pas dues à des changements de structures du réseau mais à des caractéristiques propres aux méthodes de détection de communautés. D'autres approches ont donc été proposées, avec l'idée de considérer les différents pas de temps du réseau temporel, que ce soit simultanément, ou successivement en prenant en compte les résultats obtenus à l'étape t lors de la détection des communautés à l'étape $t + 1$. Ces méthodes se basent ainsi soit sur des modifications du réseau afin de le rendre statique [132, 171, 244], de façon à pouvoir appliquer les méthodes de détection de communautés classiques, soit sur la modification des algorithmes eux-mêmes de manière à ce qu'ils intègrent la prise en compte des différents pas de temps [14, 15, 255]. Ces travaux, non-exhaustifs, montrent le dynamisme sur ce domaine de recherche qui, comme pour les réseaux statiques, passionne la communauté scientifique.

En dehors de la détection de communautés, d'autres travaux pour caractériser la structure des réseaux temporels ont vu le jour, notamment sur l'extension des mesures de la théorie des réseaux statiques [41, 179] au cas dynamique. La détection de motifs temporels a également été étudiée [264, 145, 74] dans le but de repérer des évènements particuliers, statiques ou dynamiques. Enfin, les problèmes de visualisation, encore plus importants lorsque le temps entre en jeu, ont également été abordés. Parmi ces travaux, les travaux de Xu et al. [253] ont la particularité d'utiliser le positionnement multidimensionnel, présenté au Chapitre 3, auquel une régularisation temporelle est ajoutée afin de représenter la structure principale du réseau temporel dans un espace euclidien.

1.2 Réseaux temporels et traitement du signal

De la même manière que la physique statistique joue un rôle important dans le développement des travaux sur la dynamique sur les réseaux, il existe des arguments qui laissent penser que le traitement du signal pourrait être considéré de façon pertinente dans l'étude des réseaux temporels, et apporter des idées novatrices par le développement de nouvelles méthodologies pour l'analyse de réseaux temporels.

Traitement du signal sur graphe dynamique Comme discuté dans le Chapitre 3, le domaine du traitement du signal sur graphe, en plus de proposer des méthodes d'analyse de signaux définis sur les

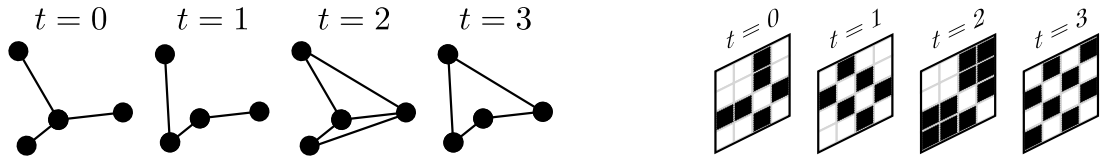
nœuds du graphe, permet de caractériser la structure des réseaux eux-mêmes. Un exemple important est la récente méthode de détection multi-échelle de communautés proposée par Tremblay [233] à l'aide d'ondelettes définies sur le graphe. De la même manière que le traitement d'images s'est naturellement étendu vers le traitement vidéo lorsqu'une composante temporelle a été prise en compte, il est tout à fait possible d'imaginer une extension du traitement du signal sur graphes statiques vers les graphes dynamiques, et donc l'apparition de nouvelles méthodes pour étudier la structure du réseau temporel basées sur le formalisme mathématique du traitement du signal. Cet objectif est néanmoins encore loin d'être rempli, en partie à cause du fait qu'on ne dispose pas encore du recul nécessaire pour envisager cette extension : le traitement du signal sur graphes est une discipline émergente qui, si elle propose déjà des avancées significatives, nécessite encore d'être étudiée intensivement avant de pouvoir envisager une prise en compte de la temporalité. Néanmoins, des premiers travaux [160] montrent qu'il est probable que ce domaine prenne beaucoup d'importance dans les années à venir.

Réseaux temporels comme signaux De façon beaucoup plus pragmatique, l'utilisation du traitement du signal pour la caractérisation des réseaux temporels est déjà envisageable, en ramenant le réseau temporel à un ensemble de grandeurs évoluant dans le temps. Une approche complète consisterait à considérer chaque graphe comme un point du signal, avec l'inconvénient que le traitement du signal n'est pas adapté pour analyser des signaux sous la forme de graphes. Au contraire, ces grandeurs peuvent être obtenues à l'aide des mesures définies dans le Chapitre 2, permettant à chaque pas de temps de caractériser un aspect de la structure du graphe, de manière à obtenir des signaux uni-dimensionnels. L'analyse de ces signaux par des techniques de traitement du signal permet ainsi de faire ressortir des mouvements périodiques ou d'étudier la stationnarité du graphe. Cette approche pose néanmoins le problème du choix des mesures qui, par nature, sont limitées pour assurer une bonne représentativité la structure du graphe. Une solution intermédiaire pour pallier les problèmes soulevés consiste ainsi à décrire le réseau non plus par un graphe ou par des mesures, mais par la collection de signaux obtenue en utilisant la méthode proposée dans le Chapitre 3. Ces signaux, qui représentent exactement la structure du réseau, peuvent ainsi se substituer au graphe et permettent l'utilisation des techniques du traitement du signal.

Cadre d'étude Cette dernière approche est l'objet de ce Chapitre 4. La Section 2 discute comment la dualité entre réseaux statiques et signaux peut être étendue dans le cas temporel. De la même manière que pour la collection de signaux représentant un réseau statique, il est possible de décrire la structure à l'aide d'analyse spectrale : une discussion sur l'observation directe de ces motifs fréquentiels temporels est d'abord réalisée puis, dans la Section 3, une méthode de décomposition des spectres est discutée à l'aide d'une factorisation en matrices non négatives, afin de pouvoir réaliser une décomposition du réseau temporel en sous-réseaux temporels. Enfin, la Section 4 propose un point de vue légèrement différent, en effectuant cette décomposition non pas sur les spectres, mais directement sur le réseau lui-même, afin de comparer les différences avec la méthode précédente.

Les travaux présentés dans ce chapitre fournissent ainsi un point d'entrée vers l'utilisation du traitement du signal pour l'analyse des réseaux temporels. Du fait du faible ancrage de la méthode de transformation dans une cadre théorique solide, il est d'autant plus compliqué d'envisager pour cette méthodologie une validation théorique suffisante. Aussi, une approche empirique est privilégiée, à travers l'illustration des outils présentés sur des réseaux temporels soit dont la structure est connue, soit dont le système est suffisamment documenté pour permettre de valider les résultats obtenus avec des comportements attendus.

1.3 Représentation d'un réseau temporel



(a) Représentation comme une séquence de graphes

(b) Représentation comme un tenseur d'adjacence $\mathcal{A}$

FIGURE 4.1 – Représentation d'un réseau temporel $\mathcal{G}$ à 4 nœuds sous la forme d'une séquence de graphes et d'un tenseur d'adjacence.

Il existe de nombreuses façons de représenter un réseau temporel, comme le soulignent Holme et al. [125]. Le réseau temporel peut ainsi être vu, comme discuté jusque-là, comme une succession d'instantanés de graphes, ou bien comme un graphe statique dans lequel une suite d'évènements affectant les liens du réseau apparaissent au cours du temps. Ces deux représentations, bien que proches, impliquent des représentations mathématiques différentes et donc des méthodes différentes, et témoignent de la difficulté de généraliser la théorie des graphes au cas temporel.

La première représentation d'un réseau temporel est retenue dans ces travaux, et peut se décrire à l'aide de séquences de contacts se produisant à des temps discrets : soit $\mathcal{G} = (\mathcal{V}, \mathcal{E}, T)$ où $\mathcal{V}$ est l'ensemble de nœuds, $\mathcal{E}$ l'ensemble de liens et T le nombre de pas de temps [1]. À la différence d'un réseau statique, l'ensemble $\mathcal{E}$ est composé de triplets composés de deux nœuds correspond aux extrémités du lien, et d'un pas de temps $t \in \{0, \dots, T-1\}$ indiquant le temps où ce lien est actif. De la même manière, l'ensemble $\mathcal{V}$ peut prendre en compte une dimension temporelle, mais dans la suite de ces travaux, les nœuds seront considérés comme présents à chaque pas de temps dans le réseau temporel.

Comme discuté plus haut, une manière équivalente pour voir cette représentation consiste à la considérer comme une séquence à temps discret de graphes (Basu et al. [20] parlent de séquences de *graphlet*) : à chaque pas de temps t, le réseau temporel se réduit à un graphe statique, que l'on notera $G^{(t)}$. De la même manière que pour les réseaux statiques, un réseau temporel peut se représenter à l'aide d'une structure d'adjacence, qui sera ici un tenseur d'adjacence $\mathcal{A} \in \mathbb{R}^{N \times N \times T}$ avec $\mathcal{A} = \{\boldsymbol{A}^{(t)}\}_{t \in \{0,\dots,T-1\}}$, où $\boldsymbol{A}^{(t)}$ est la matrice d'adjacence au temps t. La Figure 4.1 résume schématiquement un réseau temporel $\mathcal{G}$ à 4 nœuds sous la forme d'une séquence de graphes et d'un tenseur d'adjacence.

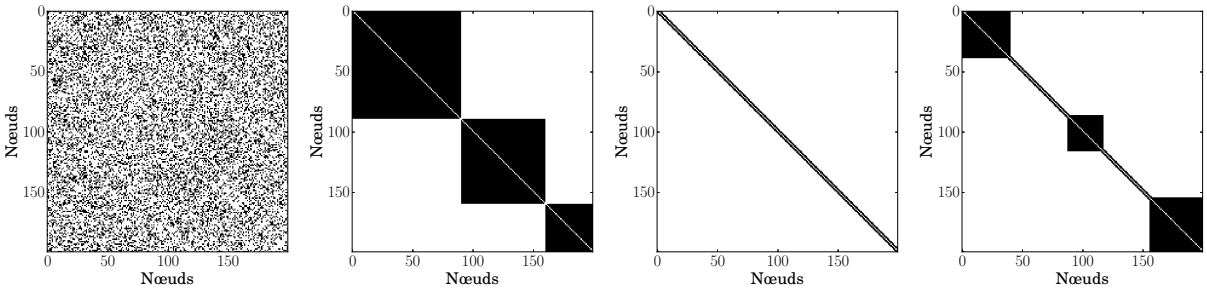
1.4 Exemples étudiés

1.4.1 Préliminaires : modèle de génération de réseaux temporels à partir de structures statiques

Il existe dans la littérature plusieurs méthodes pour obtenir un réseau temporel synthétique, de manière à pouvoir contrôler la structure au cours du temps, comme le modèle aléatoire exponentiel temporel [118] ou des modèles dynamiques à blocs stochastiques [254]. Ces modèles sont néanmoins limités à un seul type de structure, et leur combinaison peut de plus se révéler problématique. Afin de pouvoir contrôler la structure du réseau temporel au cours du temps en proposant une gamme de structures plus diverse, une méthode de génération d'un réseau temporel est proposée, de manière à construire un graphe temporel comme une succession de structures statiques. Ces structures sont actives pendant un intervalle de temps défini, et les transitions entre chaque structure sont effectuées de façon à ce que le changement de topologie du réseau ne soit pas brutal.

À chaque pas de temps, l'algorithme de construction du réseau temporel ajoute et enlève des liens en fonction de probabilités qui dépendent à la fois de la présence ou non du lien au temps précédent ainsi

1. L'utilisation de lettres rondes dénotera par la suite la temporalité de l'objet auquel il réfère.



(a) Structure 1 (b) Structure 2 (c) Structure 3 (d) Structure 4

FIGURE 4.2 – Matrice d'adjacence de chacun des quatre graphes prescrits utilisés pour la génération du réseau temporel synthétique.

que de la présence ou non du lien dans la structure prescrite. Ces probabilités sont choisies de manière à préserver une continuité temporelle de la structure du réseau mais également de faire en sorte que la structure du réseau soit très proche de la structure prescrite, à la fin de la période pendant laquelle elle est active. En choisissant plusieurs structures prescrites successives, il est ainsi possible de générer un réseau synthétique dont la structure évolue au cours du temps, tout en garantissant des transitions relativement douces entre chaque pas de temps.

Les probabilités d'ajouter un lien dans le réseau à un temps donné sont notées de la façon suivante :

— p_{11} si le lien existe dans le réseau temporel au temps précédent et dans le graphe prescrit ;
— p_{01} si le lien n'existe pas dans le réseau temporel au temps précédent mais existe dans le graphe prescrit ;
— p_{10} si le lien existe dans le réseau temporel au temps précédent mais n'existe pas dans le graphe prescrit ;
— p_{00} si le lien n'existe ni dans le réseau temporel au temps précédent ni dans le graphe prescrit.

L'Algorithme 8 donne la procédure de génération d'un réseau temporel synthétique.

Algorithme 8 : Génération d'un réseau temporel synthétique à partir de structures statiques

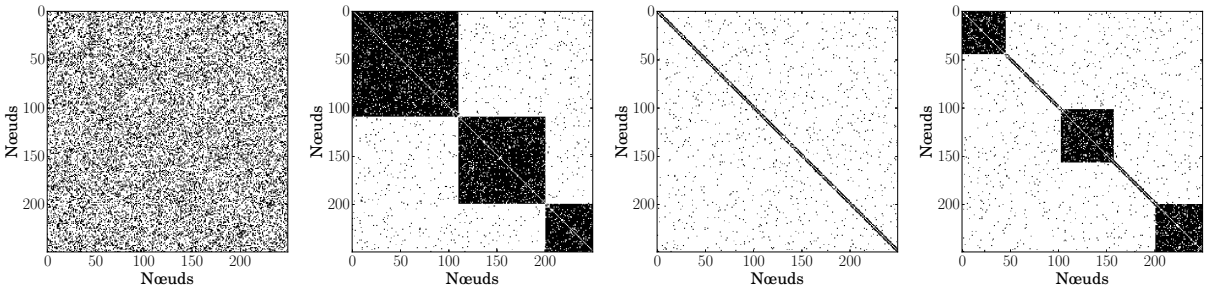
Entrées :
— $n \in \mathbb{N}^*, T \in \mathbb{R}^+$
— Une liste de graphes à n nœuds $\{G_1, G_2, \ldots, \}$ avec une liste d'intervalles associés $\{I_1, I_2, \ldots\}$ tel que $\bigcup_t I_t = \{0, \ldots, T-1\}$

Sortie : $\mathcal{G}$ un réseau temporel

début
  Définir un réseau temporel $\mathcal{G} = (\mathcal{V}, \mathcal{E}, T)$ avec $|\mathcal{V}| = n$ et $\mathcal{E} = \emptyset$
  pour $t \in \{0, \ldots, T-1\}$ **faire**
    Choisir la structure prescrite $G_k = (V_k, E_k)$ tel que $t \in I_k$, pour $k \in \{1, 2, \ldots\}$
    pour $\forall (u, v) \in V^2$ **faire**
      si $(u, v, t-1) \in \mathcal{E}$ ***et*** $(u, v) \in E_k$ **alors** Ajouter (u, v, t) à $\mathcal{E}$ avec une probabilité p_{11}
      si $(u, v, t-1) \in \mathcal{E}$ ***et*** $(u, v) \notin E_k$ **alors** Ajouter (u, v, t) à $\mathcal{E}$ avec une probabilité p_{10}
      si $(u, v, t-1) \notin \mathcal{E}$ ***et*** $(u, v) \in E_k$ **alors** Ajouter (u, v, t) à $\mathcal{E}$ avec une probabilité p_{01}
      si $(u, v, t-1) \notin \mathcal{E}$ ***et*** $(u, v) \notin E_k$ **alors** Ajouter (u, v, t) à $\mathcal{E}$ avec une probabilité p_{00}

Note : Afin de ne pas avoir un graphe déconnecté au temps $t = 0$, on suppose qu'il existe un état fictif $t = -1$ dans lequel tous les nœuds sont déconnectés.

1.4.2 Réseau temporel synthétique



(a) $t = 19$ (b) $t = 39$ (c) $t = 59$ (d) $t = 79$

FIGURE 4.3 – Matrice d'adjacence $\boldsymbol{A}^{(t)}$ du réseau temporel synthétique pour différentes valeurs de t.

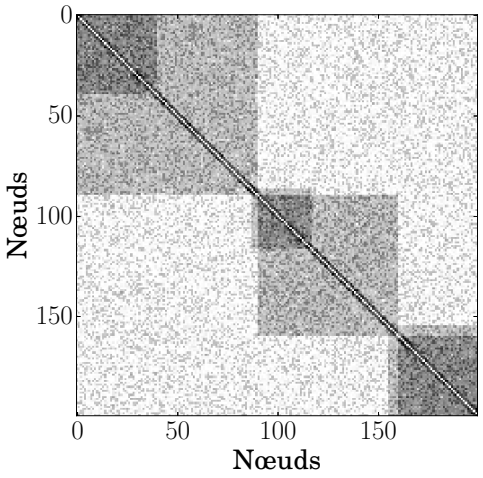


FIGURE 4.4 – Matrice d'adjacence $\boldsymbol{A}_t$ du réseau temporel synthétique agrégée sur tous les pas de temps.

Construction Un réseau temporel synthétique de 200 nœuds et 100 pas de temps est introduit, généré à partir du modèle décrit ci-dessus, afin d'illustrer la méthodologie développée dans la suite de ce chapitre. Quatre structures de 200 nœuds sont successivement choisies comme graphe prescrit :

1. pour $t \in \{0, 24\}$, un graphe aléatoire de type Erdős-Rényi avec $p = 0.4$;
2. pour $t \in \{24, 49\}$, un graphe avec 3 cliques déconnectés ;
3. pour $t \in \{50, 74\}$, un graphe 4-régulier ;
4. pour $t \in \{75, 99\}$, un graphe 4-régulier avec 3 cliques.

Afin de créer des transitions douces entre ces structures, les probabilités d'apparition d'un lien dans le réseau temporel au temps suivant sont fixées de la façon suivante : $p_{11} = 0.01$, $p_{01} = 0.5$, $p_{11} = 0.4$ et $p_{00} = 0.01$. La Figure 4.2 affiche les matrices d'adjacence [2] de chacun des quatre graphes prescrits utilisés pour la génération du modèle.

Description La Figure 4.3 affiche des instantanés du réseau temporel à la fin de chaque intervalle de temps pendant lesquels une structure est active, sous la forme de matrices d'adjacence. Les structures obtenues sont très proches de celles prescrites par le modèle, visibles à la Figure 4.2 : les blocs décrivent les communautés, alors que les diagonales représentent le graphe k-régulier. Néanmoins, contrairement

2. Les nœuds du graphe sont ordonnés de façon à faire ressortir les structures par la visualisation des matrices d'adjacences. Cet ordonnancement, qui est l'objet du Chapitre 2, est implicite par la suite.

aux matrices d'adjacence des graphes prescrits, ces structures sont bruitées, ce qui est visible par la présence de points noirs dans les zones blanches dénotant les liens présents dans le réseau temporel mais pas dans la structure prescrite, et de points blancs dans les zones blanches, dénotant les liens absents dans le réseau temporel mais présents dans la structure prescrite. La Figure 4.4 montre la matrice d'adjacence du réseau temporel agrégée sur les 100 pas de temps : les quatre types de structures apparaissent également avec cette représentation, qui permet une visualisation simple du réseau temporel, mais sans néanmoins permettre de distinguer la temporalité de ces structures.

1.4.3 Réseau temporel des interactions sociales dans une école primaire

Construction Le réseau temporel des interactions sociales dans une école primaire décrit les interactions entre des enfants d'une école primaire pendant deux jours en octobre 2009. Cette expérience, réalisée dans le cadre du projet SocioPatterns [224], à consister à enregistrer les contacts entre les individus à l'aide de capteurs RFID (Radio Frequency IDentification). Pendant des intervalles de temps de 20 secondes, un lien existe entre deux individus si un contact est enregistré, c'est-à-dire si les personnes se tiennent face-à-face à moins d'un mètre de distance.

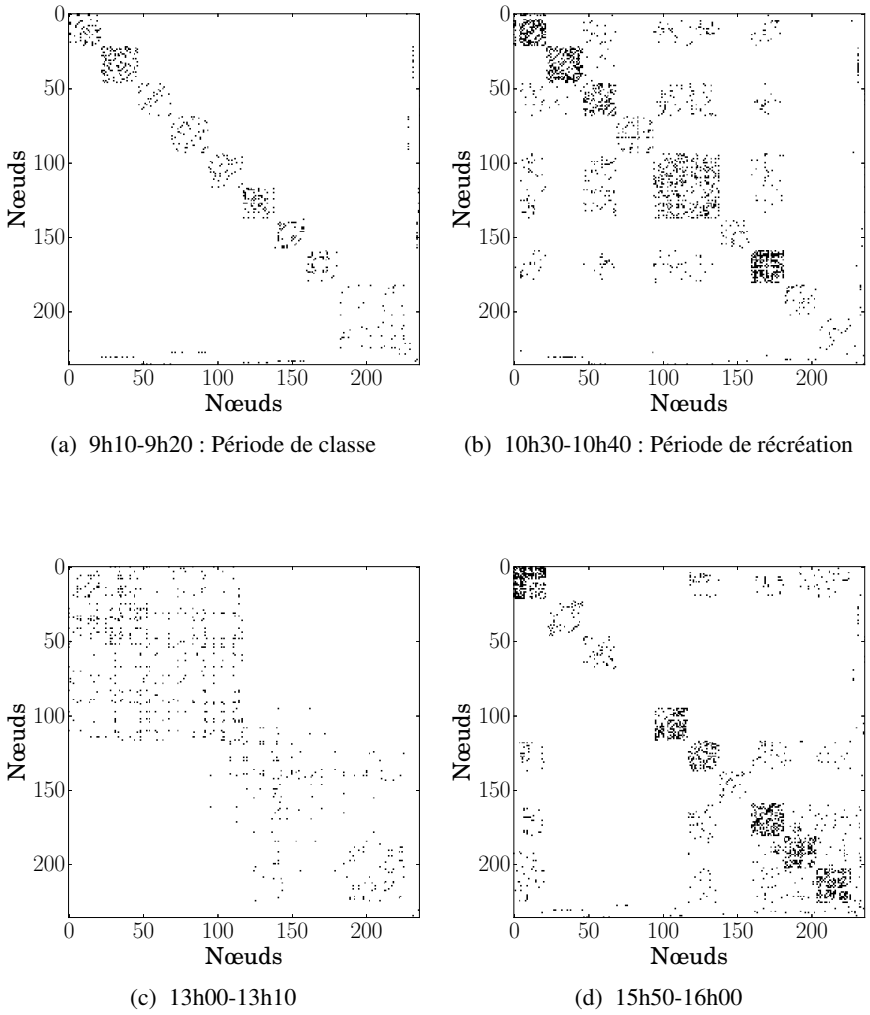
Ces données ont été principalement étudiées sous la forme de réseau agrégé dans [224]. Pour notre étude, nous proposons une représentation de ces données sous la forme d'un réseau temporel, en se limitant à la première journée de l'expérience. La journée d'école, qui commence à 8h30 et finit à 16h30, est découpée en périodes de 10 minutes. Pour chaque intervalle de temps, un réseau statique est défini, en ajoutant un lien entre deux individus s'ils ont eu au moins un contact face-à-face durant l'intervalle. 226 enfants et 10 enseignants ont participé à l'expérience, répartis en 5 niveaux (du CP au CM2), eux-mêmes décomposés en 2 classes. Le réseau temporel obtenu est composé de 236 nœuds, évoluant sur 48 pas de temps.

Description La Figure 4.5 affiche des instantanés du réseau temporel pendant différentes périodes caractéristiques de l'activité scolaire. Les nœuds du graphe sont numérotés en suivant la classe des élèves correspondants, de la plus petite section à la grande section. Les enseignants apparaissent à la suite.

Sur la Figure 4.5a, correspondant à une période où les élèves sont répartis en classes, la distinction est clairement visible à travers une structure par blocs, correspondant aux communautés. Cette structure perdure sur la Figure 4.5b au moment de la pause, mais avec un bloc comprenant la moitié des élèves, alors que la deuxième moitié reste répartie par classe. Cette structure s'explique par le fait que l'école n'est pas assez grande pour accueillir tous les élèves en pause simultanément : la récréation se fait ainsi en deux fois. Cette rotation continue pendant la pause (Figure 4.5c) pour des raisons similaires, avec cependant des répartitions de groupe différentes. Enfin, un autre type de structure en blocs apparaît pour la pause de l'après-midi (Figure 4.5d), où l'on note également la disparition de la 4^{e}, pour des raisons inconnues.

Cette description n'est évidemment pas exhaustive, et repose uniquement sur la visualisation des matrices d'adjacence à des temps arbitraires. L'évolution de la structure du réseau est inconnue, à l'exception des indices fournis par l'emploi du temps en vigueur dans l'école primaire.

Les Figures 4.6 et 4.7 montre respectivement la matrice d'adjacence du réseau temporel agrégée sur les 48 pas de temps, et la représentation sous la forme d'un graphe. Dans cette dernière figure, les nœuds colorés représentent les élèves, et la couleur des nœuds leur classe. Les nœuds noirs sont les enseignants de chaque classe. La position des nœuds est fournie dans les données, et reflète la proximité entre les nœuds. La structuration en communautés correspondant aux classes est nettement visible.



(a) 9h10-9h20 : Période de classe

(b) 10h30-10h40 : Période de récréation

(c) 13h00-13h10

(d) 15h50-16h00

FIGURE 4.5 – Matrice d'adjacence $\boldsymbol{A}^{(t)}$ du réseau temporel synthétique pour différentes valeurs de t.

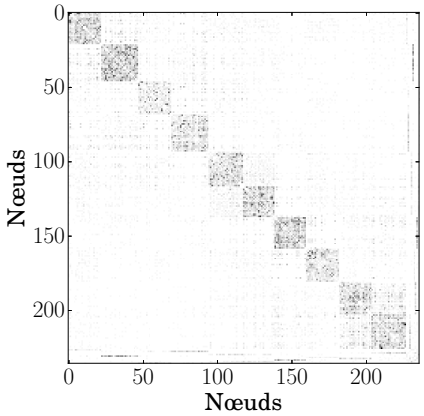


FIGURE 4.6 – Matrice d'adjacence $\boldsymbol{A}_t$ du réseau temporel des interactions sociales dans une école primaire agrégée sur tous les pas de temps.

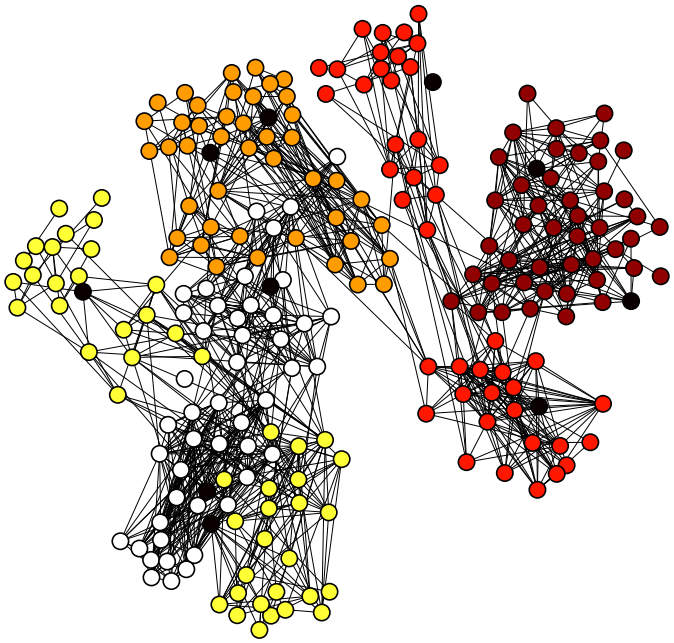


FIGURE 4.7 – Représentation du réseau temporel agrégé sur tous les pas de temps. les nœuds colorés représentent les élèves, et la couleur des nœuds leur classe. Les nœuds noirs sont les enseignants de chaque classe. La position des nœuds est fournie dans les données, et reflète la proximité entre les nœuds.

2 Extension de la dualité entre graphes et signaux aux réseaux temporels

2.1 Transformation de réseaux temporels en signaux

La transformation de réseaux en collection de signaux discutée dans le Chapitre 3 est directement extensible au cas temporel en considérant chaque pas de temps indépendamment les uns des autres.

On note la collection temporelle de signaux obtenue à partir du réseau temporel $\mathcal{G}$ par $\mathcal{X} \in \mathbb{R}^{N \times C \times T}$ avec $\mathcal{X} = \{\boldsymbol{X}^{(t)}\}_{t \in \{0,\dots,T-1\}}$, où $\boldsymbol{X}^{(t)}$ est la collection de signaux obtenue pour chaque pas de temps t :

$$\boldsymbol{X}^{(t)} = \boldsymbol{\mathcal{T}}[\boldsymbol{A}^{(t)}] \tag{4.1}$$

Comme le nombre de nœuds est invariant dans le temps par hypothèse, le nombre de composantes C et donc le nombre de fréquences F sont constants dans le temps [3]. À chaque pas de temps, l'indexation des nœuds du graphe est réalisée suivant l'algorithme d'étiquetage discutée dans le Chapitre 2, sans considérer les différents étiquetages obtenus [4].

Réciproquement, la transformation inverse est réalisée de la même manière, en appliquant la transformation inverse d'une collection de signaux vers un graphe statique, et ce pour chaque pas de temps t. En notant $\mathcal{G}^{(r)} = \{G^{(r,t)}\}_{t \in \{0,\dots,T-1\}}$ le réseau temporel reconstruit, décrit par le tenseur d'adjacence $\mathcal{A}^{(r)} = \{\boldsymbol{A}^{(r,t)}\}_{t \in \{0,\dots,T-1\}}$, cela revient à calculer :

$$\boldsymbol{A}^{(t,r)} = \boldsymbol{\mathcal{T}}^{-1}[\boldsymbol{X}^{(t)}] \tag{4.2}$$

où $\boldsymbol{A}^{(t,r)}$ représente la matrice d'adjacence du réseau temporel reconstruit au temps t.

3. Le cas où l'ensemble des nœuds $\mathcal{N}$ du graphe évolue au cours du temps est néanmoins simple à considérer, puisqu'il suffit de construire le tenseur en fixant le nombre de composantes comme étant le nombre maximal de composantes, et en ajoutant des zéros pour les pas de temps pendant lesquels le nombre de composantes est inférieur.

4. L'objectif ici est de suivre comment la structure globale du réseau temporal évolue, sans prendre en compte l'étiquette des nœuds, et l'ordre obtenu n'est pas pertinent à regarder. Néanmoins, l'étude de cet ordre donne des informations sur l'évolution d'un nœud individuel au sein de la structure globale.

2.2 Analyse spectrale des réseaux temporels

Comme précédemment, l'extension de l'analyse spectrale de signaux représentant un réseau, définis dans la Section 2.5 du Chapitre 3, est simplement réalisée en considérant indépendamment chaque pas de temps. Un tenseur des spectres temporels, noté $\mathcal{S} \in \mathbb{R}^{C\times F\times T}$, est défini, pour lequel chaque matrice $\boldsymbol{S}^{(t)}$ correspond au spectre obtenu au temps t :

$$\boldsymbol{S}^{(t)} = \boldsymbol{F}\boldsymbol{X}^{(t)} \tag{4.3}$$

avec $\boldsymbol{F}$ la matrice de Fourier, appliquée sur chaque colonne de $\boldsymbol{X}^{(t)}$.

À partir de $\mathcal{S}$, que l'on nommera dorénavant spectres temporels par analogie avec le réseau temporel, les équivalents temporels de l'énergie et des amplitudes s'obtiennent de façon identique :

— $\mathcal{M} \in \mathbb{R}^{C\times F\times T}$ représente le tenseur des amplitudes temporelles, c'est-à-dire $\boldsymbol{M}^{(t)} = |\boldsymbol{S}^{(t)}|$;

— $\mathcal{Z} \in \mathbb{R}^{C\times F\times T}$ représente le tenseur des énergies temporelles.

Visualisation des spectres temporels Comme décrit dans le Chapitre 3, le spectre est intimement relié à la structure du réseau sous-jacent. Ainsi, l'observation des énergies ou des amplitudes temporelles permet d'avoir des indices sur l'évolution de la structure du réseau temporel au cours du temps. Cette observation est réalisée par la suite en regardant les marginales des tenseurs, moyennés soit sur les fréquences, soit sur les composantes. Par la suite, nous nous concentrons sur les marginales des énergies temporelles :

— $\boldsymbol{Z}_c$ désigne la marginale de $\mathcal{Z}$ sur les composantes, c'est-à-dire :

$$\boldsymbol{Z}_c = \frac{1}{C}\sum_{c=1}^{C} \mathcal{E}_{c..} \tag{4.4}$$

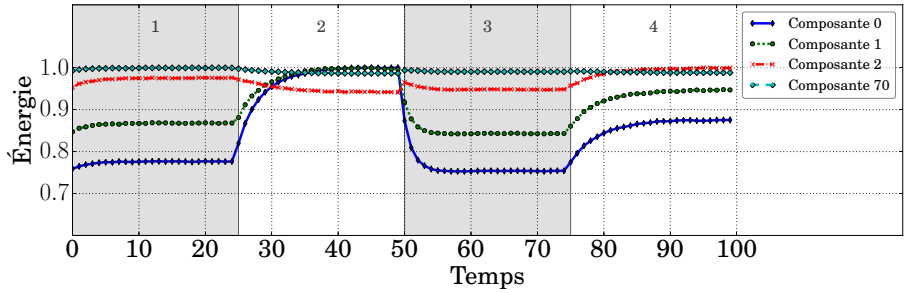
— $\boldsymbol{Z}_c(f)$ désigne la marginale de $\mathcal{Z}$ sur les fréquences, c'est-à-dire :

$$\boldsymbol{Z}_f = \frac{1}{F}\sum_{c=1}^{C} \mathcal{E}_{.f.} \tag{4.5}$$

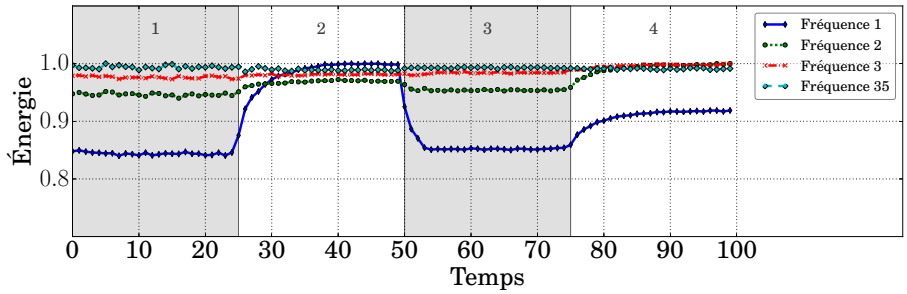
Cette représentation permet de visualiser simplement $\mathcal{Z}$, en ciblant soit des composantes, soit des fréquences particulières.

Corrélation entre spectres L'analyse spectrale peut également être utilisée afin de comparer le réseau temporel à un pas de temps fixé avec un réseau statique, dont la structure est contrôlée par exemple par un modèle de graphe. Pour cela, un coefficient de corrélation entre les énergies à chaque pas de temps est calculé. Si l'on note $\boldsymbol{Z}_m$ les énergies des signaux obtenues après transformation d'une instance $\mathcal{G}_m$ d'un modèle de graphe donné, on peut calculer un coefficient de corrélation $\rho_n^{(t)}$ entre $\boldsymbol{Z}_m$ et $\boldsymbol{Z}^{(t)}$. L'idée est ainsi de trouver parmi un ensemble de modèle paramétrique de graphe celui qui correspond au mieux au réseau à chaque pas de temps. En générant plusieurs instances du modèle pour une valeur des paramètres donnée, on peut obtenir une valeur moyenne du coefficient de corrélation.

Cette approche est testée dans les sections suivantes sur deux types de structures. Le premier modèle de graphe utilisé est un modèle à blocs stochastiques dans lequel le nombre de communautés C varie : en réglant p_{inter} à 0 et p_{intra} à 1, on obtient un réseau avec C cliques sans aucun lien entre elles. Le deuxième modèle de graphe étudié est un modèle de Watts-Strogatz, qui correspond à un graphe k-régulier en anneau, auquel plusieurs liens sont ajoutés en fixant p à 0.1 et en faisant varier le degré k.



(a) Marginales sur les fréquences $\boldsymbol{Z}_f$ pour les composantes 0, 1, 2, et 70.



(b) Marginales sur les composantes $\boldsymbol{Z}_c$ pour les fréquences 1, 2, 3 et 35.

FIGURE 4.8 – Marginales des énergies des spectres temporels du réseau temporel synthétique. Chaque courbe est normalisée par la valeur maximale. L'alternance de régions blanches et grises représentent l'alternance des structures.

2.3 Illustrations sur deux exemples

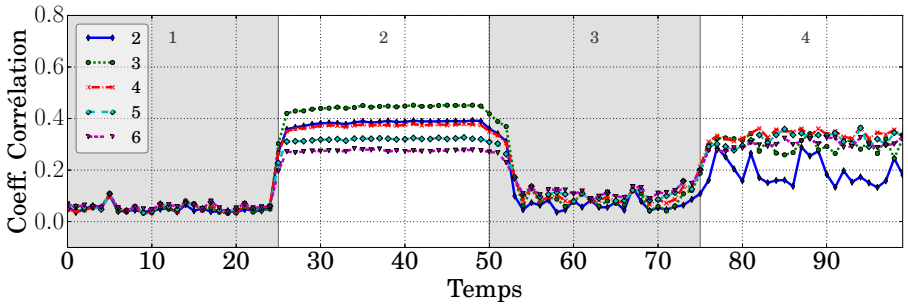
2.3.1 Réseau temporel synthétique

Le réseau temporel synthétique, décrit dans la Section 1.4, est transformé en une collection temporelle de signaux, elle-même étudiée à travers une analyse spectrale. Les énergies temporelles $\mathcal{Z}$ résultantes sont étudiées par la suite, à travers la visualisation des marginales et par l'étude des corrélations entre ces énergies et les énergies obtenues avec un modèle standard de graphes.

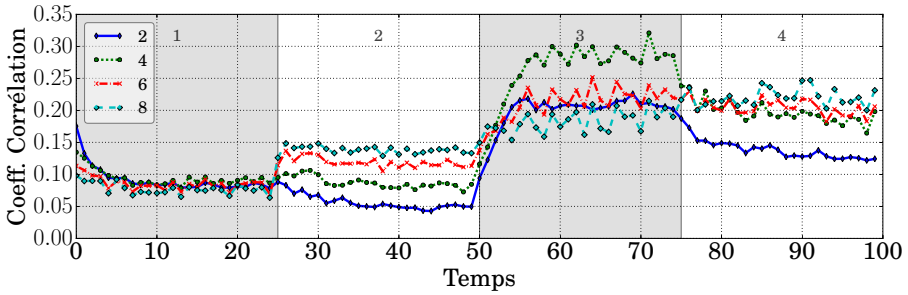
Visualisation des spectres temporels La Figure 4.8 montre $\boldsymbol{Z}_f$ pour les composantes 0, 1, 2 et 70 ainsi que $\boldsymbol{Z}_c$ pour les fréquences 1, 2, 3 et 35. L'alternance de régions blanches et grises représentent l'alternance des structures. Chaque courbe est normalisée par la valeur maximale, afin de mettre en valeur leur évolution au cours du temps. Ces deux figures révèlent la prédominance des premières composantes et des basses fréquences pour le suivi de l'organisation en communautés du réseau : dans ces zones, l'énergie est plus forte, comme discuté dans le Chapitre 3. Le suivi des autres types de structures est néanmoins plus compliqué en utilisant cette représentation.

Corrélation entre spectres Une étude sur les marginales des énergies temporelles est réalisée, en regardant leur corrélation avec l'énergie obtenue après transformation d'un réseau généré à l'aide d'un modèle de graphes.

Pour le premier modèle de graphe utilisé avec une structure en communautés, C varie de 1 à 5. La Figure 4.9a montre la corrélation moyenne obtenue après 20 répétitions pour chaque valeur de C : comme attendu, la corrélation est maximale lorsque les structures 2 et 4 sont actives, c'est-à-dire les intervalles de temps où le réseau temporel est effectivement structuré en communautés. Lorsque le réseau est uniquement structuré en communautés (structure 2), la valeur de C pour laquelle la corrélation est maximale



(a) Comparaison avec un réseau avec k communautés, pour différentes valeurs de k.



(b) Comparaison avec un graphe k-régulier en anneau, pour différente valeur de k.

FIGURE 4.9 – Corrélation entre les marginales des énergies temporelles $\mathcal{Z}$ du réseau temporel synthétique, et l'énergie obtenue après transformation d'un réseau avec une structure connue. Deux types de structures sont étudiées. La comparaison est réalisée à chaque pas de temps et pour 20 instances de graphes. Le coefficient de corrélation affiché est la moyenne obtenue sur les répétitions.

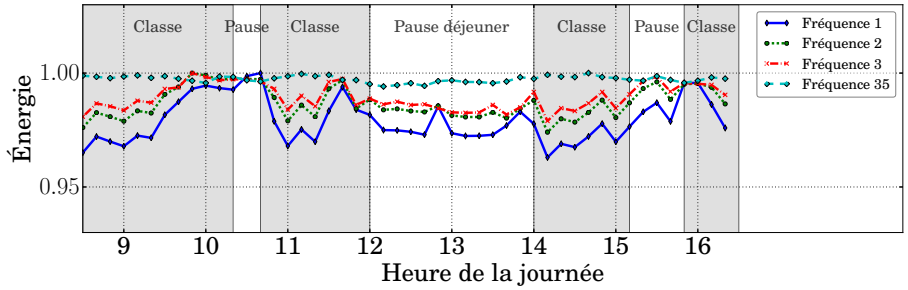
est 3, ce qui correspond au nombre effectif de communautés. Lorsque le réseau n'est pas uniquement structuré en communautés mais contient également des parties régulières, le nombre de communautés est moins clair : cela s'explique par le fait que les parties régulières entre chaque communauté peuvent être considérées comme des communautés en elle-mêmes.

Pour le deuxième modèle de graphe étudié, k prend pour valeurs 2, 4, 6 ou 8. La Figure 4.9b montre les corrélations obtenues, qui sont plus faibles que dans le cas précédent. Il est néanmoins possible de noter qu'elle augmente significativement lorsque la structure 3, un graphe régulier en anneau, est active, et en particulier le maximum de la corrélation est atteint pour $k = 4$, qui correspond bien à la structure prescrite.

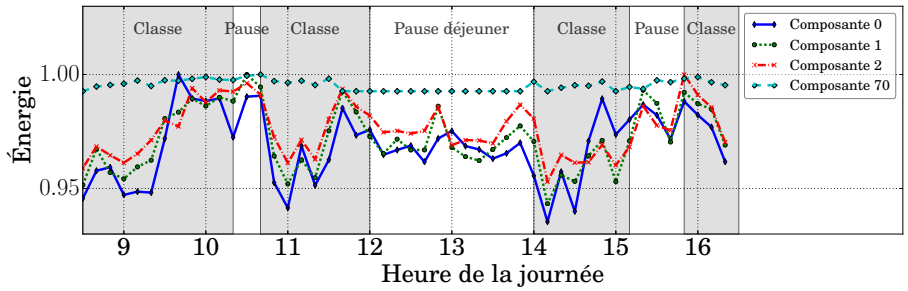
2.3.2 Réseau temporel des interactions sociales dans une école primaire

De même que pour le réseau temporel synthétique, le réseau temporel des interactions sociales dans une école primaire est transformé en une collection temporelle de signaux, elle-même étudiée à travers une analyse spectrale. Les énergies temporelles $\mathcal{Z}$ résultantes sont étudiées par la suite, à travers la visualisation des marginales et par l'étude des corrélations entre ces énergies et les énergies obtenues avec un modèle standard de graphes.

Visualisation des spectres temporels La Figure 4.10 montre les marginales des énergies temporelles. Les régions grisées correspondent aux périodes de classes, alors que les régions blanches correspondent aux pauses et à la pause de midi, selon les informations données dans [224]. Les énergies associées aux basses fréquences pour les premières composantes n'est pas réparties équitablement au cours du temps : cela indique des changements dans la structure globale du réseau temporel. Ces changements arrivent principalement de 9h30 à 10h40, incluant la récréation et la période la précédant, ainsi que pendant la



(a) Marginales sur les fréquences $\boldsymbol{Z}_f$ pour les composantes 0, 1, 2, et 70.



(b) Marginales sur les composantes $\boldsymbol{Z}_c$ pour les fréquences 1, 2, 3 et 35.

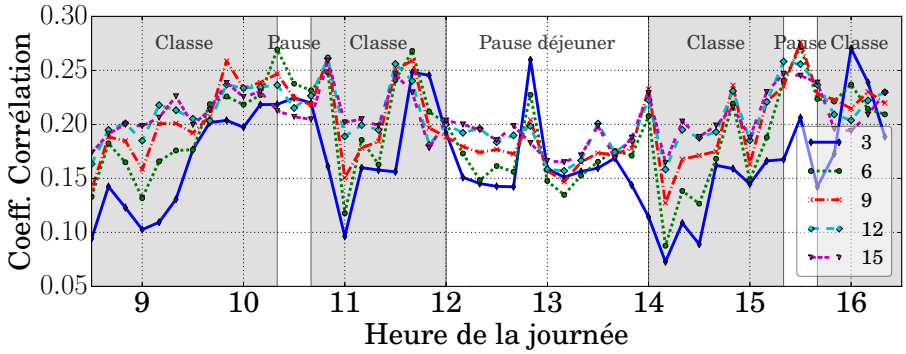
FIGURE 4.10 – Marginales des énergies des spectres temporels du réseau temporel des interactions sociales dans une école primaire. Les régions grisées correspondent aux périodes de classes, alors que les régions blanches correspondent aux pauses et à la pause de midi, selon les informations données dans [224].

pause déjeuner et la pause de l'après-midi. La journée peut ainsi être divisée en deux périodes principales : les périodes de classe et les périodes de pause, avec des différences avec les périodes théoriques pendant lesquelles ces évènements arrivent.

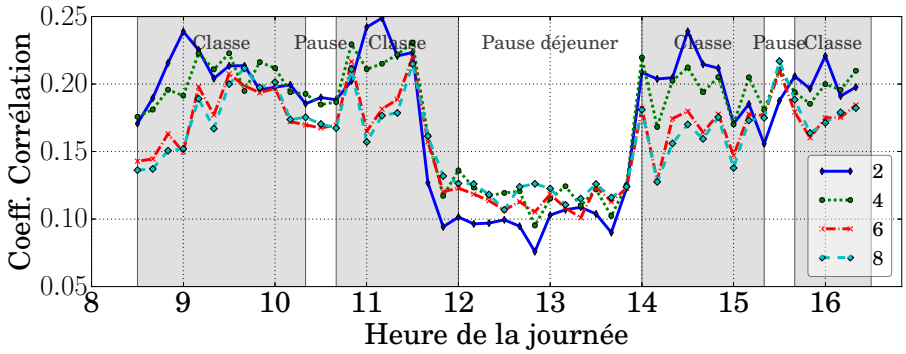
Corrélation entre spectres La comparaison du réseau temporel des interactions sociales dans une école primaire avec un réseau avec des communautés, en utilisant la corrélation entre les énergies temporelles, montre sur la Figure 4.11 que le réseau temporel n'est pas structuré en communautés distinctes, et ceci pour tous les pas de temps. Des corrélations importantes émergent néanmoins, que ce soit avec des réseaux ayant un nombre restreint de communautés (entre 3 et 6) pendant les périodes de pauses et de déjeuner, ou avec un nombre élevé de communautés (entre 9 et 15) pendant les périodes de classes. Ces observations sont cohérentes avec les informations données dans [224], qui expliquent que les classes sont séparées dans plusieurs salles de classes, alors que pendant les pauses, les classes sont mélangées, avec cependant une séparation en deux entre les classes du CP au CE2 et les classes du CM1 au CM2.

2.4 Discussions

L'étude des spectres temporels donne des indices à propos de la structure du réseau sous-jacent, même si cette approche est limitée par l'absence, dans la plupart des cas, de connaissance sur le modèle de graphe paramétrique adéquat. Dans la section suivante, nous proposons une méthode pour trouver automatiquement les structures qui caractérisent le mieux le réseau temporel, ainsi que les intervalles de temps dans lesquels la structure apparaît effectivement.



(a) Comparaison avec un réseau avec k communautés, pour différentes valeurs de k.



(b) Comparaison avec un graphe k-régulier en anneau, pour différente valeur de k.

FIGURE 4.11 – Corrélation entre les marginales des énergies temporelles $\mathcal{Z}$ du réseau temporel des interactions sociales dans une école primaire, et l'énergie obtenue après transformation d'un réseau avec une structure connue. Deux types de structures sont étudiées. La comparaison est réalisée à chaque pas de temps et pour 20 instances de graphes. Le coefficient de corrélation affichée est la moyenne obtenue sur les répétitions.

3 Décomposition de réseau temporel dans le domaine des signaux

3.1 Factorisation en matrices non-négatives (NMF)

La factorisation en matrices non-négatives (*nonnegative matrix factorization* ou NMF) est une technique permettant de décomposer une matrice de données $\boldsymbol{V}$ de dimension $F \times N$ à valeurs non-négatives, en un produit de deux matrices également non-négatives $\boldsymbol{W}$ et $\boldsymbol{H}$ de dimension respectives $F \times K$ et $K \times N$. La NMF permet une réduction de la dimensionnalité des données, en extrayant de façon non supervisée K motifs caractéristiques.

La contrainte de non-négativité des valeurs de $\boldsymbol{W}$ et $\boldsymbol{H}$ induit une représentation "par parties" qui permet d'extraire dans les colonnes de $\boldsymbol{W}$ des motifs caractéristiques des données et dans les lignes de la matrice $\boldsymbol{H}$ les coefficients d'activation de chacun des motifs pour chaque intervalle de temps. Une approche usuelle pour résoudre ce problème consiste en la résolution d'un problème d'optimisation :

$$(\boldsymbol{W}^*, \boldsymbol{H}^*) = \arg\min_{\boldsymbol{WH}} D(\boldsymbol{V}|\boldsymbol{WH})$$

où D est une mesure de dissemblance. S'il n'existe pas de solution analytique à ce problème, un algorithme d'optimisation alterné a été proposé [90] pour le cas où D est la β-divergence, dont les cas particuliers $\beta = 0$, $\beta = 1$ et $\beta = 2$ $\beta = 1$ correspondent respectivement à la divergence d'Itakura-Saito, la divergence de Kullback-Leibler et la distance Euclidienne.

Cadre d'étude dans le cas de la divergence d'Itakura-Saito La NMF est mise en œuvre afin de trouver des motifs dans les spectres des collections de signaux, obtenues après transformation du réseau temporel. Une méthodologie est mise en place par analogie avec l'analyse musicale : de la même manière qu'un signal audio peut être décomposé en plusieurs signaux audio, de façon par exemple à séparer la voix des parties instrumentales dans une chanson [89], l'objectif ici est de réaliser une décomposition du réseau temporel en sous-réseaux temporels, décomposant à chaque pas de temps la structure globale en plusieurs sous-structures. De plus, les spectres audio partagent des caractéristiques communes avec les spectres issues de graphes. Une approche similaire est ainsi proposée, en utilisant la divergence d'Itakura-Saito comme mesure, donnée par :

$$d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1 \tag{4.6}$$

Choisir cette mesure permet d'utiliser les outils développés pour la reconstruction des signaux à partir des spectres, ainsi que l'utilisation des algorithmes d'optimisation efficaces avec la possibilité d'ajouter de la régularisation temporelle [88].

La NMF retourne deux matrices $\boldsymbol{W}$ et $\boldsymbol{H}$: chaque colonne de $\boldsymbol{W}$ représente le k^{e} motif fréquentiel (normalisé), tandis que la k^{e} colonne de $\boldsymbol{H}^T$ donne les coefficients d'activation du motif fréquentiel k à chaque pas de temps. Comme décrit dans [89], utiliser la NMF avec la divergence d'Itakura-Saito permet d'avoir à disposition des moyens de reconstruction de la collection de signaux correspondant à chaque motif. Pour chaque motif k, la valeur reconstruite de ce motif est obtenue en utilisant un filtrage de Wiener, tel que les éléments $s_{cf}^{(k,t)}$ sont donnés par :

$$s_{cf}^{(k,t)} = \frac{w_{(cf)k}h_{kt}}{\sum_{l=1}^{K} w_{(cf)l}h_{lt}} s_{cf} \tag{4.7}$$

menant à la décomposition du tenseur $\mathcal{S}$:

$$\mathcal{S} = \sum_{k=1}^{K} \mathcal{S}^{(k)} \tag{4.8}$$

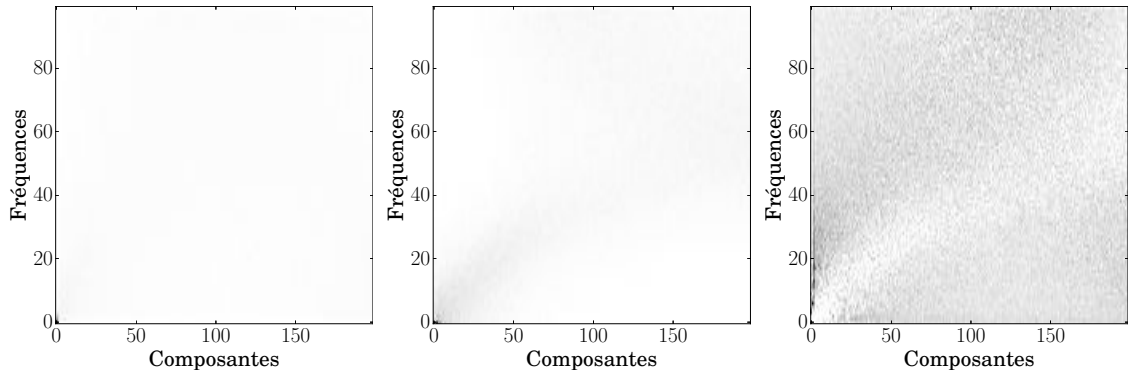
Le spectre dynamique du motif k est ainsi une fraction du spectre dynamique original. À partir de $\mathcal{S}^{(k)}$, une transformée de Fourier inverse est réalisée, menant à une collection de signaux pour chaque motif k, notée $\mathcal{X}^{(k)} \in \mathbb{R}^{N\times N\times T}$. Finalement, le tenseur d'adjacence $\mathcal{A}^{(k)}$ décrivant le réseau temporel correspondant au motif k est obtenu en utilisant la transformation inverse $\boldsymbol{\mathcal{T}^{-1}}$ décrite dans la Section 2.

de

3.2 Décomposition des spectres représentant le graphe

Comme l'entrée dans notre cas est le réseau temporel $\mathcal{S}$, représenté comme un tenseur de dimension $C \times F \times T$, une légère adaptation doit être réalisée avant d'appliquer la NMF. À chaque pas de temps t, les spectres $\boldsymbol{S}_t$ sont représentés par un vecteur $\boldsymbol{v}_t$ en ajoutant successivement bout-à-bout les colonnes de la matrice $\boldsymbol{S}_t$. Pour tout $t \in \{0, \ldots, T-1\}$, ces vecteurs composent les colonnes de la matrice $\boldsymbol{V}$, de dimension $(FC)\times T$. Le nombre de motifs K est fixé en adéquation avec les attentes sur les données, et le paramètre γ est choisi de manière empirique, de manière à assurer le lissage des coefficients d'activation.

À partir de chaque colonne $\boldsymbol{w}_k$ de $\boldsymbol{W}$, une carte composante-fréquence est reconstruite en retaillant une matrice à partir du vecteur. Afin de mettre en évidence à quoi ressemblent les sous-structures dans une représentation de graphe traditionnelle, chaque motif est transformé en réseau temporel.



(a) Motif fréquentiel 1 (b) Motif fréquentiel 2 (c) Motif fréquentiel 3

FIGURE 4.12 – Motifs fréquentiels obtenus après application de la NMF sur le réseau temporel synthétique. Les colonnes de $\boldsymbol{W}$ sont retaillées en matrice.

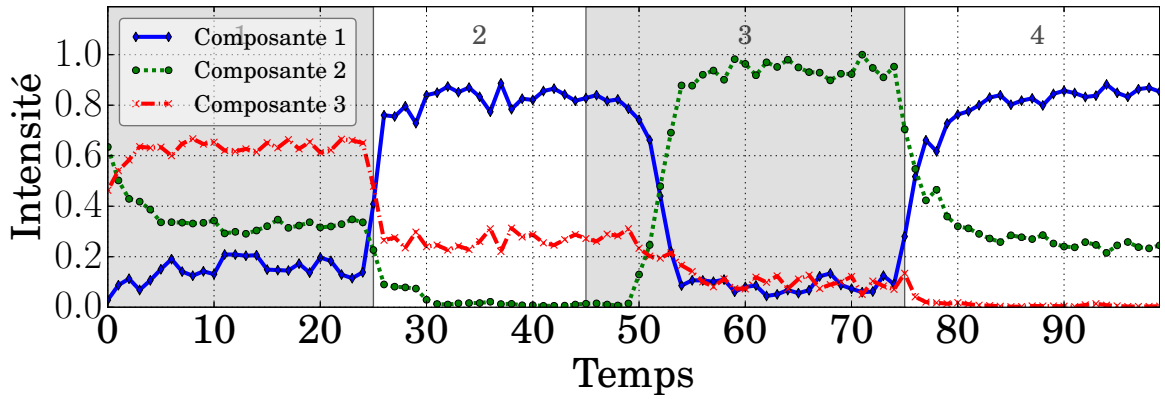


FIGURE 4.13 – Coefficients d'activation pour chacun des motifs obtenus à la Figure 4.12, correspondant aux lignes de la matrice $\boldsymbol{H}$. Les valeurs sont normalisées par la valeur maximale de $\boldsymbol{H}$.

Reconstruction des sous-réseaux temporels Dans les illustrations proposées par la suite, le réseau temporel est décrit en affichant la matrice d'adjacence agrégée sur les temps, de manière à faire ressortir les structures les plus significatives. Pour prendre en compte les coefficients d'activation, la matrice d'adjacence agrégée est calculée de la façon suivante pour le motif k :

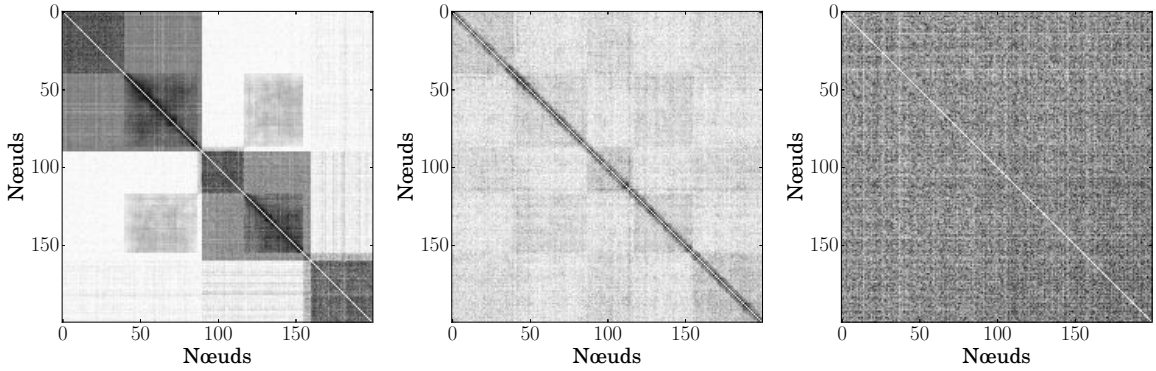
$$\boldsymbol{A}^{(k)} = \sum_{t=1}^{T} h_{kt} \boldsymbol{A}^{(k,t)} \tag{4.9}$$

avec h_{kt} désignant les éléments de $\boldsymbol{H}$.

3.3 Application au réseau synthétique

La décomposition en sous-structures temporelles est d'abord appliquée au réseau temporel synthétique. La matrice $\boldsymbol{V}$, de dimension 19900×100 est décomposée en utilisant $K = 3$ et $\gamma = 5$.

La Figure 4.13 montre les coefficients d'activation obtenus pour chacun des motifs. La première observation est que ces coefficients sont, sans surprise, en cohérence avec le découpage en quatre structures du réseau temporel synthétique. Ces coefficients ont également différents niveaux d'intensité au cours du temps. Le motif 1 est par exemple très actif pendant la période 4 et surtout la période 3, et à une intensité relativement faible pendant les périodes 1 et 2. Le motif 2 quant à lui est très actif pendant la période 1,



(a) Sous-réseau temporel 1 (b) Sous-réseau temporel 2 (c) Sous-réseau temporel 3

FIGURE 4.14 – Matrice d'adjacence agrégée sur tous les pas de temps du réseau temporel obtenue après transformation de la collection temporelle de signaux issue des motifs fréquentiels.

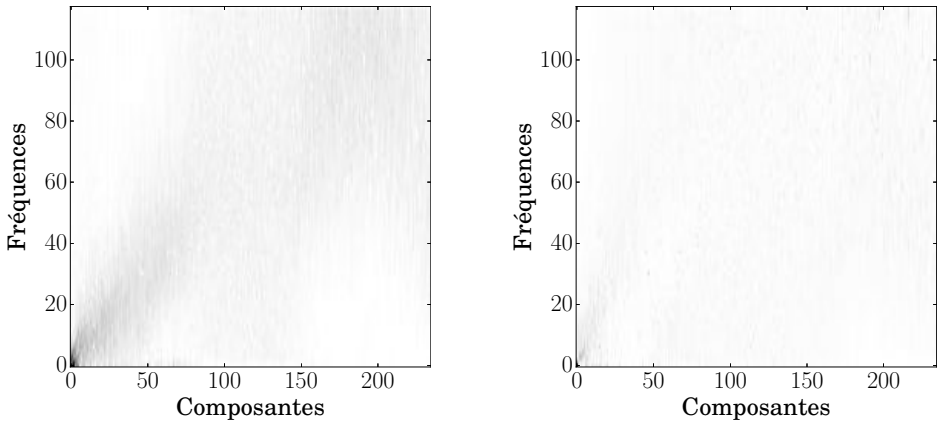
2 et 4, et inactif pendant la période 3. Finalement, le motif 3 est surtout actif pendant la période 1, ainsi que dans une moindre mesure pendant la période 2.

À la lumière des motifs fréquentiels visibles à la Figure 4.12, il est possible d'avoir des idées sur la structure du réseau temporel en établissant des parallèles entre ces motifs temporels et ceux observés dans la Section 3 du Chapitre 3 pour différents modèles de graphe. Le premier motif fréquentiel ressemble, bien que plus faiblement, au motif fréquentiel d'un graphe 4-régulier en anneau, alors que le motif fréquentiel 2 révèle une énergie concentrée sur les premières composantes pour les basses fréquences, caractéristique d'une structure en communauté. Enfin, le motif 3 est beaucoup moins structuré que les autres, ce qui le rapproche d'une structure aléatoire. Afin de confirmer ces intuitions, une reconstruction des sous-structures à partir des motifs fréquentiels est réalisée dans la partie suivante.

Reconstruction des sous-structures à partir des motifs fréquentiels La Figure 4.14 montre, pour chaque motif fréquentiel k, la matrice d'adjacence agrégée sur tous les pas de temps du réseau temporel obtenue après transformation de la collection temporelle de signaux issue des spectres temporels.

Cette représentation confirme les relations entre le spectre et la structure décrites dans la Section 2. La sous-structure à partir du premier motif fréquentiel ressemble à un graphe k-régulier, la sous-structure 2 affiche une structure en communautés, et la sous-structure 3 semble ne pas avoir de structure claire. L'obtention d'un réseau temporel permet néanmoins de ne pas se limiter à cette représentation statique, et d'aller observer à chaque pas de temps comment cette sous-structure apparaît dans le réseau. On remarque en effet que la sous-structure 1 semble capter également un type de structure en communautés, ou que la deuxième sous-structure laisse apparaître des communautés qui ne semblent pas exister dans le réseau temporel. Ce suivi temps par temps, qui ne sera pas détaillé par la suite, permet d'avoir un aperçu plus fin de ces sous-structures. D'un point de vue global, cette décomposition a néanmoins permis de révéler les trois sous-structures présentes dans le réseau temporel, ainsi que de détecter les mélanges entre ces sous-structures.

La combinaison des informations entre la structure, révélée par les motifs fréquentiels, et les périodes d'activation de chacun de ces motifs, permettent de suivre l'évolution de la structure du réseau temporel synthétique : la période 1 est ainsi structurée en une unique communauté avec une structure aléatoire, la période 2 est uniquement structurée en communautés, la période 3 est principalement composée d'une structure régulière en anneau, avec une structure aléatoire présente avec une faible intensité et qui décroît au cours du temps. Enfin, la période 4 est un mélange entre structure régulière en anneau et structure en communautés.



(a) Motif fréquentiel 1 (b) Motif fréquentiel 2

FIGURE 4.15 – Motifs fréquentiels obtenus après application de la NMF sur le réseau temporel des interactions sociales dans une école primaire. Les colonnes de la matrice $\boldsymbol{W}$ sont retaillées en matrice.

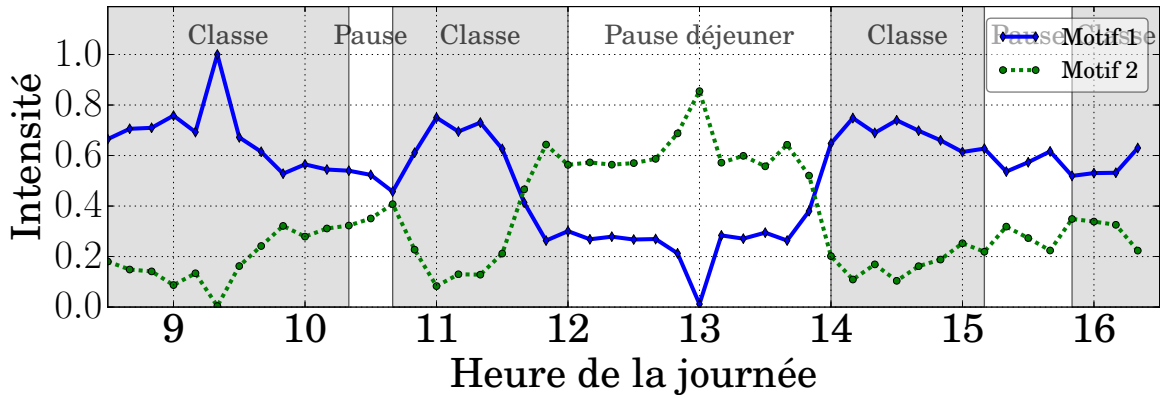
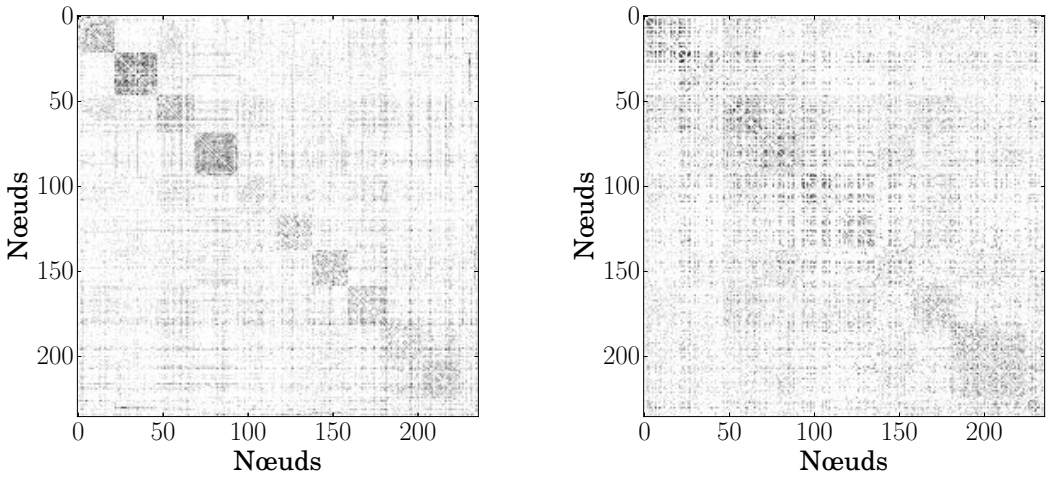


FIGURE 4.16 – Coefficients d'activation pour chacun des motifs obtenus à la Figure 4.15, correspondant aux lignes de la matrice $\boldsymbol{H}$. Les valeurs sont normalisées par la valeur maximale de $\boldsymbol{H}$.

3.4 Application au réseau temporel des interactions sociales dans une école primaire

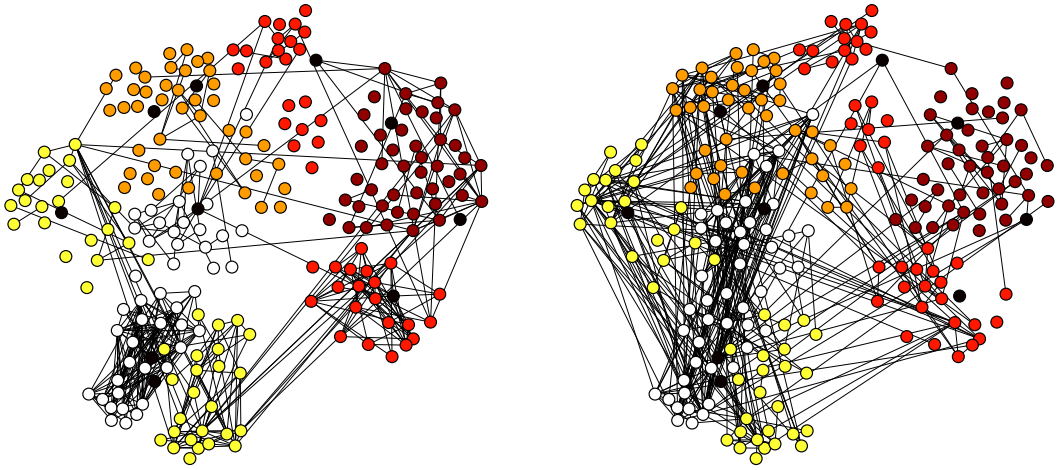
La décomposition en sous-réseaux temporels est ensuite appliquée au réseau temporel des interactions sociales dans une école primaire. À la lumière de l'étude sur les spectres temporels, deux types de période semblaient se dégager, une correspondant aux heures de classes et l'autre aux pauses du matin, du midi et de l'après-midi. Afin de valider cette hypothèse, le nombre de motifs fréquentiels est fixé à 2. Le paramètre γ est quant à lui fixé 5, encore une fois par une approche empirique.

La Figure 4.16 montre les coefficients d'activation obtenus pour chacun des motifs. La journée d'école est divisée en deux périodes, ce qui semble valider notre hypothèse. Le premier motif est actif de 8h30 à 12h, puis de 14h à 16h30, c'est-à-dire pendant les périodes de classes. Le deuxième motif est lui actif entre 10h et 10h40, puis entre 11h50 et 13h50, et enfin à partir de 15h20 jusqu'à la fin de la journée. Ces périodes correspondent environ aux pauses du matin, de midi et de l'après-midi. On obtient finalement un découpage de la journée en trois périodes : la première pendant laquelle seule le motif 1 est actif, correspondant aux périodes de classes. Les périodes pendant lesquelles les deux motifs sont actifs, correspondant aux pauses du matin et de l'après-midi. Enfin, la dernière période concerne la pause déjeuner, pendant laquelle seul le motif 2 est actif. Ce découpage est cohérent avec la vérité de terrain, à savoir que pendant les périodes de classes, les groupes sont séparés, pendant les pauses, une autre structure apparaît, même si la structure en classe est toujours présente à cause de la rotation de la récréation,



(a) Motif 1 (b) Motif 2

FIGURE 4.17 – Matrice d'adjacence agrégée sur tous les pas de temps du réseau temporel obtenue après transformation de la collection temporelle de signaux issue des motifs fréquentiels.



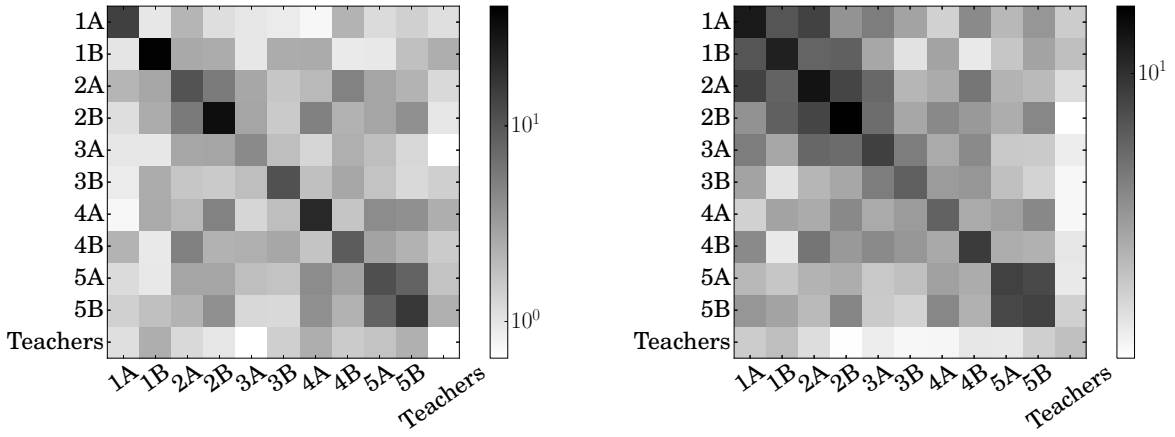
(a) Motif 1 (b) Motif 2

FIGURE 4.18 – Graphe obtenu après agrégation sur tous les pas de temps et seuillage de la matrice d'adjacence, en utilisant la même représentation qu'à la Figure 4.7.

et enfin pendant la pause déjeuner, il n'y a plus de structuration en classes.

La Figure 4.17 affiche la matrice d'adjacence agrégée sur tous les pas de temps du réseau temporel obtenue après transformation de la collection temporelle de signaux issue des motifs fréquentiels, alors que la Figure 4.18 affiche le graphe correspondant, en utilisant la même représentation qu'à la Figure 4.7. La première observation que l'on peut faire est que le motif 1 semble correspondre à une structure avec les classes. Le graphe correspondant semble confirmer cette représentation, en affichant des liens entre les nœuds d'une même couleur. Le motif 2 présente une matrice d'adjacence beaucoup moins structurée, même si quelques éléments de structure en communautés semblent ressortir. Les reconstructions ne permettent cependant pas de saisir de manière claire les structures sous-jacentes.

La Figure 4.19 affiche pour chaque motif la matrice de contact entre chaque classe : les entrées de la diagonale donnent le nombre de contacts entre les élèves à l'intérieur de chaque classe, alors que les autres entrées donnent le nombre de contact entre les classes. La couleur est codée sur une échelle de gris,



(a) Motif 1

(b) Motif 2

FIGURE 4.19 – Matrice de contact entre chaque classe pour chaque motif : les entrées de la diagonale donnent le nombre de contacts entre les élèves à l'intérieur de chaque classe, alors que les autres entrées donnent le nombre de contact entre les classes. La couleur est codée sur une échelle de gris, en utilisant pour des raisons de lisibilité une échelle logarithmique.

en utilisant pour des raisons de lisibilité une échelle logarithmique. Ces matrices de contacts donnent des informations un peu plus claire sur la structure des deux motifs : le premier motif affiche de manière distincte la structure par classe, du fait de la forte coloration des éléments diagonaux. Le motif 2 illustre également une structure en communautés, mais qui concerne un regroupement de classes plutôt qu'une classe unique. Plusieurs niveaux de communautés semblent ainsi apparaître. Ces observations confirment les observations réalisées dans la Section 1.4, c'est-à-dire une séparation entre les périodes de classe et les périodes de pause.

3.5 Discussions

Ces résultats permettent de mettre en valeur l'intérêt à décomposer un réseau temporel en plusieurs sous-réseaux, qui peuvent être étudiés indépendamment les uns des autres. Cette décomposition est rendue possible par la représentation du réseau temporel sous la forme d'une collection temporelle de signaux. L'association d'un motif fréquentiel à chaque pas de temps permet, en décomposant le tenseur des motifs fréquentiels à l'aide de la NMF, de repérer les motifs fréquentiels, et ainsi les structures de graphe, les plus significatifs. Sans connaissance a priori sur la structure, les différentes périodes d'activité dans l'école primaire sont révélées, et peuvent ainsi guider l'analyse du système en restreignant l'analyse sur plusieurs intervalles de temps. De la même manière, la décomposition du réseau temporel synthétique a permis de révéler les trois structures, même lorsque ces structures se combinent.

Quelques limites à cette approche peuvent néanmoins être soulevées. La première concerne le choix des paramètres de la factorisation, et notamment du nombre de motifs. S'il existe des méthodes pour trouver un nombre raisonnable du nombre de composantes, comme celle proposée dans la section suivante, il n'a pas été possible pour les deux exemples de faire ressortir un nombre pertinent vis à vis de l'application. Le choix le plus judicieux semble ainsi de se laisser guider par l'application. La deuxième limite vient dans la reconstruction des motifs fréquentiel, qui ne permet pas d'obtenir de manière claire la structure représentée par le motif.

Ces limites empêchent l'utilisation de cette méthode sur le réseau temporel obtenu à partir des données Vélo'v, qui est un réseau avec des structures complexes et naturellement peu identifiables, ne permettent pas d'appliquer ce processus au réseau Vélo'v. S'il est possible d'obtenir des motifs temporels

cohérents avec la réalité des déplacements, la reconstruction des sous-réseaux temporel se révèle problématique et ne permet pas d'obtenir des motifs spatiaux clairs et distincts.

Ce échec relatif, qui ne doit cependant occulter les avantages que cette approche laisse apercevoir, a permis néanmoins de réfléchir à la décomposition de réseau temporel avec le point de vue de la factorisation de matrice. Il a ainsi à l'origine des travaux présentés dans la section suivante, dans lesquels la NMF est appliquée directement sur le tenseur d'adjacence.

4 Application aux les données vélo'v

4.1 Décomposition de réseau dynamique dans le domaine des graphes

4.2 Principe de la méthode

Une approche différente est discutée dans cette dernière section, réalisée en collaboration avec Cédric Févotte. L'idée consiste à décomposer directement le tenseur d'adjacence $\mathcal{A}$ plutôt que de passer par la transformation en collection de signaux.

Plusieurs approches ont déjà été proposées pour adapter la NMF au domaine des réseaux, qu'ils soient statiques [263] ou temporels [106]. Dans cette dernière approche, le tenseur d'adjacence est décomposé par un produit tensoriel de vecteurs de rang 1 à l'aide d'une variante de la NMF appelée factorisation tensorielle non-négatives (NTF).

L'approche proposée ici consiste à transformer le tenseur d'adjacence en une matrice et à appliquer la NMF classique sur cette matrice. Pour cela, pour chaque intervalle de temps $[t, t + \Delta_t]$, la matrice d'adjacence correspondante est transformée en un vecteur colonne par empilement bout-à-bout des colonnes. Ce vecteur constitue la t^{e} colonne de la matrice $\boldsymbol{V}$. Après décomposition par NMF, les colonnes de $\boldsymbol{W}$ sont ainsi des matrices d'adjacence, obtenues suivant le procédé inverse. À la différence de la décomposition introduite à la Section 3, les structures obtenues sont statiques.

4.3 Application aux données Vélo'v

4.3.1 Construction du réseau temporel Vélo'v

Le réseau temporel est construit à partir des trajets individuels des utilisateurs : un trajet correspond à un mouvement d'une station m à une station n, partant à l'instant t_m et arrivant à l'instant t_n.

Soit $\mathcal{S}$ l'ensemble des stations du réseau et $\mathcal{T}$ la durée de temps considérée. Le temps est tout d'abord divisé en I intervalles de temps réguliers de longueur Δ_t, définissant l'ensemble des intervalles $\mathcal{I} = \{[i, i + \Delta_t]\}_{i\in\{0,\dots,I-1\}}$. Le temps $t \in \mathcal{T}$ appartient à l'intervalle i si $t \in [i, i + \Delta_t]$.

Un déplacement est ainsi un élément appartenant à l'ensemble $\mathcal{S} \times \mathcal{S} \times \mathcal{I} \times \mathcal{I}$. Un premier réseau temporel est défini par $\mathcal{G} = (\mathcal{S}, \mathcal{E}, \mathcal{I}, \mathcal{W})$, où l'ensemble des nœuds $\mathcal{S}$ correspond à l'ensemble des stations du système, l'ensemble des liens $\mathcal{E} = (\mathcal{S}\times\mathcal{S})$ est l'ensemble des couples « Origine/Destination » possibles, $\mathcal{I}$ est l'ensemble des intervalles temporels considérés et $\mathcal{W} = \{w_{ei}\}_{e\in\mathcal{E},i\in\mathcal{I}}$ donne le poids du lien $e = (m, n)$, correspondant au nombre de vélos partant de la station m vers la station n pendant l'intervalle i. À partir de $\mathcal{G}$, un réseau temporel moyen $\tilde{\mathcal{G}}$ sur la semaine est défini.

Soit $D : \mathcal{I} \to \{1, \dots, 7\}$ une fonction donnant le jour de la semaine dans lequel se trouve l'intervalle $i \in \mathcal{I}$ (on supposera dans la suite qu'un intervalle appartient à un seul jour). On a alors $\tilde{\mathcal{G}} = (\mathcal{S}, \mathcal{E}, \mathcal{J}, \tilde{\mathcal{W}})$, où $\mathcal{J}$ est l'ensemble des intervalles de temps de longueur Δ_t sur une semaine, et $\tilde{\mathcal{W}} = \{\tilde{w}_{ej}\}_{e\in\mathcal{E},j\in\mathcal{J}}$ est défini par :

$$\tilde{w}_{ej} = \frac{1}{N_j} \sum_{i\in\mathcal{I}} \mathbf{1}_{D(i)=D(j)} w_{ei} \tag{4.10}$$

avec $N_j = \sum_{i\in\mathcal{I}} \mathbf{1}_{D(i)=D(j)}$.

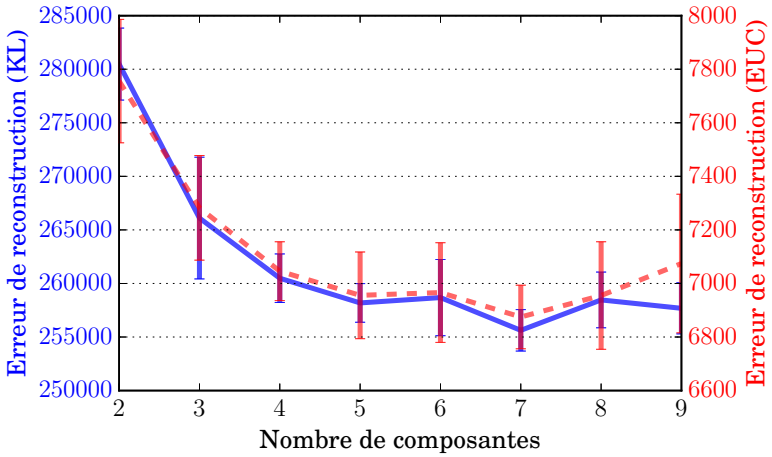


FIGURE 4.20 – Erreur de reconstruction après validation croisée pour K variant de 2 à 9. L'erreur est calculée en utilisant la divergence de Kullback-Leibler (solide bleue) et la distance euclidienne (pointillés rouges).

Le réseau temporel $\tilde{\mathcal{G}}$ est représenté par un tenseur d'adjacence $\mathcal{A}$, où chaque tranche t du tenseur $\mathcal{A}^{(t)}$ représente la matrice d'adjacence au temps t. Contrairement aux cas étudiés jusque-là, chaque instantané du réseau temporel est un graphe pondéré et dirigé.

4.3.2 Décomposition de la matrice d'adjacence

Analyse temporelle des rythmes hebdomadaires Deux paramètres nécessitent d'être réglés dans l'algorithme de la NMF : la mesure de dissemblance D, contrôlée ici par la valeur de β, et le nombre de motifs K. Le choix de β est guidé par le modèle probabiliste dans lequel nous nous plaçons par rapport à la nature des données : dans notre application sur le système Vélo'v, les éléments v_{it} de la matrice $\boldsymbol{V}$ décrivent un nombre moyen de vélos au départ d'une station vers une autre station pendant un intervalle de temps, qui peuvent se modéliser par une loi de Poisson :

$$v_{it} \sim \mathcal{P}(v_{it}, \sum_{k=1}^{K} w_{ik} h_{kt}) \tag{4.11}$$

où $\mathcal{P}(x, \lambda) = e^{-\lambda} \frac{\lambda^x}{\Gamma(x+1)}$ avec $\Gamma(x+1)$ la fonction Gamma. Comme décrit dans [239], sous des hypothèses d'indépendance des éléments de $\boldsymbol{W}$ et $\boldsymbol{H}$, maximiser la vraisemblance du modèle revient à minimiser la divergence de Kullback-Leibler, c'est-à-dire la β-divergence pour $\beta = 1$. Il n'existe en revanche pas de choix naturel pour le nombre de motifs K, que nous proposons de choisir par validation croisée : pour une valeur de K fixée, la NMF est réalisée sur une matrice $\boldsymbol{V}_{VC}$ construite en supprimant aléatoirement 50 % des valeurs de $\boldsymbol{V}$. La matrice $\tilde{\boldsymbol{V}}$ obtenue est comparée avec les valeurs de $\boldsymbol{V}$ sur les données manquantes de $\boldsymbol{V}_{VC}$, en utilisant la distance euclidienne et la divergence de Kullback-Leibler. Ce processus est répété 15 fois afin d'évaluer la dispersion de l'erreur de reconstruction. La Figure 4.20 affiche l'erreur de reconstruction pour K variant de 2 à 9, afin de conserver un nombre raisonnable de motifs dans l'interprétation. Pour les deux fonctions d'erreur, la valeur de K qui minimise l'erreur de reconstruction est $K = 7$, choix qui sera retenu par la suite pour la décomposition du réseau temporel.

La Figure 4.21 affiche pour chacun des motifs les coefficients d'activation pour chaque intervalle au cours de la semaine. Les pics d'intensité permettent de classer les motifs suivant deux critères : la période de la journée (matin, midi, après-midi, soirée et nuit) et le type de jour (jour de semaine ou week-end), ce qui est réalisé dans la Table 4.1. La description temporelle des motifs montre que la NMF extrait

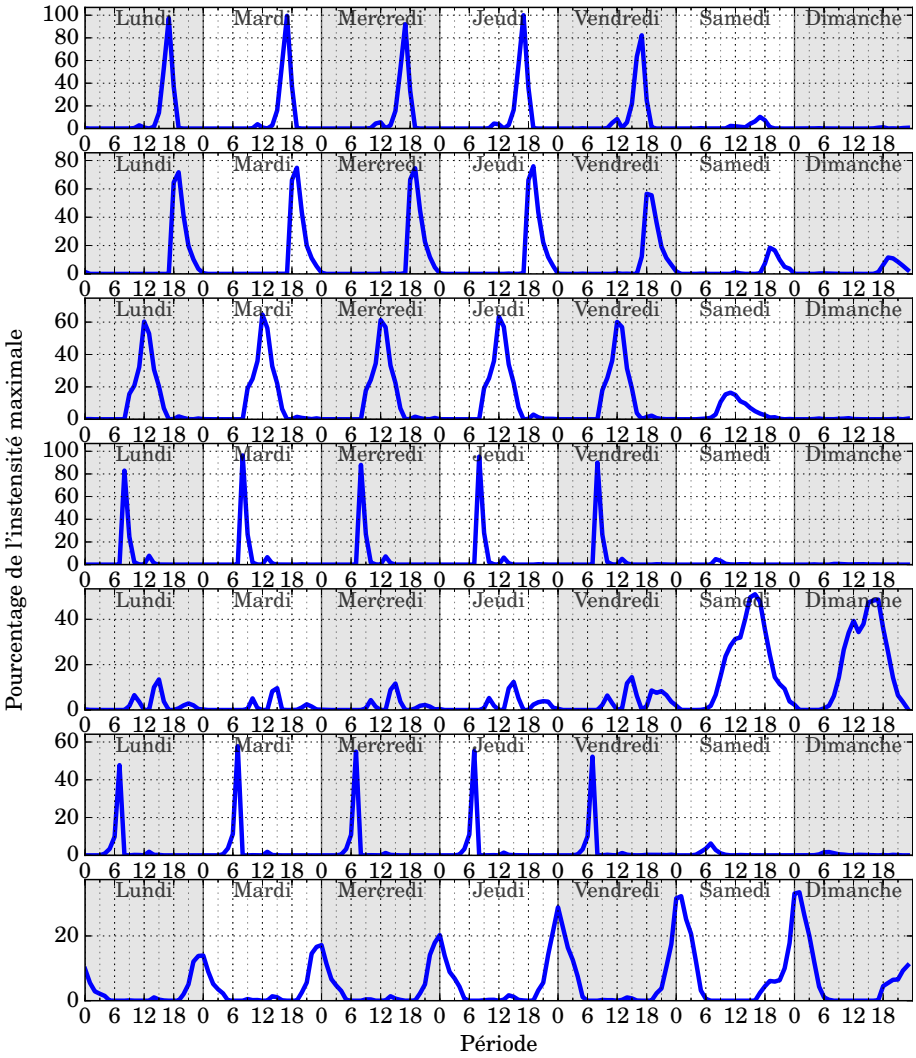


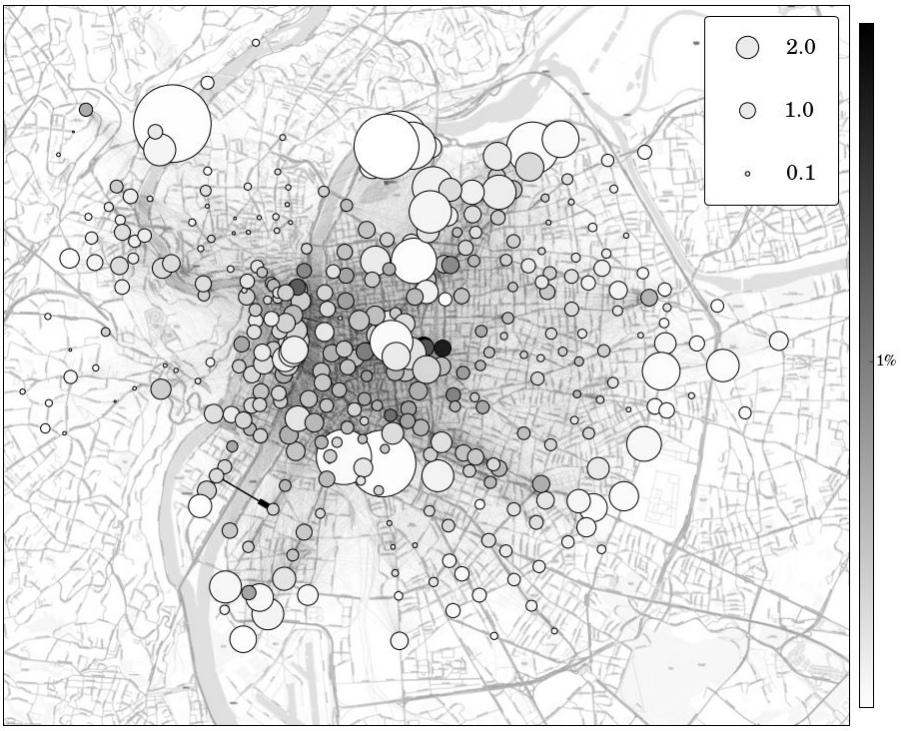
FIGURE 4.21 – Coefficients d'activation pour chaque motif au cours de la semaine.

de manière automatique les périodes pertinentes d'un point de vue socio-économique, à savoir les pics d'activité du système explicités dans le Chapitre 1.

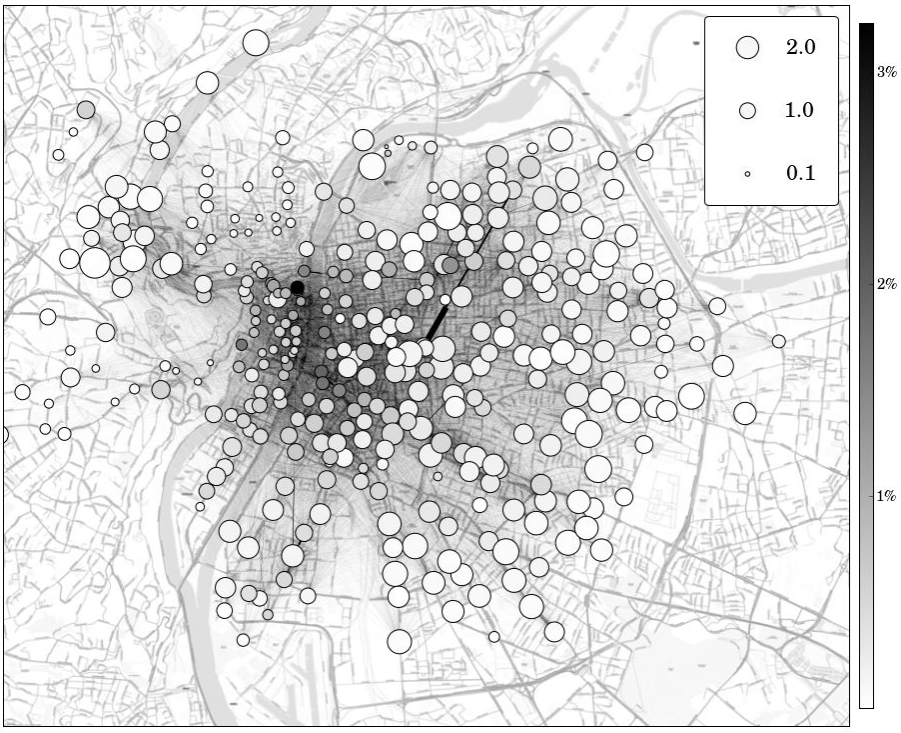
On peut noter que la modulation du coefficient d'activation permet d'associer un même comportement, décrit dans la section suivante, avec des variations dans l'intensité de ce comportement. Par exemple, le motif 7 exprime une activité présente la nuit entre 23h et 3h du matin, similaire pour chaque jour de la semaine, mais avec une intensité plus forte le jeudi, vendredi et samedi, jours où l'activité nocturne est plus intense.

Analyse spatiale des rythmes hebdomadaires Par souci de visualisation, la Figure 4.22 affiche ces matrices sous la forme d'un graphe déployé dans l'espace géographique lyonnais, où les nœuds du graphe sont les stations. Le degré sortant w_{out} (respectivement le degré entrant w_{in}) d'un nœud est défini comme la somme des poids des liens partant du nœud (respectivement la somme des poids des liens arrivant vers le nœud). La couleur indique le degré sortant du nœud sur une échelle de blanc à noir. Comme les motifs sont normalisés, ce degré représente un pourcentage de l'activité totale. La taille du point indique le ratio $\frac{w_{\text{in}}}{w_{\text{out}}}$ pour chaque nœud : plus le nœud est grand, plus la station se remplit. Inversement, plus le nœud est petit, plus la station se vide.

L'interprétation spatiale des motifs est ici limitée à deux motifs. La Figure 4.22a affiche le réseau



(a) Motif 4



(b) Motif 7

FIGURE 4.22 – Représentation des motifs sous la forme d'un graphe déployé dans l'espace géographique lyonnais. Les nœuds correspondent aux stations. La couleur des nœuds indique le degré sortant du nœud sur une échelle de blanc à noir, représentant un pourcentage de l'activité totale. La taille des nœuds indique le ratio entre degré entrant et degré sortant. Les arcs entre les nœuds sont représentés et leur épaisseur est proportionnelle au poids associé à chaque arc.

Motif	Période
1	Semaine - Après-midi (17h)
2	Semaine - Début de soirée (18h-19h)
3	Semaine - Midi (11h-14h)
4	Semaine - Matin (8h)
5	Week-end - Journée
6	Semaine - Matin (7h)
7	Nuit (23h-3h)

TABLE 4.1 – Classification des motifs en fonction des pics d'intensité des coefficients d'activation suivant deux critères : la période de la journée (matin, midi, après-midi, soirée et nuit) et le type de jour (jour de semaine ou week-end).

correspondant au motif 4, principalement actif le matin les jours de semaine. L'étude de ce réseau permet de retrouver des éléments déjà connus sur les rythmes matinaux dans la ville de Lyon. Tout d'abord, l'activité est concentrée sur quelques stations dans des zones spécifiques telles que Part-Dieu, regroupant la gare, un centre commercial et le quartier des affaires, ainsi que dans le centre de la presqu'île. Les comportements des stations sont également cohérents avec les analyses socio-économiques déjà réalisées : les stations qui se remplissent sont situées autour des campus (La Doua, Université Jean-Moulin, etc.) ou des zones à forte activité commerciale. Parallèlement, les stations qui se vident sont situées dans les zones résidentielles (8ème arrondissement, ouest du 3ème arrondissement, etc.) et les zones en altitude (Croix-Rousse, Fourvière). Ce réseau souligne ainsi la disparité géographique des déplacements urbains dans une période de grande activité. Les mouvements sont mieux répartis pendant les périodes de faible activité, comme le montre la Figure 4.22b correspondant au motif 7, dont les coefficients d'activation présentent des pics d'intensité la nuit entre 23h et 3h du matin chaque jour de la semaine. Si l'activité est concentrée comme attendu autour des zones dynamiques telles que les quartiers de l'Opéra, de Saint Jean ou de la Guillotière, regroupant de nombreux bars et restaurants, les stations situées au centre sont enclines à se vider au profit de celles situées à la périphérie de la ville. Cela souligne ainsi l'usage du vélo comme moyen de transport alternatif lorsque les métros et bus ne fonctionnent plus.

5 Conclusion et perspectives

Les résultats de ce travail ont ouvert la voie à une analyse conjointe de la dynamique d'un réseau temporel et des motifs d'interactions caractérisant cette dynamique, avec pour objectif de pouvoir séparer en temps et dans le domaine des graphes les caractéristiques du réseau temporel.

Dans une première partie, une nouvelle approche de suivi de la structure d'un réseau temporel a été proposée, à partir de la dualité entre réseaux et signaux étudiée dans le chapitre 3. À chaque pas de temps, le réseau temporel est représenté par un graphe, auquel un motif fréquentiel, caractéristique de sa topologie, est associé. La dynamique de ces motifs temporels permet de suivre également la structure du réseau et, grâce aux techniques de traitement du signal développées, notamment pour l'analyse musicale, il est possible d'extraire de manière automatique les structures temporelles les plus pertinentes. Le réseau temporel se représente alors comme la combinaison de structures, variant dans l'espace des graphes et dans le temps. Une validation empirique, sur deux réseaux temporels à la structure connue soit par construction pour le réseau temporel synthétique, soit par connaissance du système pour le réseau temporel des interactions sociales dans une école primaire, a permis d'apprécier la capacité de la méthode pour automatiquement dégager les structures pertinentes.

Néanmoins, des limites apparaissent lorsque qu'il s'agit de reconstruire des sous-réseaux à la structure plus complexe. Les tests sur le réseau temporel issu des données Vélo'v, montrent en effet que si la structure temporelle est identifiable, les sous-réseaux temporels correspondants restent difficile à construire. Plusieurs options pour pallier ce problème pourraient être envisagées : la première serait de construire le réseau statique correspondant au motif fréquentiel extrait à partir de la NMF. Pour cela, une reconstruction en signaux nécessiterait d'associer à ce motif une matrice de phases, qui pourrait être sélectionnée comme étant les phases obtenues au temps pendant lequel le motif est à son maximum d'activation. L'inconvénient de cette structure est qu'elle ne permet pas d'avoir des structures dynamiques, et en particulier dans l'exemple du réseau temporel synthétique, il ne serait pas possible de reconstruire les deux types de structures en trois communautés obtenus. Une deuxième option serait de considérer des phases dynamiques, en développant une méthode de reconstruction des signaux adaptée aux motifs fréquentiels.

Néanmoins, l'application de la NMF sur le tenseur d'adjacence a permis de valider la représentation des données Vélo'v sous la forme d'un réseau temporel, en mettant en évidence de manière totalement non supervisée des comportements connus, à la fois des analyses quantitatives et des études socio-économiques déjà réalisées. Ces résultats fournissent des outils innovants pour la caractérisation des réseaux temporels, et appellent à la réalisation de nouvelles études, par exemple en étendant l'analyse des résultats pour rechercher des comportements inattendus.

Conclusion

Deux motivations sont à l'origine de cette thèse. La première a été d'étudier différents aspects du système de vélo en libre-service à Lyon, à travers l'analyse des données de déplacements. La deuxième a consisté, étant donné la représentation naturelle de ces données sous la forme d'un réseau temporel, de proposer une méthode d'analyse de l'évolution de la structure du réseau, basée sur les méthodes du traitement du signal.

Tout au long de cette thèse, je me suis concentré sur ces deux objectifs. J'ai tout d'abord contribué à l'étude de l'utilisation des vélos partagé, en réalisant des analyses variées sur différents aspects du système Vélo'v. J'ai également développé une méthode d'analyse des réseaux temporels en utilisant une approche originale couplant la théorie des réseaux avec le domaine du traitement du signal, de manière à exploiter la dimension spatio-temporelle du système Vélo'v. Pour ces deux aspects de mon travail, je propose en guise de conclusion un rappel des principales contributions, leurs limites, ainsi que les principales perspectives qu'elles permettent d'envisager.

Dans le contexte des villes intelligentes présenté en introduction, la compréhension du système de vélos en libre-service lyonnais permet de mieux appréhender comment ce nouveau moyen de transport est utilisé par ses usagers, ainsi que de saisir les rythmes urbains à travers l'étude des déplacements des usagers.

Ma première contribution à ce projet a été de mettre en place la base de données des déplacements réalisés en 2011. Ce processus, qui a nécessité plusieurs étapes, a commencé par une visite de trois jours au centre de recherche et développement de JCDecaux, qui exploite le système Vélo'v, afin de récupérer les données brutes. Ce séjour m'a également permis d'acquérir une connaissance précieuse sur le fonctionnement du système, tant sur les aspects techniques que sur les pratiques des gestionnaires de tels systèmes. La deuxième étape a été de nettoyer les données, en développant les tests permettant de détecter les erreurs d'enregistrement et les incohérences. J'ai enfin activement travaillé sur la documentation de cette base, de manière à proposer aux autres chercheurs membres du projet une base de données fiable, compréhensible et facilement utilisable. Cette première contribution, qui a pris un temps non négligeable en raison du volume assez important du nombre d'enregistrements (il y a environ 7 millions de déplacements en 2011) et de leur hétérogénéité (plusieurs formats existaient, ainsi que plusieurs types de données), a été une étape essentielle pour la suite des travaux. Même si ce travail n'est pas à proprement parler de la recherche, il est à mon sens indispensable, lorsque l'on traite de données massives, de comprendre et maîtriser ces mécanismes de stockage et de nettoyage.

À partir de cette base de données, une typologie des usagers du système Vélo'v a été mis en place, en collaboration avec Julien Barnier, Isabelle Mallon et Marie Vogel du centre Max Weber, Luc Merchez du laboratoire Environnement, Ville, Société, et Patrice Abry et Guillaume Lozenguez du laboratoire de Physique. Ces travaux ont permis la caractérisation de différents groupes d'utilisateurs à travers l'étude de l'intensité et de la régularité de leur pratique du système de vélos en libre-service. Ces travaux ont été l'occasion pour moi de mettre en pratique l'interdisciplinarité, et de me rendre compte de la richesse, mais également des difficultés, qu'elle induit. Enfin, une extension plus méthodologique sur différentes manières de normaliser les données a été proposée. Cette dernière étude montre que le travail interdisciplinaire amène également à des réflexions ciblées sur des détails techniques. La principale perspective

de ces travaux, si je me limite aux aspects d'analyse de données, consiste à affiner la typologie en définissant de nouveaux profils pour les usagers. Il est par exemple intéressant de regarder, en plus de la régularité temporelle, la régularité spatiale, c'est-à-dire quelle est la fréquence d'un trajet donné parmi les trajets réalisés par un utilisateur. L'utilisation d'autres méthodes de classification pourrait également être envisagée, permettant d'avoir des distinctions plus nettes entre les groupes d'utilisateurs.

Ma troisième contribution a été d'étudier les stations sous différents points de vue. Le premier a été de regarder s'il était possible d'expliquer les flux entrants et sortants de vélos à l'aide de facteurs socio-économiques associés à la zone autour des stations. Ces travaux, décrits dans l'annexe A, ont donné les premiers éléments nécessaires à la compréhension de ces flux, notamment à travers une méthode de répartition originale des variables socio-économiques sur les stations.

Un deuxième travail a consisté à développer une méthode afin de détecter les périodes pendant lesquelles les stations sont pleines et vides, à partir des seuls déplacements. Ces moments constituent en effet des moments importants, car ils témoignent d'un dysfonctionnement dans le système. D'une part, ils réduisent l'activité en ne satisfaisant pas l'offre, et d'autre part ils provoquent un ressenti négatif des utilisateurs sur le système. La méthode proposée, présentée dans l'annexe C, permet de repérer ces moments à partir des seuls déplacements. À partir de cette détection, il est possible de classer les stations en plusieurs catégories. Ces travaux restent pour le moment inachevés car jusqu'à récemment, je n'avais pas à disposition les données de disponibilités des stations qui, dans une certaine mesure, permettent de valider les résultats obtenus par la méthode développée. La mise à disposition récente de données de déplacements et de disponibilités des stations sur une même période de temps permet d'envisager cette validation. Ces informations pourraient ainsi être utilisées pour les modèles de régression linéaires, afin d'expliquer à l'aide de facteurs socio-économiques ces moments pendant lesquels les stations sont vides ou pleines. Elles peuvent également servir de base pour la création de profils de station, en vue de faire des classifications, comme cela a été abordé à travers le travail de Maximilien Thess, en stage au laboratoire de Physique.

Une dernière contribution, que je n'ai pas évoquée dans le Chapitre 1, a été de proposer des outils de visualisation des données Vélo'v. J'ai ainsi développé un outil pour visualiser de manière dynamique les résultats présentés dans l'Annexe C de façon intuitive. J'ai travaillé également sur une interface pour réaliser des tris croisés sur les données, en fonction de nombreux paramètres portant sur les trajets, les usagers ou des indicateurs spatio-temporels. Ces outils ont ainsi été utiles pour explorer la base de données, en permettant d'obtenir rapidement des informations à partir des données lors de réunions de travail.

Le deuxième objectif de mes travaux de thèse a été de proposer une méthode d'analyse de l'évolution de la structure des réseaux temporels. Elle se décompose en plusieurs étapes, et met en jeu des méthodes à la fois de théorie des graphes, de théorie des réseaux, de statistique et de traitement du signal. Dans le Chapitre 3, une dualité entre réseaux et signaux est présentée, permettant de représenter un réseau statique en une collection de signaux puis, à partir de cette collection de signaux, d'obtenir un réseau. Cette transformation inverse est qualifiée de robuste car lorsqu'une perturbation vient affecter la collection de signaux, une perturbation affecte également le réseau correspondant, dans des proportions mesurées comme similaires. Cette caractéristique est importante, car elle permet d'agir sur le réseau en agissant directement sur la collection de signaux la représentant.

Cette double représentation d'un réseau, à la fois sous forme de graphe et sous forme de signaux est utilisée pour permettre l'utilisation des outils de traitement du signal pour l'étude des réseaux. Le problème de l'indexation des signaux, réalisée à l'aide de la numérotation des nœuds, a néanmoins souligné le besoin de disposer d'un étiquetage des nœuds cohérents avec la structure des graphes, afin d'obtenir des signaux lisses, et ainsi de pouvoir dégager des propriétés spectrales pertinentes. Ce problème a été résolu en développant une heuristique pour l'étiquetage des nœuds, qui a été exposée dans le Chapitre 2.

En se basant sur la résolution d'un problème classique de la théorie des graphes, une heuristique a été développée et est capable d'obtenir, pour un graphe quelconque, un étiquetage des nœuds cohérent avec la structure, fournissant alors une solution au problème de l'indexation des signaux par les nœuds. La transformation ainsi définie permet de décrire le réseau à travers la description des signaux, notamment en utilisant l'analyse spectrale. La définition d'une analyse composantes-fréquences m'a permis d'établir des liens entre les motifs obtenus dans ce plan et la structure du graphe correspondant.

Cette dualité permet également de préciser la sémantique utilisée jusque-là, en différenciant les notions de réseau et de graphe. Tout au long de ce manuscrit, je n'ai fait qu'une faible distinction entre les deux termes, parlant plutôt de graphe pour parler des nœuds et des liens, et de réseau pour désigner l'ensemble des relations. Ce raccourci, communément utilisé dans la littérature, est néanmoins réducteur car il amalgame le réseau, qui représente le système réel, avec le graphe, qui n'en est que la simplification, ou le modèle. En définissant une nouvelle représentation d'un réseau, ces travaux viennent confirmer la différence entre ces deux notions, à savoir que le réseau est la réalité, représenté de façon approximative soit par un graphe, soit par une collection de signaux.

Ce travail dans le cas statique a été étendu naturellement au cas dynamique, en considérant une succession de motifs fréquentiels. Le Chapitre 4 présente cette extension, et ainsi qu'une validation empirique sur des réseaux temporels à la structure simple mais néanmoins non triviale. J'ai proposé également une méthode de décomposition du réseau temporel à l'aide de la factorisation non négatives, en établissant un parallèle ces réseaux temporels et des signaux musicaux. Il en ressort que la méthode permet de retrouver les structures principales du réseau temporel, même lorsque ces structures sont mélangées. Les limites évoquées, notamment sur la reconstruction des sous-structures, m'ont empêché d'appliquer cette méthode sur le réseau temporel issu des données Vélo'v, et soulignent le travail qui reste à parcourir avant de pouvoir étudier pour des systèmes complexes. Néanmoins, les résultats encourageants obtenus laissent entrevoir les avantages à coupler science des réseaux et traitement du signal.

Mon apprentissage des techniques de factorisation en matrices non négatives a été réalisé notamment grâce à l'aide apporté par Cédric Févotte, chercheur au Laboratoire Lagrange à Nice. Cette collaboration, qui s'est poursuivie pendant un séjour de recherche de deux semaines à Nice grâce au soutien du GdR ISIS, m'a également permis également de travailler plus spécifiquement sur une méthode de décomposition du réseau temporel des déplacements Vélo'v, réalisée dans le domaine des graphes. L'analyse des motifs spatio-temporels obtenus, en accord avec les connaissances déjà acquises sur le système, soulignent la pertinence de l'utilisation du traitement du signal pour l'analyse de réseaux, qui représente à mon avis une piste prometteuse pour mieux comprendre la dynamique des systèmes complexes.

Enfin, je souhaiterais terminer ce manuscrit en évoquant les activités qui ne sont pas directement reliées aux travaux présentés dans ce manuscrit de thèse, mais qui constituent à mon sens un complément indispensable au travail de recherche. Ces activités concernent notamment l'enseignement, que j'ai pu pratiquer pendant les trois années de ma thèse à l'ENS Lyon et l'INSA Lyon, ainsi que les activités de vulgarisation scientifique auxquelles j'ai participé, par exemple lors de Fête de la science. Ces évènements sont l'occasion d'échanges enrichissants sur la recherche scientifique, ainsi que sur les préoccupations actuelles autour des réseaux et de l'analyse de données, permettant une prise de conscience des risques, mais surtout des opportunités que cette révolution des données apporte.

ANNEXE A

Modèles de régressions linéaires sur les données Vélo'v

1 Présentation des données et uniformisation spatiale

Pour chaque station, on dispose des données suivantes pour la période du 01/11/2005 au 01/06/2006 :

— les flux totaux entrants et sortants durant différentes périodes de temps :

1. les jours de semaine de 7h à 9h afin d'étudier les trajets «Domicile -> Travail / École» ;
2. les jours de semaine de 12h à 14h afin d'étudier les déplacements pendant la pause déjeuner ;
3. les jours de semaine de 18h à 20h afin d'étudier les trajets «Travail -> École - Domicile » ;
4. les jours de week-end de 14h à 16h afin d'étudier les déplacements «Loisirs».

— la capacité ;
— l'altitude ;
— la durée totale pendant laquelle la station a été vide sur les mêmes périodes de temps que les flux ;
— la durée totale pendant laquelle la station a été pleine sur les mêmes périodes de temps que les flux ;
— la localisation géographique.

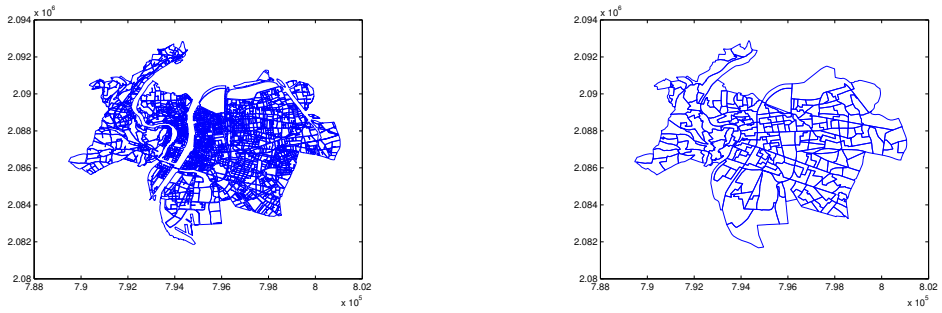
Les durées pendant lesquelles la station a été pleine ou vide sont obtenues en utilisant la méthode présentée en Annexe C.

La ville de Lyon est découpée en morceaux en Ilots (voir Figure A.1a), eux-mêmes regroupés en IRIS (Ilots Regroupés pour l'Inférence Statistique) (voir Figure A.1b).

Pour chaque IRIS, on dispose des données suivantes :
— les coordonnées du centre ;
— la géographie (sous la forme des coordonnées des sommets d'un polygone) ;
— la valeur de 68 variables démographiques et socio-économiques.

Pour chaque Ilots, on dispose des données suivantes :
— les coordonnées du centre ;
— la géographie [1] ;
— la valeur de de 159 variables démographiques et socio-économiques.

1. La géographie n'a pas pu être utilisée pour des raisons techniques. Un diagramme de Voronoi appliqué aux centre des Ilots a permis de disposer d'une géométrie approximative



(a) (b)

FIGURE A.1 – (a) Carte des Ilots de Lyon-Villeurbanne. (b) Carte des IRIS de Lyon-Villeurbanne.

Les niveaux de description des données sont différents :
— au niveau des stations ;
— au niveau des IRIS ;
— au niveau des Ilots.

Il est donc nécessaire d'uniformiser les données afin qu'une comparaison des données ait un sens.

L'uniformisation sur les Ilots n'est pas envisageable car ceux-ci sont trop nombreux ; on compte plus de 2500 îlots sur les villes de Lyon-Villeurbanne et seulement 184 stations : une répartition des flux sur les îlots serait fastidieuse et peu pertinente. Il reste donc le choix de travailler soit sur les IRIS, soit sur les stations. Ces deux approches sont présentées dans les paragraphes suivants.

Première approche : répartition sur les IRIS

La première approche utilisée lors de la genèse du projet a été de travailler sur les IRIS. Comme les stations sont le plus souvent sur les axes majeurs de circulation, frontières des IRIS, associer une station à l'IRIS le plus proche n'est pas une solution envisageable. Une technique, inspirée des modes *raster* en géographie, consiste à affecter les flux de Vélo'v au IRIS par un noyau de lissage en posant la conservation du nombre total de trajets effectués. On considère ainsi que les utilisateurs d'une station Vélo'v vont rayonner autour de leur point de départ pour trouver une station Vélo'v jusqu'à une distance critique R_0 (typiquement $R_0 = 100$ mètres) : au-delà de cette distance critique la probabilité de se déplacer va baisser. Une densité exponentielle a été proposée comme noyau de lissage (Figure A.2a). Le flux F_i affecté à l'IRIS i s'obtient, en notant G_s le flux de la station s et r_{is} la distance entre le centre de l'IRIS i et la station s, de la façon suivante :

$$F_i(t) = \frac{\sum_{s\in\{\text{Stations}\}} G_s(t).e^{\frac{-r_{is}}{R_0}}}{\sum_{j\in\{IRIS\}} e^{\frac{-r_{js}}{R_0}}} \tag{A.1}$$

Cette réaffectation des déplacements en Vélo'v aux IRIS conserve ainsi les flux de Vélo'v :

$$\sum_{i\in\{IRIS\}} F_i(t) = \sum_{s\in\{Stations\}} G_s(t) \tag{A.2}$$

La même stratégie est adoptée pour répartir les variables Ilots sur les IRIS en utilisant les centres des Ilots pour calculer la distance entre un IRIS et un Ilot. Cette approche est satisfaisante car elle permet de répartir les flux d'une station sur les IRIS et donc de prendre en compte la présence d'IRIS éloignés de toute station ; ces IRIS se retrouvent avec un flux de vélos très faibles et ne sont donc pas utilisés pour l'analyse statistique. C'est également une méthode simple à mettre en œuvre. Elle a néanmoins plusieurs défauts :

— l'interprétation du flux de vélos pour un IRIS n'est pas naturelle : le flux concerne plutôt une station qu'un IRIS ;
— le calcul de la distance entre un IRIS et une station est réduit dans la méthode proposée à un calcul entre la station et le centre de l'IRIS : la géométrie irrégulière des IRIS n'est pas prise en compte ;
— la différence de nature entre les IRIS n'est pas prise en compte, ce qui conduit par exemple à répartir équitablement le flux entre un IRIS avec une très forte densité et un IRIS avec très peu de population.

Pour ces raisons, si la méthode proposée est une bonne approximation simple à mettre en œuvre, elle reste cependant limitée si on souhaite étudier plus finement la relation entre variables économiques sur les IRIS et les flux de Vélo'v.

Deuxième approche : répartition sur les stations

Pour répondre aux critiques de la première approche pour harmoniser les données disponibles, une deuxième méthode a été élaborée et implémentée durant ce stage. Cette approche consiste à répartir cette fois les variables IRIS sur les stations qui servent donc de niveau d'étude.
Pour un IRIS i on procède aux étapes suivantes :

1. Tirage aléatoire de n points $(x_j)_{j=1,..,n}$ à l'intérieur de l'IRIS ;
2. Pour chaque point $x_j, j \in (1, .., n)$:
   a) Calcul d'une probabilité pour qu'un utilisateur partant de ce point se déplace vers une station Vélo'v : si on note R_0 la distance typique que pourra parcourir un utilisateur pour rejoindre une station (typiquement $R_0 = 100$ mètres) et D la distance de la station la plus proche, on a alors :

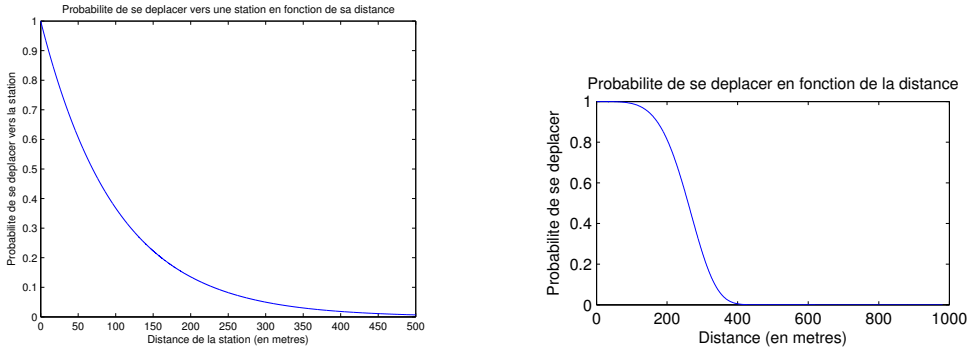
$$P_{\text{Déplacement}}(x_j) = \int_D^{+\infty} k(\frac{r}{r_0}).dr \tag{A.3}$$

   avec $k(x)$ noyau de lissage à déterminer. Le choix d'une densité de Weibull pour fonction k correctement paramétrée permet d'obtenir $P_{\text{Déplacement}}(x_j) \simeq 1$ si la station la plus proche est entre 0 et 100 mètres et $P_{\text{Déplacement}}(x_j) \simeq 0$ si la station la plus proche est à plus de 300 mètres du point x_j (voir Figure A.2b). La forme de cette loi de probabilité est arbitraire : l'idée est d'avoir une probabilité de se déplacer très forte lorsque l'utilisateur se trouve à une distance entre 0 et 100 mètres de la station la plus proche et qui chute au-delà, pour atteindre une probabilité quasi-nulle lorsqu'une distance autour de 400 mètres est atteinte ;
   b) Calcul des probabilités pour qu'un utilisateur rejoigne chaque station : en utilisant la probabilité de déplacement, on a

$$P_{\text{Déplacement}}(x_j) = \sum_{s \in \{Stations\}} P_{\text{Déplacements vers la station s}} = \sum_{s \in \{Stations\}} p(x_j, s). \tag{A.4}$$

   Le calcul des $p(x_j, s)$ se fait en utilisant le noyau de lisage $k(x)$: on répartit $P_{\text{Déplacement}}(x_j)$ sur chaque station en fonction de leur distance pour calculer les probabilités $p(x_j, s)$.
3. Le coefficient de répartition d'un IRIS sur les stations se calcule en moyennant les probabilités obtenues pour chaque point sur les stations :

$$P(i, s) = \int p(x, s).dx = \frac{1}{n} \sum_{j=1}^{n} p(x_j, s)$$



(a) (b)

FIGURE A.2 – (a) Densité exponentielle. (b) Probabilité de déplacement en fonction de la distance (loi de Weibull)

2 Nettoyage et classification des variables socio-économiques

L'uniformisation des données dans l'espace a permis de disposer pour chaque station de la valeur de 227 variables socio-économiques. La matrice des corrélations de ces régresseurs (voir Figure A.3) permet de se rendre compte visuellement des corrélations entre les variables. On remarque la présence de groupe de variables fortement corrélées entre elles. Ces fortes corrélations sont à prendre en compte dans la stratégie de résolution du problème de régression linéaire.

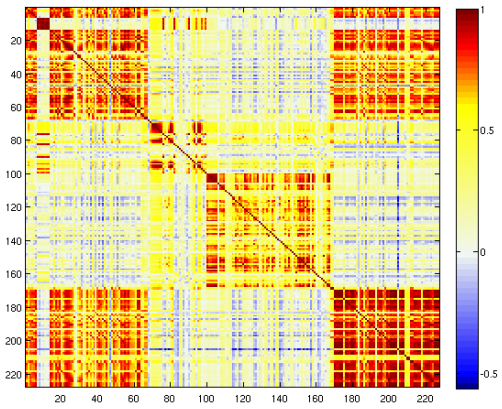


FIGURE A.3 – Représentation de la matrice des corrélations des régresseurs

Avant cette résolution, il est nécessaire de supprimer les variables qui n'ont pas beaucoup de sens dans l'étude effectuée. Une étape de classification de ces variables est également proposée afin de mettre en évidence des catégories de variables.

Nettoyage des variables socio-économiques

Avant de mettre en place le modèle de régression linéaire, il est nécessaire de procéder à un nettoyage des régresseurs. En effet, les problèmes suivants apparaissent si on regarde de près les variables :

— présence de doublons : certaines variables représentent la même donnée mais pour deux recensements différents. Pour notre étude, deux recensements ont été utilisés : celui de 1999 et celui 2006. En cas de doublon, seules les variables du recensement de 2006 ont été conservées ;

— présence de variables peu significatives : certaines variables sont très peu significatives par rapport aux autres. Par exemple de nombreuses variables représentent des populations (population de 18 à 30 ans, population d'ouvriers, etc.). Ces variables sont néanmoins à considérer en fonction de la population totale : par exemple la population d'agriculteurs est très faible à Lyon (moins de 50 individus) et n'est donc pas à considérer. Enlever ces variables est nécessaire car la méthode de régression considère chaque variable comme étant indépendante et la normalisation va gommer le fait qu'une très faible population n'est pas à considérer. Ainsi, dans le cas de la population d'agriculteurs, la présence de beaucoup d'agriculteurs pour une station à très fort flux tromperait la méthode qui verrait une forte corrélation entre agriculteurs et flux, ce qui n'est pas cohérent avec la réalité ;
— présence de variables de densité : des informations concernent des densités (par exemple d'emploi ou de population) sur les IRIS ou Ilots considérés. Ces densités n'ont plus de sens lorsque l'on parle des stations. Elles sont donc enlevées.

Une méthode rapide de nettoyage se basant sur le type de données (« Population », « Établissements », etc.) et sur la moyenne de la variable sur toutes les stations permet de faire un tri grossier des variables en se concentrant sur les points évoqués précédemment. Une analyse plus fine serait nécessaire si l'on souhaitait améliorer la pertinence des données ; le manque de temps a empêché cette analyse, d'autant plus que les perspectives sur les données sont l'utilisation des données socio-économiques du recensement de 2010. Il a ainsi été décidé d'attendre la récupération des données les plus récentes avant d'analyser les données plus finement. La réalisation d'un premier tri grossier sur les données permet d'éliminer 96 variables.

Classification des variables socio-économiques

Une idée intéressante est de procéder à une classification des variables. Le choix s'est porté sur une classification *K-means*, qui permet de rapprocher les variables suivant une distance choisie ici comme étant la corrélation. La Table A.1 présente les résultats de cette classification.

No	Nom de la classe	Nombre de variables
1	Population	21
2	Scolaire	14
3	Population	66
4	Population active sur lieu de travail	6
5	Emploi	24

TABLE A.1 – Classification des régresseurs à l'aide de la méthode *kmeans*.

Les noms donnés aux classes correspondent à la catégorie de la majorité des variables de la classe. Certaines variables n'ont cependant pas de rapport avec le nom de leur classe.

3 Techniques de régressions linéaires

Le cadre de l'étude est le suivant : on cherche à expliquer une variable par plusieurs variables explicatives. Les notations adoptées sont les suivantes :

— X est une matrice de taille $(n \times p)$; $X_i, i = 1, ..., n$ est un vecteur colonne de taille p représentant l'individu i ; $X^j, j = 1, ..., p$ est un vecteur colonne de taille n représentant la variable j ; X est appelée matrice des régresseurs, des prédicteurs ou des données.
— y est un vecteur colonne de taille n ; y est appelée la variable à expliquer ;
— β est un vecteur colonne de taille p contenant les coefficients de régression ;

— ϵ est un vecteur colonne de taille n appelé résidus ; $(\epsilon_i)_{i=1,...,n}$ sont en général supposées être des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) gaussiennes centrées de variance σ^2 connue mais ces hypothèses sont souvent prises en défaut en réalité ;
— A^t est la transposée de la matrice A ;
— $\|.\|^2$ désigne la norme ℓ_2 au carré : $\|\beta\|^2 = \sum_{i=1}^{p} \beta_i^2$.

On considère que les données sont générées suivant le modèle de régression linéaire suivant :

$$y = X\beta + \epsilon.$$

Sous forme indicielle, le modèle est le suivant :

$$y_i = X_i\beta + \epsilon_i, i = 1, ..., n.$$

On considère la matrice X centrée et le vecteur y centré également.

On se place dans un problème de grande dimension, c'est-à-dire que p le nombre de régresseurs est plus grand que n le nombre d'individus. On suppose que le paramètre β est parcimonieux (*sparse* en anglais), c'est-à-dire que peu de composantes $\beta_j, j \in \{1, ..., p\}$ sont différentes de zéro. L'intérêt de supposer la parcimonie est de permettre une interprétation plus simple du modèle qui est aussi plus facile à manipuler.

Méthode des moindres carrés

Définition La méthode usuelle pour estimer le paramètre $\beta \in \mathbb{R}^p$ est celles des moindres carrés. Elle consiste à chercher une valeur $\hat{\beta}$ du paramètre qui minimise la somme des carrés des résidus :

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - X_i\beta)^2.$$

On a ainsi :

$$\hat{\beta}^{MC} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

ou sous forme matricielle :

$$\hat{\beta}^{MC} = \arg\min_{\beta} \text{argmin}(y - X\beta)^t(y - X\beta).$$

Solution La différenciation de RSS en fonction de β permet d'obtenir :

$$\frac{\partial RSS}{\partial \beta} = -2X^t(2X^tX)$$

et

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^t} = 2X^tX$$

Si la matrice X^tX est inversible, on a alors $\hat{\beta}^{MC}$ défini de manière unique et s'écrivant sous la forme :

$$\hat{\beta}^{MC} = (X^tX)^{-1}X^ty.$$

La matrice X^tX n'est pas inversible dans les deux cas suivants :
— pour $p > n$: la matrice X^tX est alors au plus de rang n ;
— lorsque les colonnes de X sont liées (par exemple $X_2 = 2 * X_1$).

Dans ces deux cas, le paramètre β n'est pas identifiable. Dans le problème de grande dimension que nous traitons, X^tX n'est pas inversible. Plusieurs techniques existent pour pallier à ce problème, comme le recours à des algorithmes d'optimisation ou l'utilisation de la pseudo-inverse.

Commentaires Le principal inconvénient de la méthode des moindres carrés apparaît lorsque les données sont très corrélées. Des valeurs propres de la matrice X^tX sont alors très proches de zéro, entraînant un mauvais conditionnement de la matrice X^tX. Le calcul de l'inverse de cette matrice est alors instable, rendant inexploitable l'estimateur des moindres carrés.

Régression *ridge*

Définition

La régression *Ridge* ou régularisation de Thikonov ([229]) a été introduite pour pallier au problème de la méthode des moindres carrés lorsque des régresseurs sont fortement corrélés entre eux. Elle consiste à régulariser les coefficients $\hat{\beta}_j, j \in \{1, ..., p\}$ en imposant une pénalité sur leur taille de type ℓ_2, c'est-à-dire que la somme des carrés des coefficients est pénalisée :

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \{\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\} \tag{A.5}$$

λ est un paramètre qui permet de contrôler le niveau de régularisation : plus la valeur de λ est grande, plus la régularisation est forte. On peut également écrire ce problème sous la forme :

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 \text{ sous la contrainte } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

Il existe une correspondance entre ces deux problèmes : pour tout choix de λ, la solution au problème est tel qu'il existe t pour lequel elle est aussi solution du problème contraint. L'avantage de cette dernière écriture est de pouvoir contrôler la valeur de la somme à l'aide du paramètre t.

De la même manière que la méthode des moindres carrés, on peut exprimer la somme des carrés résiduels sous forme matricielle :

$$RSS(\beta, \lambda) = (y - X\beta)^t(y - X\beta) + \lambda\beta^t\beta.$$

Solution

De la même manière que pour la méthode des moindres carrés, on peut déterminer analytiquement la solution, cette fois pour toute matrice X :

$$\hat{\beta}^{ridge} = (X^tX + \lambda I_p)^{-1}X^ty$$

où I_p est la matrice identité de taille $p \times p$.

Commentaires

On peut noter que le terme $X^tX + \lambda I$ est toujours inversible, même si la matrice X ne l'est pas. L'ajout de la pénalité permet d'éviter l'inversion de la matrice X^tX. De plus, la régularisation permet de mieux conditionner les données en augmentant les petites valeurs propres de la matrice X^tX.

Une autre interprétation pour justifier de l'ajout d'une pénalité est la suivante : lorsque les variables sont très corrélées entre elles, leur coefficient peuvent présenter une grande variance : un coefficient largement négatif sur une variable peut être annulé par un coefficient largement positif sur une variable corrélée. En imposant une contrainte de taille sur les coefficients, le problème est évité.

La relation entre convexité de la pénalité *ridge* et bon comportement face aux variables corrélées est explicitée dans la partie sur l'*elastic net*.

Régression *lasso*

Définition

Introduit par Tibshirani [228], le *lasso* est une méthode de moindres carrés avec pénalisation, qui impose une contrainte sur la norme ℓ_1 des coefficients de régression permettant de forcer la parcimonie du vecteur de régression estimé.

On considère un modèle linéaire de régression avec n observations et p variables (prédicteurs).

Le vecteur des coefficients de régression $\hat{\beta}^{lasso}$ est obtenu en minimisant le critère *lasso* suivant :

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \|Y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{A.6}$$

où $\lambda \in \mathbb{R}^*$. On peut montrer que ceci équivaut au problème de minimisation contraint

$$\hat{\beta}^{lasso} = \arg\min_{\beta} (\|Y - X\beta\|^2) \text{ sous la contrainte } \|\beta\|_1 \leq t$$

Comme pour la régression *ridge*, il existe une correspondance entre λ et t.

La solution est explicite dans le cas où la matrice X est orthogonale, c'est-à-dire $X^tX = I_p$ où I_p est la matrice identité de taille p. On a alors pour tout $j = 1, .., p$:

$$\hat{\beta}_j^{lasso} = signe(\hat{\beta}_j^{MC})(|\hat{\beta}_j^{MC}| - \gamma)\mathbb{1}_{|\hat{\beta}_j| \geq \gamma}$$

où $\hat{\beta}^{MC}$ est l'estimateur des moindres carrés.

La démonstration pour y parvenir, en notant que $\hat{\beta}^{MC} = \hat{\beta}_0 = X^ty = (y^tX)^t$, est la suivante :

$$\begin{aligned}
& \min_{\beta} (y - X\beta)^t(y - X\beta) + \lambda \sum_{i=1}^{p} |\beta_i| \\
= & \min_{\beta} -2y^tX\beta + \beta^t\beta + \lambda \sum_{i=1}^{p} |\beta_i| \\
= & \min_{\beta} -2\hat{\beta}_0^t\beta + \beta^t\beta + \lambda \sum_{i=1}^{p} |\beta_i| \\
= & \min_{\beta} \sum_{i=1}^{d} -2\hat{\beta}_{0,i}\beta_i + \beta_i^2 + \lambda|\beta_i|
\end{aligned}$$

Il est ensuite nécessaire de s'occuper de chaque indice i séparément. Pour chaque indice i, on doit alors résoudre

$$\min \left\{ \min_{\beta \geq 0} (-2\hat{\beta}_0\beta + \beta^2 + \lambda\beta), \min_{\beta \leq 0} (-2\hat{\beta}_0\beta + \beta^2 - \lambda\beta) \right\}$$

On a donc respectivement :

$$\begin{aligned}
\min_{\beta \geq 0} (-2\hat{\beta}_0\beta + \beta^2 + \lambda\beta) &= \begin{cases} -(\hat{\beta}_0 - \frac{\lambda}{2})^2 & \text{si } \hat{\beta} - \frac{\lambda}{2} > 0 \\ 0 & \text{si } \hat{\beta} - \frac{\lambda}{2} \leq 0 \end{cases} \\
\min_{\beta \leq 0} (-2\hat{\beta}_0\beta + \beta^2 - \lambda\beta) &= \begin{cases} -(\hat{\beta}_0 + \frac{\lambda}{2})^2 & \text{si } \hat{\beta} + \frac{\lambda}{2} > 0 \\ 0 & \text{si } \hat{\beta} + \frac{\lambda}{2} \leq 0 \end{cases}
\end{aligned}$$

En remplaçant $\frac{\lambda}{2}$ par γ on obtient facilement le résultat énoncé précédemment.

Commentaires

L'estimateur ainsi obtenu correspond à un seuillage doux (*soft thresholding*) de l'estimateur des moindres carrés. Les coefficients $\hat{\beta}_j$ sont remplacés par $\phi_\lambda(\hat{\beta}_j)$ où $\phi_\lambda : x \to signe(x)(|x| - \lambda)_+$.

Dans le cas où la matrice X n'est pas orthogonale, on utilise des procédures d'optimisation pour trouver la solution. Un algorithme appelé LARS a également été proposé ; il est détaillé dans la section suivante.

Le *lasso* permet d'une part d'améliorer la précision du modèle par minimisation du biais, et d'autre part, du fait de la nature de la pénalité ℓ_1, de réduire à 0 des coefficients. Le modèle est ainsi à la fois précis et parcimonieux, ce qui permet de faire une sélection de variables efficace.

La solution est dite parcimonieuse (*sparse* en anglais) car elle comporte beaucoup de coefficients nuls. Le degré de parcimonie (ou *sparsity*) se règle en fonction du paramètre λ ou t suivant l'écriture choisie. La Figure A.4 permet de comprendre, dans le cas où $p = 2$, les raisons de la nullité des coefficients dans le cas de la norme ℓ_1. La somme résiduelle a des contours elliptiques centrés sur la solution des moindres carrés. La région de contrainte pour la norme ℓ_2 est le disque $\beta_1^2 + \beta_2^2 \leq t$ tandis que celle de la norme ℓ_1 est le losange $\beta_1 + \beta_2 \leq t$. Dans le cas d'une régression *lasso*, on cherche l'intersection entre les résidus (représentés par des contours elliptiques centrés en la valeur β_{MC} solution des moindres carrés) avec la région de contrainte. Contrairement au disque, le losange contient des sommets, où un des paramètres est égal à zéro. L'intersection est donc probable sur ces sommets. Quand $p > 2$, le losange devient un rhomboïde avec beaucoup de sommets, d'arêtes et de faces : les possibilités d'annulation des coefficients augmentent.

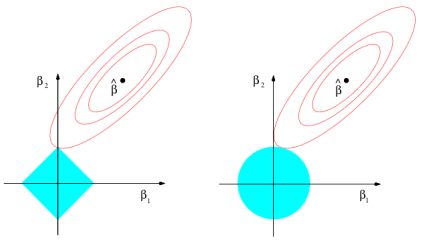


FIGURE A.4 – Représentation des régions de contrainte pour la norme ℓ_1 (à gauche) et pour la norme ℓ_2 (à droite). [98]

Malgré les bons résultats de la régression *lasso* pour obtenir de la parcimonie, cette méthode a plusieurs limitations :

1. dans le cas où $p > n$, le *lasso* va sélectionner au plus n variables avant de saturer du fait de la nature du problème d'optimisation convexe ;
2. s'il y a un groupe de variables parmi lequel la corrélation est très forte, alors le *lasso* tend à ne sélectionner qu'une seule variable de ce groupe et ne s'occupe pas de savoir laquelle ;
3. pour la situation $p < n$, si il y a de fortes corrélations entre les régresseurs, il a été montré par Tibshirani (1996) que la performance de prédiction du *lasso* est dominée par celle de la régression *ridge*.

Pour remédier à ces limitations, une nouvelle méthode de régularisation a été proposée par Zou et Hastie ([265]) et consiste en l'utilisation d'une pénalité définie comme la somme pondérée de la norme ℓ_1 et du carré de la norme ℓ_2 du vecteur des coefficients β. Le premier terme permet d'obtenir la parcimonie de la solution tandis que le second permet de prendre en compte les corrélations des variables. On obtient une méthode couplant les pénalités *ridge* et *lasso*.

Régression *elastic net*

Définition de la régression *naïve elastic net*

Une première version, appelée *naïve elastic net* est d'abord proposée. Pour $\lambda_1 > 0$ et $\lambda_2 > 0$ fixées, le critère est le suivant :

$$L(\lambda_1, \lambda_2, \beta) = \|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

où

$$\begin{aligned} \|\beta\|^2 &= \sum_{j=1}^{p} \beta_j^2 \\ \|\beta\|_1 &= \sum_{j=1}^{p} |\beta_j| \end{aligned}$$

On a alors l'estimateur du *naïve elastic net* $\hat{\beta}^{\text{nen}}$ solution de

$$\min_{\beta}\{L(\lambda_1, \lambda_2, \beta)\}.$$

Si on pose $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, on peut réécrire de façon équivalente le problème précédent sous la forme contrainte :

$$\hat{\beta}^{\text{nen}} = \arg\min_{\beta} |y - X\beta|^2 \quad \text{sous la contrainte } (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \leq t.$$

La fonction $(1-\alpha)\|\beta\|_1 + \alpha\|\beta\|^2$ est appelée la pénalité *elastic net* et est une combinaison convexe de la pénalité *lasso* et *ridge*. On peut noter les cas particuliers suivants :

— $\alpha = 1$: la régression *ridge* ;

— $\alpha = 0$: la régression *lasso*.

On peut noter que pour $\alpha > 0$, la pénalité *elastic net* est strictement convexe (alors qu'elle est seulement convexe pour $\alpha = 0$) et que pour $\alpha \in [0, 1[$, la pénalité *elastic net* présente une singularité en 0.

Solution du *naïve elastic net*

Une méthode pour résoudre le problème du *naïve elastic net* a ensuite été proposée. L'idée est de reformuler le problème du *naïve elastic net* en un problème *lasso* que l'on sait résoudre.

On définit un nouveau couple de données (y^*, X^*) de la façon suivante :

$$\begin{aligned} X^*_{(n+p)\times p} &= \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I_p \end{pmatrix} \\ y^*_{n\times p} &= \begin{pmatrix} y \\ 0_p \end{pmatrix} \end{aligned}$$

avec les notations I_p matrice identité de taille p et 0_p vecteur de zéros de taille p.

On pose ensuite $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$ et $\beta^* = \sqrt{1+\lambda_2}\beta$. On a alors :

$$\begin{aligned} L(\lambda_1, \lambda_2, \beta) &= \|y - X\beta\|^2 + \lambda_2\|\beta_2\|^2 + \lambda_1\|\beta\|_1 \\ &= \|y - \frac{1}{\sqrt{1+\lambda_2}} X\beta^*\|^2 + \lambda_2\|\frac{1}{\sqrt{1+\lambda_2}}\beta^*\|^2 + \lambda_1\|\frac{1}{\sqrt{1+\lambda_2}}\beta\|_1 \end{aligned}$$

On remarque que :

$$\lambda_1 \| \frac{1}{\sqrt{1+\lambda_2}} \beta^* \|_1 = \gamma \|\beta^*\|_1$$

et

$$\lambda_2 \| \frac{1}{\sqrt{1+\lambda_2}} \beta^* \|^2 = \| 0_p - \frac{\sqrt{\lambda_2}}{\sqrt{1+\lambda_2}} I_p \beta^* \|^2$$

On obtient ainsi une pénalité *lasso* :

$$\begin{aligned} L(\lambda_1, \lambda_2, \beta) &= L(\gamma, \beta^*) \\ &= \|y^* - X^* \beta^*\|^2 + \gamma \|\beta^*\|_1 \end{aligned}$$

Si on note l'estimateur de ce problème *lasso* $\hat{\beta}^*$ tel que :

$$\hat{\beta}^* = \arg\min_{\beta} L(\gamma, \beta^*)$$

alors on obtient la solution $\hat{\beta}^{\text{nen}}$ de la façon suivante :

$$\hat{\beta}^{\text{nen}} = \frac{1}{\sqrt{1+\lambda_2}} \hat{\beta}^*.$$

Commentaires

On peut noter que dans le cas qui nous intéresse en grande dimension, c'est-à-dire $p > n$, la matrice X^* est de rang p, ce qui signifie que le *naïve elastic net* peut potentiellement sélectionner tous les prédicteurs p ce qui n'était pas le cas de la méthode *lasso* dont une des limitations était de ne pouvoir sélectionner qu'au plus n prédicteurs. L'utilisation de la méthode *lasso* met en évidence le fait que la régression *naïve elastic net* va permettre de faire de la sélection de variables de façon similaire au *lasso*.

On peut montrer également que contrairement au *lasso*, la régression *naïve elastic net* va permettre de sélectionner des variables corrélées. On se place pour cela dans le cadre d'une pénalité générique :

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda J(\beta)$$

où $J(.)$ est une pénalité positive pour $\beta \neq 0$.

Une méthode de régression permet de prendre en compte les corrélations entre les régresseurs si les coefficients de régression de variables fortement corrélées ont tendance à être égaux (au signe près selon le signe de la corrélation). En particulier, dans la situation extrême où les régresseurs sont exactement identiques, la méthode de régression doit assigner des coefficients identiques aux régresseurs. Supposons $X^i = X^j, i, j \in \{1, ..., p\}$. On a les résultats suivants :

1. si $J(.)$ est strictement convexe, alors $\hat{\beta}_i = \hat{\beta}_j, \forall \lambda > 0$;
2. si $J(\beta) = \|\beta\|_1$ ($J(.)$ est convexe), alors $\hat{\beta}_i \hat{\beta}_j \geq 0$ et $\hat{\beta}^*$ est un autre minimiseur du problème, avec

$$\hat{\beta}^*_k = \begin{cases} \hat{\beta}^*_k & \text{si } k \neq i \text{ et } k \neq j \\ (\hat{\beta}^*_i + \hat{\beta}^*_j).(s) & \text{si } k = i \\ (\hat{\beta}^*_i + \hat{\beta}^*_j).(1-s) & \text{si } k = j \end{cases}$$

pour tout $s \in [0, 1]$.

Le cas 1 correspond au cas de la régression *naïve elastic net* alors que le cas 2 correspond au cas de la régression *lasso*. On observe que la stricte convexité de la pénalité garantit le bon comportement face à des variables corrélées.

Définition de la régression *elastic net*

Des preuves empiriques permettent de montrer que la régression *naïve elastic net* n'est pas satisfaisante à moins qu'elle soit très proche de la régression *ridge* ou *lasso* (voir les sections 4 et 5 de l'article de Zou et Hastie [265]). Le problème vient du fait que la procédure consiste à appliquer successivement la pénalité *ridge* puis la pénalité *lasso*. Cette double opération a pour résultat d'introduire un biais supplémentaire sans pour autant réduire la variance de façon conséquente. Les performances de la régression peuvent être améliorées en corrigeant ce problème.

En reprenant les notations de la partie précédente, on pose le problème *naïve elastic net* comme un problème de type *lasso* :

$$\hat{\beta}^* = \arg\min_{\beta} \|y^* - X^*\beta^*\|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}\|\beta^*\|_1$$

La solution de l'*elastic net* corrigé est définie par

$$\hat{\beta}^{\text{en}} = \sqrt{1+\lambda_2}\hat{\beta}^*.$$

On rappelle que la solution du *naïve elastic net* était :

$$\hat{\beta}^{\text{nen}} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*$$

on a donc :

$$\hat{\beta}^{\text{en}} = (1+\lambda_2)\hat{\beta}^{\text{nen}}.$$

L'*elastic net* est une simple mise à l'échelle du *naïve elastic net*. Les propriétés de parcimonie et de comportement face aux variables corrélées ne sont pas affectées, les bonnes propriétés du *naïve elastic net* sont conservées pour l'*elastic net*.

4 Algorithme de régularisation

Algorithme LARS

L'algorithme LARS ([82]) est un outil développé afin d'implémenter la méthode du *lasso*. Il repose sur un algorithme de base, appelé LAR (Least Angle Regression), auquel s'ajoute des modifications mineures qui permettent d'implémenter le *lasso* (LARS) et l'*elastic net*. L'un des avantages de l'algorithme LARS par rapport à des algorithmes d'optimisation et qu'il permet de trouver le chemin entier, c'est-à-dire que quelque soit la valeur du paramètre de régularisation, on peut connaître les coefficients de régression.

Principe de l'algorithme LAR

L'algorithme LAR est très proche de l'algorithme de sélection de modèles de type *forward* (*forward stepwise* [122]) : partant d'un modèle sans variables, celui-ci ajoute les variables au modèle une par une ; à chaque étape, il identifie la meilleure variable (au sens d'un critère à déterminer, comme par exemple le critère BIC, AIC ou la plus faible *p-value* d'un test de Fisher) à inclure et met à jour l'estimateur des moindres carrés en prenant en compte cette nouvelle variable.

LAR agit de façon similaire : à chaque itération, l'algorithme identifie la variable X^j la plus corrélée avec les résidus ; il modifie alors le coefficient β_j associé à cette variable vers la valeur $\hat{\beta}_j^{MC}$ qu'une méthode des moindres carrés aurait calculé, sans toutefois l'atteindre : en effet modifier le coefficient vers cette valeur va avoir pour effet de diminuer la corrélation de cette variable avec les nouveaux résidus calculés (on parlera de résidus actualisés). De même les corrélations des autres variables vont également

varier, si bien qu'à un moment une des variables $X^k_{k\neq j}$ aura une corrélation égale à celle de X^j. Les coefficients β_j et β_k associés aux variables retenues sont alors modifiés conjointement vers la solution des moindres carrés, jusqu'à ce que comme précédemment une variable $X^l_{l\neq k\neq j}$ ait la même corrélation que X^k et X^j, et ainsi de suite jusqu'à ce que toutes les variables soient incluses dans le modèle.

L'algorithme suivant décrit le fonctionnement de LAR étape par étape :

1. Standardiser la matrice X afin qu'elle soit centrée et réduite. Commencer avec les résidus $R = y - \hat{y}$ où $\hat{y} = X\beta$ avec $\beta_1 = \beta_2 = ... = \beta_p = 0$.
2. Trouver le prédicteur X^j le plus corrélé avec R.
3. Modifier β_j de 0 vers son coefficient des moindres carrés jusqu'à ce qu'un autre prédicteur X^k ait la même corrélation avec les résidus actualisés que X^j.
4. Bouger β_j et β_k dans la direction définie par leur coefficient des moindres carrés des résidus actualisés R par $< X^j, X^k >$, jusqu'à ce qu'un autre prédicteur X^l ait la même corrélation avec les résidus actualisés.
5. Continuer de la même façon jusqu'à ce que les p prédicteurs soient pris en compte.

Le nombre d'itérations de cet algorithme est de $\min(n-1, p)$. En effet, si $p > n-1$, l'algorithme LAR arrive à des résidus nuls après $n-1$ étapes (-1 car les données sont centrées).
L'étape 4 peut être explicitée : supposons que l'algorithme se trouve au début de la $k^{\text{ième}}$ étape et que β^k soit le vecteur des coefficients pour ces variables à cette étape. Il y a donc $k-1$ valeurs non-négatives dans ce vecteur et la variable qui vient juste de rejoindre l'ensemble a un coefficient égal à 0. Si les résidus actualisés sont égaux à $R_k = y - X\beta^k$ alors la direction pour cette étape est

$$\delta_k = (X^T X)^{-1} X^T R_k$$

Le nouveau vecteur de coefficients est alors donné par $\beta^k(\alpha) = \beta^k + \alpha\delta_k$.

Formalisme mathématique

On note $\mathcal{A}$ un sous-ensemble d'indices $\{1, 2, ..., p\}$ et $X_\mathcal{A} = (...s_j X^j ...)_{j\in\mathcal{A}}$ où $s_j = \pm 1$. On note également

$$\mathcal{G}_\mathcal{A} = X^t_\mathcal{A} X_\mathcal{A} \text{ et } A_\mathcal{A} = (1^t_\mathcal{A} \mathcal{G}^{-1}_\mathcal{A} 1_\mathcal{A})^{-\frac{1}{2}}$$

avec $1_\mathcal{A}$ vecteur de 1 de taille $|\mathcal{A}|$, taille du vecteur $\mathcal{A}$. On remarque que $\mathcal{G}_\mathcal{A}$ représente la matrice de covariance de $\mathcal{A}$.
On définit un vecteur équi-angulaire comme :

$$u_\mathcal{A} = X_\mathcal{A} w_\mathcal{A} \text{ avec } w = A_\mathcal{A} \mathcal{G}^{-1}_\mathcal{A} 1_\mathcal{A}$$

w est le vecteur unitaire faisant des angles égaux inférieurs à 90 degrés avec les colonnes de $X_\mathcal{A}$:

$$X^t_\mathcal{A} u_\mathcal{A} = A_\mathcal{A} 1_\mathcal{A} \text{ et } \|u_\mathcal{A}\|^2 = 1.$$

À l'aide de ces notations, il est maintenant possible de décrire l'algorithme LAR. On pose $\hat{\mu_0} = 0_p$ et on construit $\hat{\mu}$ par étapes. Supposons que $\hat{\mu}_\mathcal{A}$ est l'estimation courante de l'algorithme LARS. On pose $\hat{c} = X^t(y - \hat{\mu}_\mathcal{A})$ le vecteur des corrélations entre $X_\mathcal{A}$ et les résidus actualisés. $\mathcal{A}$ correspond à l'ensemble des variables actives, c'est-à-dire les variables avec les plus grandes corrélations actualisées en valeur absolue :

$$\hat{C} = \max_j \{\hat{c}_j\} \text{ et } \mathcal{A} = \{j \ : |\hat{c}_j = \hat{C}\}$$

On pose $s_j = \text{signe}(\hat{c}_j)$ pour $j \in \mathcal{A}$. On calcule ensuite $X_\mathcal{A}$, $A_\mathcal{A}$ et $u_\mathcal{A}$ tels que définis précédemment et le produit scalaire $X^t u_\mathcal{A}$. $_\mathcal{A}$ donne la direction équi-angulaire dans laquelle l'algorithme va aller. L'étape suivante de l'algorithme LARS met à jour $\hat{\mu}_\mathcal{A}$:

$$\hat{\mu}_{\mathcal{A}_+} = \hat{\mu}_\mathcal{A} + \hat{\gamma} u_\mathcal{A}$$

avec

$$\hat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{A_\mathcal{A} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_\mathcal{A} + a_j} \right\}$$

$\min^+$ indique que le minimum est pris sur les composantes positives pour chaque indice j.

On peut tenter une explication de cette dernière équation. Posons $\mu(\gamma) = \hat{\mu}_\mathcal{A} + \gamma u_\mathcal{A}$. Pour $\gamma > 0$, la corrélation actualisée est

$$c_j(\gamma) = X_j^t (y - \mu(\gamma)) = \hat{c}_j - \gamma a_j$$

Pour $j \in \mathcal{A}$, on obtient :

$$|c_j(\gamma)| = \hat{C} - \gamma A_\mathcal{A}$$

ce qui permet de se rendre compte que le maximum des corrélations actualisées en valeur absolue diminue de façon identique. Pour $j \in \mathcal{A}^c$, on voit que $c_j(\gamma)$ égale la valeur maximal pour $\gamma = \frac{\hat{C} - \hat{c}_j}{A_\mathcal{A} + a_j}$. De manière similaire, $-c_j(\gamma)$, la corrélation actualisée pour la variable $-X^j$, atteint son maximum pour $\gamma = \frac{\hat{C} + \hat{c}_j}{A_\mathcal{A} + a_j}$. Ainsi le choix de $\hat{\gamma}$ est la plus petite valeur positive de γ tel qu'un nouvel indice $\hat{j}$ s'ajoute à $\mathcal{A}$. Le nouvel ensemble des variables actives $\mathcal{A}_+ = \mathcal{A} \cup \{\hat{j}\}$. Le nouveau maximum pour les corrélations actualisées est en valeur absolue $\hat{C}_+ = \hat{C} - \hat{\gamma} A_\mathcal{A}$.

Exemples

Exemple 1 La Figure A.5 illustre un exemple du comportement de l'algorithme LAR dans le cas où le nombre de régresseurs est égal à $m = 2$, $X = (X^1, X^2)$. Cet exemple permet de justifier le nom de la méthode, *Least Angular Regression*, de manière graphique.

On note $\hat{\mu} = X\hat{\beta}$, et $\hat{\mu_k} = X\hat{\beta^k}$ avec $\hat{\beta}$ le vecteur des coefficients de régression à l'itération k. On cherche à déterminer la projection $\hat{y}_3$ de y dans un espace de dimension 3 engendré par X^1 et X^2. L'algorithme démarre avec $\hat{\beta^0} = (0,0)^t$, c'est-à-dire $\hat{\mu_0} = (0,0)^t$. La première étape consiste à trouver la variable la plus corrélée avec les résidus. Graphiquement, cela revient à déterminer avec quelle variable les résidus $y - \hat{\mu_0}$ font le plus petit angle. Les résidus font un angle plus petit avec X^1 qu'avec X^2, c'est-à-dire que la corrélation des résidus est plus élevée avec X^1 qu'avec X^2. LAR augmente ainsi $\hat{\mu}$ dans la direction de X^1 de la façon suivante :

$$\hat{\mu}_1 = \hat{\mu}_0 + \hat{\gamma}_1 X_1$$

Le choix du $\hat{\gamma}_1$ est fait de façon à ce que les résidus $Y - \hat{\mu}_1$ soient corrélés de manière identique avec X^2. On a donc $\hat{Y}_2 - \hat{\mu}_1$ bissectrice de l'angle entre X^1 et X^2. On note u_2 le vecteur unité selon cette bissectrice. La prochaine estimation de LAR est

$$\hat{\mu}_2 = \hat{\mu}_1 + \hat{\gamma}_2 u_2$$

Le choix de $\hat{\gamma}_2$ se fait de sorte à minimiser les résidus $y - \hat{\mu}_2$.

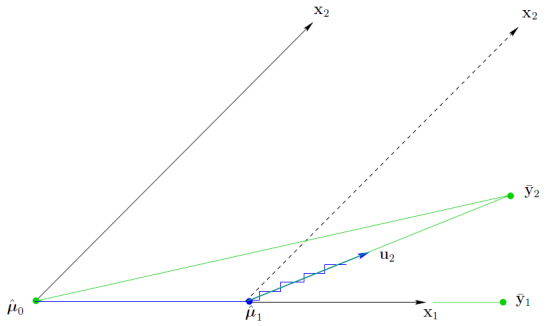


FIGURE A.5 – L'algorithme LAR avec 2 régresseurs ([98])

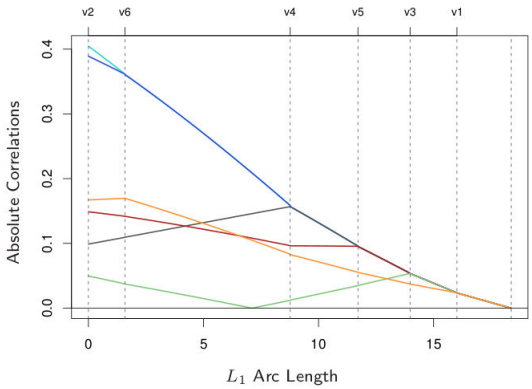


FIGURE A.6 – Évolution des valeurs absolues des corrélations à chaque étape de l'algorithme LAR avec 6 prédicteurs [98]

Exemple 2 La figure A.6 illustre l'évolution des valeurs absolues des corrélations avec les résidus actualisés au cours de l'exécution de l'algorithme LAR sur un échantillon avec 6 prédicteurs. De gauche à droite, les étiquettes en haut indiquent quelles variables sont retenues à chaque étape. On observe bien la décroissance des corrélations des variables actives avec les résidus et l'ajout d'une nouvelle variable active lors du croisement de ces corrélations.

La Figure A.7a illustre l'évolution des coefficients sur le même jeu de données lors de l'exécution de l'algorithme LAR. A chaque étape, une nouvelle variable est modifiée en plus des précédentes. On peut remarquer qu'une variable active n'a pas nécessairement de coefficient non-nul : le coefficient associé à la deuxième variable à devenir active (courbe bleue foncée) s'annule en effet dans la dernière étape de l'algorithme avant de changer de signe.

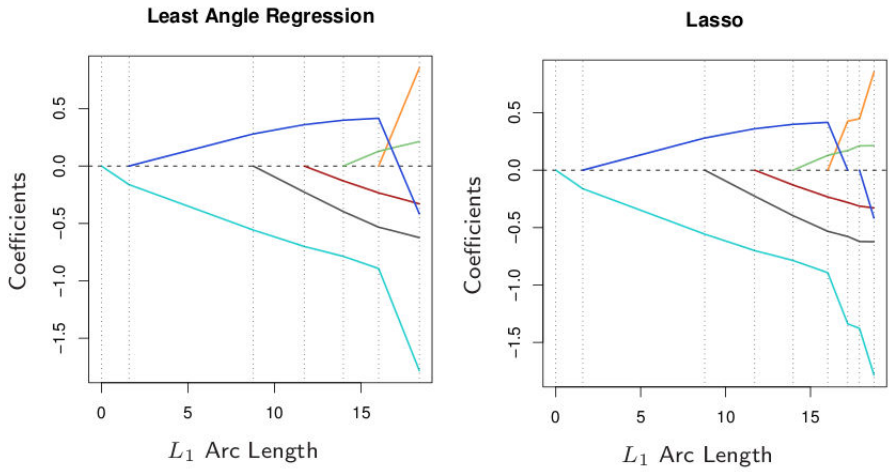
Modification de LAR pour implémenter le *lasso*

Afin d'implémenter la régression *lasso*, l'algorithme présenté précédemment suffit avec une modification, qui consiste à rajouter la contrainte suivante à chaque étape :

— si un coefficient non-nul atteint zéro, alors le prédicteur associé est retiré de l'ensemble des variables actives et une nouvelle direction est calculée.

Cet algorithme est un moyen efficace de réaliser une régression *lasso*. La Figure A.7b montre l'exécution de l'algorithme modifié sur l'exemple traité précédemment : nous voyons que le prédicteur dont le coefficient devenait nul est désactivé (son coefficient n'est plus modifié à l'étape d'après). Il recommence à être modifié lorsque le prédicteur est retenu comme lors d'une étape normale.

Cette modification se justifie en montrant que le signe de n'importe quel coefficient $\hat{\beta}_j$ différent de 0 doit être le même que s_j, signe de la corrélation actualisée de ce coefficient avec les résidus actualisés



(a) Algorithme LAR avec 6 prédicteurs (b) Algorithme LARS (*lasso*) avec 6 prédicteurs

FIGURE A.7 – Évolution des coefficients à chaque étape pour 6 prédicteurs. (a) Algorithme LAR. (b) Algorithme LARS (*lasso*). [98]

$\hat{c}_j = X^j(y - \hat{\mu})$. On a donc la condition suivante : signe$(\hat{\beta}_j) =$ signe$(\hat{c}_j) = s_j$ L'algorithme LAR ne respecte pas cette condition mais la modification permet de la lui faire respecter.

En effet, supposons que l'on a complété une étape LARS, et que l'on a $\mathcal{A}$ l'ensemble des variables actives. On suppose également que $\hat{\mu}_\mathcal{A}$ correspond à une solution du *lasso*. On a $w_\mathcal{A} = A\mathcal{A}\mathcal{G}\mathcal{A}^{-1}1\mathcal{A}$ un vecteur de longueur égale à la taille de l'ensemble $\mathcal{A}$. On note $\hat{d}$ la direction vers laquelle l'algorithme LARS va aller à la prochaine étape, on a donc $\hat{d}_j = s_j w\mathcal{A}j$ pour $j \in \mathcal{A}$ et $\hat{d}_j = 0$ sinon. La prochaine étape du LARS est :

$$\mu\gamma = X\beta(\gamma) \text{ avec } \beta_j(\gamma) = \hat{\beta}_j + \gamma\hat{d}_j$$

pour $j \in \mathcal{A}$. Le signe de $\beta_j(\gamma)$ va changer pour

$$\gamma_j = -\frac{\hat{\beta}_j}{\hat{d}_j}$$

ce qui arrive pour $\tilde{\gamma} = \min_{\gamma_j > 0}\{\gamma_j\}$ pour un régresseur $X^{\tilde{j}}$. Lorsqu'il n'y a aucun $\gamma_j > 0$, on prend $\tilde{\gamma}$ par définition.

Si $\tilde{\gamma} < \hat{\gamma}$, $\beta_j\gamma$ qui ne peut pas être une solution *lasso* pour $\gamma > \tilde{\gamma}$ puisque la condition sur les signes évoquée précédemment n'est pas respectée : $\beta_{\tilde{j}}(\gamma)$ a changé de signe alors que $c_{\tilde{j}}$ a gardé le même signe. En effet $c_{\tilde{j}}$ est une fonction continue qui ne peut pas changer de signe dans une étape LARS puisque $|c_{\tilde{j}(\gamma}| = \hat{C} - \gamma A_\mathcal{A} > 0$).

La modification pour implémenter le *lasso* est donc que si $\tilde{\gamma} < \hat{\gamma}$, alors l'algorithme LARS s'arrête pour $\gamma = \tilde{\gamma}$ et enlève $\tilde{j}$ pour calculer la prochaine direction équi-angulaire, c'est-à-dire :

$$\hat{\gamma}_{A_+} = \hat{\mu}_\mathcal{A} + \tilde{\gamma}u_\mathcal{A} \text{ et } \mathcal{A}_+ = \mathcal{A} - \{\tilde{j}\}.$$

Les solutions trouvées par l'algorithme LARS sont les solutions *lasso* exactes.

Modification de LAR pour implémenter le *lasso* positif

Imposer une contrainte de positivé sur les coefficients de régression peut être utile lorsque l'on souhaite développer des modèles génératifs. Il est très facile d'implémenter cette contrainte sur l'algorithme LARS.

Le problème est le suivant : on souhaite minimiser les résidus sous la contrainte $|\beta|_1 \leq t$ et pour tout $j \in \{1, .., p\}$, $\hat{\beta}_j \geq 0$.

La modification a effectuer est de remplacer la valeur absolue des corrélations $|\hat{c}|$ par les corrélations $\hat{c}$, de mettre $s_j = 1$ pour tout j et de changer l'étape où $\hat{gamma}$ était calculé par l'étape

$$\hat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{A_{\mathcal{A}} - a_j} \right\}$$

Modification de LAR pour implémenter l' *elastic net*

La résolution d'un problème *elastic net* se base sur la résolution d'un problème *lasso*. Il n'y a donc pas de modification à effectuer pour implémenter l'*elastic net*, si ce n'est effectuer la transformation comme précisé dans la partie détaillant l'*elastic net*. Il est néanmoins possible d'optimiser l'algorithme en prenant en compte que la matrice X^* a une structure parcimonieuse afin d'accélérer l'algorithme (voir [265] pour plus de détails).

5 Algorithmes proximaux

La résolution d'un problème du type *elastic net* implique la minimisation d'un critère convexe composé d'une somme de deux fonctions non nécessairement différentiables. Parmi la gamme des méthodes disponibles, les algorithmes proximaux (c'est-à-dire utilisant un opérateur proximal) possèdent des garanties de convergence et ont été utilisés pour implémenter l'*elastic net* ([68]). Sans entrer dans les détails techniques, une introduction à l'opérateur proximal et aux algorithmes proximaux permet de comprendre le principe de ces méthodes.

On se place dans le cadre d'un problème sous contrainte :

$$\hat{\beta} = \arg\min_{\beta} \left(f(\beta) + g(\beta) \right)$$

où f est une fonction différentiable et g est une fonction non-différentiable.

Opérateur proximal

Un opérateur proximal est associé à une fonction g de classe $\mathcal{H}$ en un point $u \in \mathcal{H}$ et se note $\text{prox}_g u$. Il correspond à l'unique point qui minimise $g + \|. + v\|^2$, c'est-à-dire :

$$\text{prox}_g \ : \mathcal{H} \to \mathcal{H} \ : u \mapsto \arg\min_{v \in \mathcal{H}} \frac{1}{2}\|v - u\|^2 + g(v)$$

Cet opérateur généralise la notion de projection P sur un ensemble convexe fermé non vide $C \subset \mathcal{H}$ telle que $\text{prox}_C = P_c$.

Dans le cas qui nous intéresse, l'opérateur proximal associé à une norme ℓ_1 est, pour $\lambda \in]0, +\infty[$, $g = \lambda|.|$ et $u \in \mathbb{R}$:

$$prox_g u \quad = \quad \begin{cases} u - x & \text{si } u > \lambda \\ 0 & \text{si } u \in [-\lambda, \lambda] \\ u + x & \text{si } u < -\lambda \end{cases}$$

Cet opérateur proximal correspond à un seuillage doux (*soft thresholding*), notion rencontrée lors de la présentation de la méthode *lasso*. On peut montrer que l'opérateur proximal est un opérateur contractant (si le pas de l'algorithme est convenable).

Algorithme proximal

Dans le cas où les fonctions f et f sont toutes les deux différentiables, la résolution du problème est simple. On calcule :

— $\beta_{n+\frac{1}{2}} = \beta_n - \gamma_n \nabla f$;

— $\beta_{n+1} = \beta_{n+\frac{1}{2}} - \alpha_n \nabla g$.

où γ_n et α_n sont les pas des descentes de gradient.

Le problème ici concerne la non différentiabilité de la fonction g.

L'idée d'un algorithme proximal est de chercher la solution au problème en deux étapes. A chaque itération :

1. on construit une approximation majorante de la fonction f ;
2. on construit localement la fonction g au point trouvé à l'étape 1 et on minimise la fonction g.

On a donc

$$\beta_{n+1} = \arg\min_{\beta} f_n(\beta) + g(\beta)$$

où f_n est une approximation majorante de la fonction g :

$$f_n(\beta) = f(\beta_n) + \nabla f(\beta_n)^t(\beta - \beta_n) + \frac{1}{2t_n}\|\beta - \beta_n\|^2.$$

On a ainsi

$$\beta_{n+1} = \arg\min_{\beta} f(\beta_n) + g(\beta) + \frac{1}{t_n}\|\beta - \beta_n + t_n \nabla f(\beta_n)\|^2 - \frac{t_n^2}{2t_n}\|\nabla f(\beta_n)\|^2$$

Les termes en β_n ne participent pas à la minimisation :

$$\begin{aligned} \beta_{n+1} &= \arg\min_{\beta} g(\beta) + \frac{1}{t_n}\|\beta - \beta_n + t_n \nabla f(\beta_n)\|^2 \\ &= \text{prox}_g(\beta_n - \gamma_n \nabla f(\beta_n)) \end{aligned}$$

pour une bonne valeur γ_n, pas de l'algorithme.

L'étude de convergence de cet algorithme est montrée dans ([193]). Il existe de nombreuses améliorations de cet algorithme dans la littérature , voir par exemple ([21]).

Une propriété intéressante des opérateurs proximaux est qu'il est aisé d'ajouter des contraintes. L'ajout par exemple d'une contrainte de positivité est une simple projection vers un convexe.

6 Prédiction des trajets Vélo'v

Modèle de régression linéaire

On souhaite expliquer les flux entrants et sortants des stations à différentes périodes de temps. On dispose pour cela des données évoquées précédemment auxquelles on rajoute un « intercept » (variable égale à 1 pour toutes les stations) et un indicateur « Part-Dieu » qui permet de prendre en compte le comportement atypique de la station située près de la gare. On intègre également les durées pendant lesquelles les stations sont vides ou pleines (on impose un coefficient négatif pour ces variables, qui agissent comme des inhibiteurs du flux). On a ainsi $n = 165$ le nombre d'individus et $p = 142$ le nombre de régresseurs. Ces données sont normalisées de la façon suivante pour chaque variable $j, j \in \{1, ..., p\}$:

$$\sum_{i=1}^{n} (X_i^j)^2 = 1.$$

Cette normalisation permet de comparer les coefficients entre eux afin de déterminer quels régresseurs ont le plus d'importance.

Le modèle linéaire que l'on souhaite développer nécessite de respecter deux types de contraintes particulières :

— contrainte de parcimonie : afin de faire de la sélection de variables ;
— contrainte de positivité : on cherche à ce que la présence d'une forte valeur pour une station d'une variable socio-économique entraîne plus de flux dans la station et non l'inverse, afin de faciliter l'interprétation. Le modèle est ainsi un modèle strictement génératif.

La méthode utilisée est la régression *elastic net* en utilisant un algorithme proximal. L'utilisation de cet algorithme au lieu de l'algorithme LARS se justifie par la contrainte de positivité que l'on souhaite obtenir. L'implémentation dans un algorithme proximal est une simple opération de projection qui garantit de bonnes propriétés de convergence vers la solution, alors que l'implémentation d'une telle contrainte dans l'algorithme LARS, si elle est possible et semble donner de bons résultats, n'a fait l'objet d'aucune recherche pour justifier la bonne convergence de la solution.

Cette partie est encore l'objet de recherches pour arriver d'une part à réaliser des modélisations correctes et d'autre part à interpréter les résultats. Les résultats présentés ci-dessus se concentrent sur la période « Jours de semaine - 7h à 9h », situation qui a été la plus étudiée et discutée.

Flux entrant - Jours de semaine 7h à 9h

Les résultats suivants ont été obtenus [2] :

```
|4|PLT99         |Pop 1999 active au lieu de travail               | 1115
|4|PLT99_CSP2    |Pop 1999 active au lieu de travail Artis. commerç| 563
|4|PLT99_CSP5    |Pop 1999 active au lieu de travail Employés      | 2807
|4|PLT99_CSP3    |Pop 1999 active au lieu de travail PIS           | 398
|4|PLT99_CSP4    |Pop 1999 active au lieu de travail Prof. Interm. | 1701
|2|S99_SUP18     |Population 1999 scolaire > 18 ans                | 1456
|2|M99_AM24      |Mén 99 <= 24 ans                                 | 271
|2|P99_A_CSP3    |Pop 99  15-29 ans PIS                            | 275
|1|P99_C_CSP3    |Pop 99  50 ans ou + PIS                          | 460
|2|Gs07_2003     |Etab. 2003 enseignement supérieur                | 2777
|5|Hotels        |Etab. 2007 hotels                                | 555
|1|Res99         |Résidences secondaires 1999                      | 599
|*|Part-Dieu     |Part-Dieu                                        | 6282
|*|-Altitude     |- Altitude de la station                         | 1727
|*|Capacite      |Capacite de la station                           | 5849
|*|-TimeCC Full |- Durée pendant laquelle la station a été pleine |
1828
|*|-TimeCC Empty|- Durée pendant laquelle la station a été vide   |
2024

R2 = 0.77
```

On remarque tout d'abord un bon ajustement malgré les contraintes imposées. Les commentaires que l'on peut faire sont :

— la présence de deux classes dominantes (au sens de la somme des coefficients), celle regroupant les variables « scolaire » (classe 2) et celle regroupant les variables « emploi » (classe 4) ;
— l'absence de l'« intercept » ;

2. La présence de (-) devant le nom de la variable indique que c'est l'opposé de la variable qui a été pris.

— la présence des deux variables concernant les contraintes de capacité.

L'interprétation que l'on peut faire de ces résultats est que les flux entrants semblent s'expliquer par sa proximité avec des établissements scolaires et professionnels : plus des variables ayant trait aux domaines scolaire ou professionnel sont élevées, plus le flux entrant dans la station sera important. L'absence de l'« intercept » peut s'expliquer par la présence de la capacité de la station qui permet de donner un flux moyen à la station en fonction de sa capacité.

Flux sortant - Jours de semaine 7h à 9h

Les résultats obtenus ont été obtenus :

```
|*|Intercept    |Intercept                                          | 995
|1|PA99_T2      |Pop 1999 active Marche à pied                      | 29
|1|PA99_T6      |Pop 1999 active Plusieurs modes                    | 853
|4|PLT99_CSP5   |Pop 1999 active au lieu de travail Employés        | 608
|1|IMM99        |Immeubles 1999                                     | 651
|2|M99_A2529    |Mén 99 25-29 ans                                   | 354
|2|M99_AM24     |Mén 99 <= 24 ans                                   | 1675
|2|P99_A_CSP3   |Pop 99  15-29 ans PIS                              | 1910
|3|P99_A_CSP4   |Pop 99  15-29 ans Prof. interm.                    | 484
|1|P99_B_CSP3   |Pop 99  30-49 ans PIS                              | 732
|2|Gs07_2003    |Etab. 2003 enseignement supérieur                  | 697
|5|Gc04_2006    |Etab. 2006 de date création non renseignée         | 508
|5|Hotels       |Etab. 2007 hotels                                  | 885
|1|Revm05       |Revenu moyen par mén. 06                           | 1155
|1|Res99        |Résidences secondaires 1999                        | 2630
|2|M99_voit0    |Mén 99 sans voiture                                | 324
|2|M99_csp8     |Mén 99 Inactifs                                    | 632
|1|M99_csp3     |Mén 99 PIS                                         | 486
|1|P99_csp3     |Pop 99 PIS                                         | 466
|2|Ev_men9905   |Evol. Mén. de 99 à 06                              | 278
|*|Part-Dieu    |Part-Dieu                                          | 8031
|*|-Altitude    |- Altitude de la station                           | 3159
|*|Capacite     |Capacite de la station                             | 3456
|*|-TimeCC Full |- Durée pendant laquelle la station a été pleine   |
22263
|*|-TimeCC Empty|- Temps pendant laquelle la station a été vide     | 818

R2 = 0.74
```

Ces résultats mettent en évidence, par rapport aux résultats précédents, de nouvelles classes dominantes pour expliquer les flux sortants, les classes 1 et 2. La classe 1 concerne les données de population. De plus, si la classe 2 représente des données scolaires, les variables sélectionnées concernent plutôt des données de population. La présence d'une forte population près des stations entraîne donc une hausse du flux sortant. On remarque l'apparition de l'« intercept » parmi les variables sélectionnées. Par rapport au cas précédent, on observe que le coefficient associé à la capacité de la station diminue. On peut donc supposer que ces deux variables interviennent pour donner un flux moyen à la station, auquel vont se rajouter des flux en fonction des régresseurs socio-économiques. La présence d'un fort coefficient pour la durée pendant laquelle la station a été pleine s'explique de la façon suivante : si la durée pendant laquelle la station a été pleine est élevée, alors cela signifie que le flux sortant est faible.

Discussion

Les modèles développés ont permis pour la première fois depuis le début de l'étude du système Vélo'v de mettre en évidence des différences d'explication entre les trajets entrants et sortants des stations pour la période « Jours de semaine - 7h à 9h ». Ce résultat est une avancée majeure et permet de valider les efforts effectués pour améliorer la qualité des données, que ce soit dans l'uniformisation des données au niveau de la station ou dans leur nettoyage pour éliminer les variables peu significatives. Retrouver des résultats qui sont cohérents avec les études déjà réalisées sur le système Vélo'v en sociologie permet d'acquérir de la confiance dans le modèle développé, et ainsi autoriser une analyse plus fine qui va permettre de trouver des effets pour des marqueurs sociologiques dont l'importance n'était pas attendue. Les pistes de poursuites possibles sont nombreuses : on peut déjà citer l'étude approfondie des autres périodes de la journée afin de déterminer s'il existe par exemple un effet de symétrie entre le matin et le soir. On peut également imaginer un travail sur les régresseurs plus poussé en amont de l'analyse, en proposant une classification des variables à la fois basée sur la corrélation mais également sur ce qu'elles représentent (variable de population, d'activité, etc.). Un travail sur ces régresseurs après l'analyse est également envisageable : on peut par exemple chercher à tester l'importance des régresseurs à l'aide de tests statistiques, en se basant sur les résultats de la sélection de variable mais également sur les intuitions de spécialistes du transport urbain. Les modèles linéaires développés se sont pour le moment concentré sur l'explication des flux entrants et sortants. On peut également envisager d'expliquer les durées pendant lesquelles les stations ont été sous contrainte de capacité pendant les périodes considérées. Il serait intéressant de voir si les variables socio-économiques permettent d'expliquer de façon significative le fait qu'une station soit vide ou pleine. Dans une autre optique, on peut également envisager le développement d'un modèle non-linéaire non-stationnaire pour expliquer les flux : on s'est en effet pour le moment contenté de considérer les durées d'activation de la contrainte de capacité comme de simples variables qui viendraient diminuer le flux. Il peut être intéressant d'arriver à développer un modèle linéaire qui prendrait en compte le fait que les contraintes de capacité vont tronquer la demande en vélos ou en emplacements.

ANNEXE B

Normalisation des profils des usagers Vélo'v

1 Describing users according to their practice of Bike sharing systems

Many cities have developed bicycle sharing system (BSS) program over recent years. These systems offer bikes that can be hired in any of the fully automated stations, spread over all the urban area, and returned at any other station. Records of all trips made by users highlight the activity over time and space of people in the city. In our research program an interdisciplinary approach is deployed to analyze BSS and its users in Lyon. Thanks to a partnership with the "Grand Lyon" City Hall and the operator Cyclocity, all the records of the Vélo'v system [1] in Lyon, France, were made available to us for the year 2011, as well as anonymized data about users. Most of recent research about BSS focuses on the stations [92]. According to a social science approach, we investigate users and users practices and first define for each user a profile based on the intensity of their use and their regularity of practice over the year and over the week. Then a typology of users is built from the profiles using the k-means clustering method [158].

The profile for each user is defined to quantify the intensity and the regularity of the use over the week and the year. For that, let us define the 21 following features, noted $X_i = (x_0^i, x_1^i, \ldots, x_{20}^i)$:

x_0^i Averaged number of trips per week calculated for all the weeks where the user has at least one trip.

$x_1^i, \ldots, x_7^i$ Average number of trips per day in the week, sorted in increasing order for working days $x_1^i \ldots x_5^i$ (x_5^i being the most intensive day).

x_8^i Total number of trips over the year.

$x_9^i \ldots x_{20}^i$ Number of trips over sorted months.

Because of the sorting, profiles are analyzed disregarding months or weekdays ranks. They only take into account the difference of use between weekends and weekdays.

This approach gives rise to methodological questions concerning the choice of the normalization of the profile. Comparison of three methods of normalization of profiles and their impact on the resulting clustering are discussed.

1. http://www.velov.grandlyon.com

2 Possible normalizations of users' profiles

Two remarks can be pointed out : first, the profiles do not provide no clearcut groups of users from which a typology arises. Second, each feature has a broad distribution, with many outliers or extreme values that can have high influence on obtained clustering.

Three normalizations of profiles are studied : all of them use the interquartile range IQR of a distribution as factor of normalization, which is equal to the difference between the lower and upper quartile of a distribution to set all values except outliers between 0 and 1.

— *Regular Normalization* : all features are normalized using the IQR value

$$\forall l \in \{0, \dots, 20\}, \qquad \bar{x}_l^i = \frac{x_l^i}{1.5 \times IQR(F_l)}$$

— *Maximal Normalization* : features x_0^i and x_8^i are computed as previously while others features are normalized using the maximal IQR value :

$$\forall l \in L,\ L \in \left\{ \begin{array}{l} \{1, \dots, 7\}, \\ \{9, \dots, 20\} \end{array} \right\}, \qquad \bar{x}_l^i = \frac{x_l^i}{1.5 \times \max\limits_{l' \in L}(IQR(F_l'))}$$

— *Proportional Normalization* : features x_0^i and x_8^i are computed as previously while for each user, day and month features are divided by their sum to get the distribution of the activity over the week and over the year

$$\forall l \in L,\ L \in \left\{ \begin{array}{l} \{1, \dots, 7\}, \\ \{9, \dots, 20\} \end{array} \right\}, \qquad \bar{x}_l^i = \frac{x_l^i}{\sum\limits_{l' \in L}(x_{l'}^i)}$$

Regular Normalization normalizes all features independently, while *Maximal Normalization* takes into account that orders of magnitudes of features are different between feature 0 (respectively 8) and features 1 to 7 (respectively 9 to 20). Finally *Proportional Normalization* removes the intensity for the features 1 to 7 (respectively 9 to 20), considering them as distribution describing regularity over the week (respectively over the year).

3 Clustering of users and discussion

We used the k-means algorithm [158] to cluster users according to their profiles. A preliminary study based on the clustering using only features 0 to 7 (practice over the week) and then features 8 to 20 (practice over the year), leads to 9 expected clusters, crossing both levels of intensity and regularity. For better optimization of the clusters, the algorithm is repeated 10 times with random initialization and the clustering which minimizes the within-cluster sum of squares is retained.

Table B.1 gives the percentage of variance explained by each principal components, for the six first components, obtained using PCA on profiles. For *Regular Normalization*, only one component explains 96% of variance, meaning that the features are not well normalized for clustering. Indeed, considering that feature 1 (less active weekday) and features 9, 10 and 11 (less active months) are often equal to 0 because most users have a day in the week and months where they are inactive. It induces high values for the highly intensive users who do not have inactive days or months, resulting in a gap too wide between small numbers of highly active users and the others. For the two others normalizations, percentages of the explained variances by principle components of PCA are more in line with what is expected for correct clustering.

For analysis, the clusters are sorted from $C0$ to $C8$ according to the average number of trips per year (feature 8). $C8$ is the group of the most intensive users.

Figure B.1 gives the profile means of each cluster for *Maximal Normalization* and *Proportional Normalization*, while Table B.2 gives the resulting table of contingency for groups with Maximal and Proportional Normalizations.

Regular Normalization	0.96	0.04	0.00	0.00	0.00	0.00
Maximal Normalization	0.79	0.09	0.03	0.02	0.02	0.02
Proportional Normalization	0.77	0.08	0.05	0.03	0.02	0.01

TABLE B.1 – Percentage of variance explained by each principal components using PCA, for the six first components

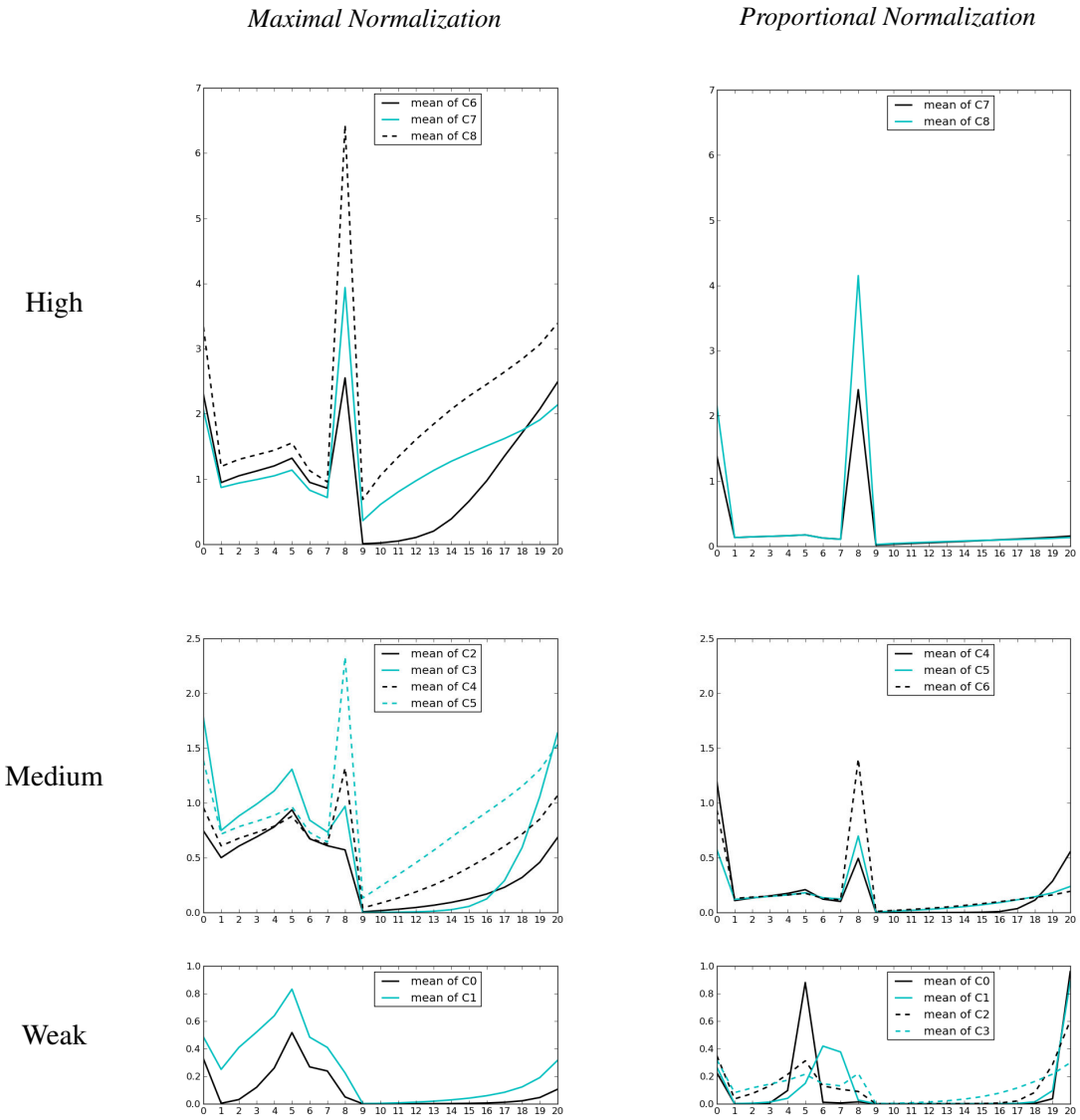


FIGURE B.1 – Cluster Means in 3 year intensity levels : High, Medium, Weak.

As expected, clusters are clearly separated according to intensity of use over the year (feature 8). For high and medium intensity profiles (covering more than 80% of movements), we observe that average behaviors are regular over the week (features 1 to 7) but more intense for weekdays, which makes sense.

Clustering using *Maximal Normalization* and *Proportional Normalization* return similar clusters (Table B.2), mainly differentiated by the features for intensity (features 0 and 8). Still, we observe some

differences in distribution of users. *Maximal Normalization* offers more clusters of high and medium intensity of use whereas clusters obtained using *Proportional Normalization* have more clusters with a lower intensity.

Cluster	Prop.	$C8$	$C7$	$C6$	$C5$	$C4$	$C3$	$C2$	$C1$	$C0$
Max.	size	560	2226	4749	8772	2602	16926	10146	2428	2080
$C8$	473	**470**	3							
$C7$	1595	88	**1507**							
$C6$	988	2	412	536	38					
$C5$	3257		304	**2953**						
$C4$	5785			1035	**4733**	17				
$C3$	2396			216	397	**1774**	2	7		
$C2$	11409				3633	718	5696	1340	22	
$C1$	14581				9	55	**9398**	4837	215	67
$C0$	9996						1830	3962	2191	2013

TABLE B.2 – Table of contingency between clustering generated using *Maximal Normalization* and *Proportional Normalization*

Maximal Normalization separates different types of medium intensity of uses during the year : C3 and C6 have months some off and some with intense use, $C2$, $C4$, $C5$, $C7$ and $C8$ have more regular uses in the months. This distinction is less clear in *Proportional Normalization* in clusters $C4$ to $C8$. However, considering weak intensity, *Proportional Normalization* leads to a separation between groups with regular uses during the week ($C2$, $C3$) or with limited uses during the weekends ($C1$) or only one weekday of use ($C0$).

These typologies show that the clustering results are impacted by the choice of the normalization of profiles. On the case of BSS' users, we probe this impact on the interpretation of the obtained topology. They all give us a precious knowledge about the variety of behaviors in transportation systems, whose further studies will refine social and economic characterization of BSS. Depending on the type of classes one wants to emphasize, one normalization or the other could be preferred., and the communication will show that the clustering with Proportional Normalization makes sense for a sociological analysis of the obtained typology of users.

ANNEXE C

Détections des problèmes de capacité des stations

Du fait de l'inégale répartition géographique des activités et des populations, les systèmes VLS sont sujets à des effets indésirables qui perturbent le bon fonctionnement : certaines zones de la ville, pour des raisons socio-économiques et géographiques, vont avoir tendance à attirer et retenir les gens, alors que d'autres vont au contraire être des zones que les usagers quittent. Une des manifestations les plus évidentes de ces dysfonctionnements est la présence de stations vides ou pleines : les stations étant de taille limitée, ainsi que la quantité de vélos en circulation, les stations les plus attractives se remplissent, n'offrant plus d'espace disponible pour déposer un vélo, alors que les stations les moins attractives se vident. La régulation des systèmes de vélos en libre-service est ainsi un enjeu majeur dans la maintenance et l'exploitation pour l'exploitant de tels systèmes, mais qui est complexe à cause de l'opposition entre les différents objectifs. Du point de vue des usagers, la régulation doit permettre d'éviter de se retrouver devant une station sans vélo lorsque l'on souhaite en louer un, ou bien dans un cas plus problématique, de se retrouver avec un vélo dans une station pleine sans possibilité de le déposer. Du point de vue de l'exploitant, l'objectif est de pouvoir mettre en place une régulation peu coûteuse afin de minimiser les dépenses, étant donné que le modèle économique ne repose pas sur le volume des mouvements, tout en garantissant un fonctionnement minimal du système défini contractuellement avec les institutions. Ces modèles de régulation sont néanmoins très souvent inadaptés à la réalité du terrain, la plupart du temps pour des raisons économiques : le fonctionnement optimal nécessite des moyens trop importants.

1 Détection des moments d'activation de la contrainte de capacité : approche naïve

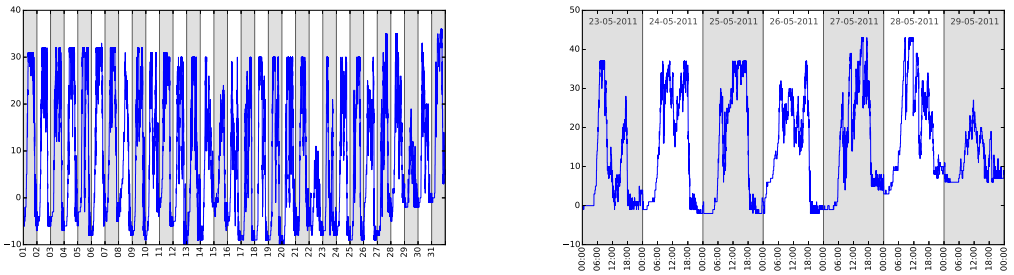
La détection de Une approche descriptive est proposée afin de détecter les moments pendant lesquels la contrainte de capacité agit sur une station. Il existe deux cas où la contrainte de capacité est activée : lorsque la station est pleine (toutes les bornettes sont sont occupées par un vélo) et lorsque la station est vide (toutes les bornettes sont libres, et donc aucun vélo n'est disponible). À partir des données "mouvements", l'état de chaque station est inconnu.

La détection des périodes d'activation de la contrainte de capacité va permettre d'une part de quantifier l'importance de ces saturations sur le fonctionnement du système Vélo'v et d'autre part de permettre de prendre en compte ces périodes dans une analyse statistique ; de telles périodes induisent en effet que la demande en Vélo'v n'est pas satisfaite : il en résulte que le flux de vélos observé est inférieur aux flux que l'on aurait observé si le système était optimal.

Préliminaire : construction de l'évolution du nombre de vélos par station

Une transformation des données brutes en une courbe d'évolution du nombre de vélos au cours du temps pour chaque station est réalisée, pour une période fixée que l'on souhaite étudier. Le nombre de vélo est fixé à zéro au début de la période considérée puis, à chaque fois qu'un vélo arrive dans la station, un incrément de 1 est réalisée sur l'évolution et réciproquement, à chaque fois qu'un vélo quitte la station, l'évolution est décrémentée de 1.

On note $E_s(t)$ le nombre de vélos de la station s au temps t. Intuitivement on peut considérer que l'évolution calculée correspond au nombre de vélos dans la station auquel on a soustrait le nombre de vélos initial dans la station qui est inconnu. En utilisant la notation introduite précédemment, on a $E_s(t) = N_s(t) - N_s(0)$ où $N_s(t)$ est le nombre de vélos présents au temps t dans la station s. En calculant E_s, on a ainsi une idée du comportement de N_s.Si on trace E_s en fonction du temps, on obtient graphiquement des courbes en escalier illustrant l'évolution du nombre de vélos dans la station, par rapport à la valeur initiale du nombre de vélos au début de la période. La figure C.1) donne un exemple pour la station 7016, pendant la période le mois de mai 2011.



(a) Mois de mai 2011

(b) Semaine du 23 au 29 mai 2011

FIGURE C.1 – Évolution du nombre de vélos dans la station 3001 pendant le mois de mai : chaque incrémentation correspond à une arrivée d'un vélo dans la station ; chaque décrémentation à un départ d'un vélo de cette station ; ces événements sont datés dans le temps

Algorithme « naïf » de détection

Un premier algorithme de détection des périodes d'activation de la contrainte de capacité est proposé : à partir de l'évolution des déplacements, il consiste à repérer le maximum et le minimum de E_s et de calculer l'écart entre ces deux valeurs. Si cette valeur est égale à la capacité de la station, alors le maximum correspond à un moment où la station est pleine et le minimum correspond à un moment où la station est vide. Ainsi tous les temps où E_s est égal au maximum ou au minimum de E_s sont a priori des périodes d'activation de la contrainte de capacité respectivement dans le cas où la station est pleine ou dans le cas où la station est vide (voir Algorithme9)

Cet algorithme est dit « naïf » car s'il décrit correctement les situations pendant lesquelles la station est vide ou pleine, à savoir des périodes pendant lesquelles le nombre de vélos est minimal ou maximal, il se révèle en pratique non adapté à l'analyse de données réelles. Il suppose en effet les hypothèses suivantes :

1. pendant la période de temps considérée, la station est au moins une fois vide et une fois pleine, de sorte que les extrémums correspondent au moment où la station est soit pleine soit vide ;
2. la capacité de la station ne varie pas pendant la période de temps considérée ;
3. il n'y a pas d'enregistrements manquants dans les données.

Algorithme 9 : Détection des contraintes de capacité : approche naïve

Entrées : $\{E_s(t)\}_{t\in\{0,\dots,T-1\}}$: Évolution du nombre de vélos pour une station s E_s pendant T pas de temps

Sorties :

début
  $M \leftarrow \max(E)$;
  $m \leftarrow \min(E)$;
  pour $t \in 0, \dots, T-1$ **faire**
    si $E_s(t) = M$ **alors** La station est pleine au temps t;
    si $E_s(t) = m$ **alors** La station est vide au temps t;

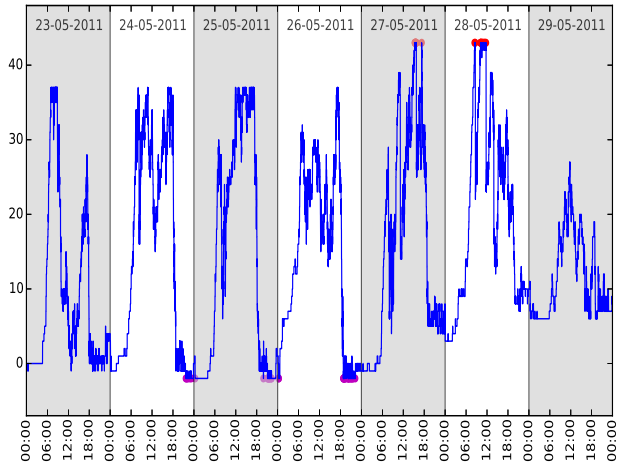


FIGURE C.2 – Exemple de représentation graphique de l'évolution : l'évolution moyenne croît au cours du temps ; l'écart entre le minimum et le maximum varie ; variation brusque de l'évolution

Chacune de ces trois hypothèses illustre les difficultés qui apparaissent lors de la détection des moments de contraintes de capacité. L'hypothèse 1 n'est pas réaliste pour un certain nombre de stations, qui soit sont très proches de l'équilibre, soit sont très souvent soit pleine soit vide, mais rarement les deux. L'hypothèse 2 peut paraître comme une hypothèse réaliste si l'on ne tient compte que de la capacité annoncée de la station : la ré-allocation de bornettes d'une station vers une autre est en effet possible mais reste ponctuel dans le temps et dans l'espace. En revanche, la capacité réelle de la station varie significativement au cours du temps, même à des échelles de temps courtes, à cause par exemple de bornettes cassées ou de vélos disponibles mais inutilisables. Enfin, l'hypothèse 3 n'est pas réaliste car comme tout système d'information, les données sont manquantes ou corrompues. Malgré la bonne qualité des données disponibles, la méthode n'est pas assez robuste.

La visualisation de la courbe d'évolution du nombre de vélos dans une station illustre (voir la figure C.2) les problèmes suivants :

1. la valeur moyenne de l'évolution n'est pas stationnaire de la valeur moyenne de l'évolution ;
2. écart entre le minimum et le maximum sur des périodes courtes (de l'ordre du jour) et longues (plusieurs semaines) non stationnaire ;
3. variations brutales de l'évolution ;

La deuxième cause explique de manière triviale le problème 2 : la variation de l'écart entre le maxi-

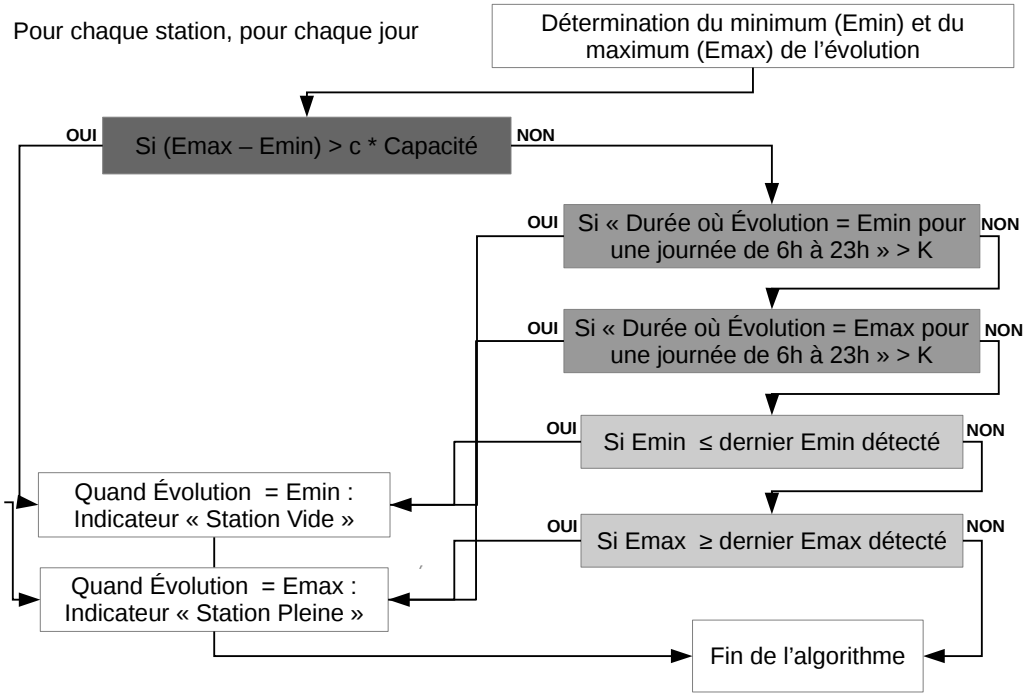


FIGURE C.3 – Algorithme de détection des périodes sous contraintes de capacité indiquant les trois critères successifs (gris foncé : critère 1 ; gris : critère 2 ; gris clair : critère 3)

mum et le minimum de l'évolution sur une période fixée, en supposant que le maximum et le minimum relevés sont des moments où la station a été pleine et vide, est liée à la capacité de station qui elle-même varie en fonction des aléas techniques (court-terme) et des modifications structurelles des stations (long-terme). Elle est facilement maîtrisable car elle n'entraîne pas de dérives dans les données ; l'algorithme proposée doit juste prendre en compte cette particularité.

La première cause explique les problèmes rencontrées et se traduit par la présence d'une dérive (ou tendance) sur les données difficilement corrigibles car ces problèmes ont des origines multiples : ajout de déplacements inexistants, mauvaises stations de départ et/ou d'arrivée, oubli de déplacements existants, etc. La dérive engendrée reste néanmoins assez faible et permet une exploitation des données en prenant en compte que les données ne sont pas parfaites.

2 Procédure multi-critères de détection des périodes d'activation de la contrainte de capacité

La procédure de détection des périodes d'activation de la contrainte de capacité se base sur l'analyse des différentes erreurs et des différentes situations pouvant survenir. Pour chacune des trois situations répertoriées, un critère de détection est proposée. Cette segmentation permet de prendre en compte la variété des comportements des déplacements des stations qui sont très hétérogènes. La procédure de détection agit station par station et jour par jour.

La Figure C.3 donne une vue d'ensemble de la procédure. Chaque critère s'applique successivement et est expliqué avec un exemple d'application dans les sous-sections suivantes.

Critère 1 : stations « équilibrées »

Hypothèse La méthode 1 prend pour hypothèse que pendant la période considérée, la station a des moments pendant lesquels elle est pleine et des moments pendant lesquels elle est vide dans la même journée. Si l'on se place dans une C'est le cas par exemple des stations proches des écoles, qui sont pleines le matin pour le début de la journée et qui sont vides le soir après la fin des cours pendant les jours de semaine.

Stratégie La stratégie employée consiste, sur chaque jour, à repérer la valeur minimale de l'évolution (eMin) et la valeur maximale de l'évolution (eMax) et de calculer l'écart eMax - eMin. Si cet écart est supérieur ou égal à la capacité de la station multipliée par une constante c comprise entre 0 et 1, alors les moments pendant lesquels l'évolution est minimale correspondent à des moments où la station est vide et les moments pendant lesquels l'évolution est maximale correspondent à des moments où la station est pleine.
Le choix de la constante c va permettre de prendre en compte des fluctuations de la capacité estimée (problème 3) et les erreurs d'enregistrements (problème 1) en augmentant la tolérance d'une station à ne pas respecter sa capacité réelle. L'idée de cette méthode reprend celle de l'algorithme « naïf » présenté dans la section précédente en intégrant la présence d'une baisse ponctuelle de la capacité de la station.

Exemple d'application La Figure C.4a donne un exemple de cas où la méthode 1 est efficace.

Critère 2 : stations « déséquilibrées »

La méthode 2 s'applique si la méthode 1 n'a pas détecté de saturations.

Hypothèse L'hypothèse retenue ici est que si une station reste longtemps à son état d'évolution minimal ou maximal pendant la journée, alors cet état d'évolution minimal ou maximal correspond à un moment pendant lequel la station est vide ou pleine. C'est le cas des stations qui sont déséquilibrées comme les stations situées en altitude (Croix-Rousse) qui se vident sans atteindre de maximum mais qui restent longtemps à l'état vide.

Stratégie La stratégie est la suivante : on calcule le temps passé par une station à son état minimal et maximal pendant la journée de 6h à 23h (pour ne pas comptabiliser la nuit où peu de déplacements ont lieu). Si un de ces temps est supérieur à une constante K, alors les moments pendant lesquels l'évolution est minimale ou maximale sont des moments pendant lesquels la station est vide ou pleine.

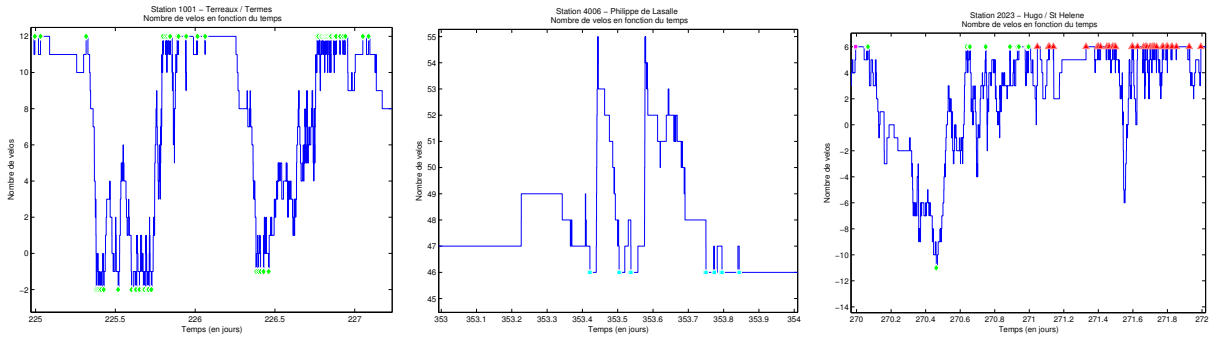
Exemple d'application La Figure C.4b donne un exemple de cas où la méthode 3 est efficace.

Critère 3 : effet mémoire

Hypothèse On prend pour hypothèse qu'à quelques jours d'intervalles, le minimum et le maximum de l'évolution ne vont pas fluctuer. Ainsi pour les stations où une saturation a été détectée un jour J mais où pour les jours suivants le critère 1 ou 2 ne s'applique pas, si le minimum de E_s ou le maximum de E_s est du même ordre que le minimum ou le maximum détecté au jour J, alors on dit qu'on a détection de saturation plein ou vide selon le cas rencontré.

Stratégie La stratégie consiste à partir du principe que si pendant le jour, il y a un minimum de l'évolution égal ou inférieur au dernier minimum détecté comme moment de saturation, alors ce minimum correspond également à un moment de saturation. De même pour le maximum s'il est supérieur ou égal au dernier maximum détecté comme moment de saturation.

Exemple d'application La Figure C.4c donne un exemple de cas où la méthode 2 est efficace.



(a) (b) (c)

FIGURE C.4 – Exemples de cas d'application des trois méthodes. (a) Critère 1 : l'écart entre le minimum et le maximum de l'évolution est de l'ordre de la capacité, ce qui entraîne la détection (carreaux verts). (b) Critère 2 : le temps pendant lequel l'évolution est égale à son minimum est important, ce qui entraîne la détection (carrés cyans). (c) Critère 3 : il utilise la détection sur le jour précédent (carreaux verts) pour détecter des contraintes de capacité (triangles rouges).

Liste de publications

Articles dans revues à comité de lecture

— M. Vogel, R. Hamon, G. Lozenguez, L. Merchez, P. Abry, J. Barnier, P. Borgnat, P. Flandrin, I. Mallon, and C. Robardet « From bicycle sharing system movements to users : a typology of Vélo'v cyclists in Lyon based on large-scale behavioural dataset » Journal of Transport Geography, Volume 41, Pages 280-291, Décembre 2014
http://www.sciencedirect.com/science/article/pii/S0966692314001537

— R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet « Discovering the structure of complex networks by minimizing cyclic bandwidth sum » IMA Journal of Complex Networks (soumis 2014, en révision)
http://arxiv.org/abs/1410.6108

— R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet « From graphs to signals and back : Identification of graph structures using spectral analysis » Physica A (soumis 2015, en révision)
http://arxiv.org/abs/1502.04697

— R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet « Duality between temporal networks and signals : Extraction of the temporal network structures » IEEE Transactions on Signal and Information Processing over Networks (soumis 2015)
http://arxiv.org/abs/1505.03044

Communications à conférences avec actes et comité de lecture

— R. Hamon, P. Borgnat, P. Flandrin, C. Robardet « Transformation de graphes dynamiques en séries temporelles non stationnaires » 24ème Colloque GRETSI sur le Traitement du Signal et des Images, Brest, 3-6 septembre 2013
https://hal-ens-lyon.archives-ouvertes.fr/ensl-00875085

— R. Hamon, P. Borgnat, P. Flandrin, C. Robardet « Tracking of a dynamic graph using a signal theory approach : application to the study of a bike sharing system », European Conference on Complex Systems, Barcelone (Espagne), 16-20 septembre 2013
https://hal-ens-lyon.archives-ouvertes.fr/ensl-00875187

— R. Hamon, P. Borgnat, P. Flandrin, C. Robardet « Networks as Signals, with an Application to Bike Sharing System », IEEE GlobalSIP 2014, Austin (Texas, USA), 1-4 décembre 2013
https://hal-ens-lyon.archives-ouvertes.fr/ensl-00875187

— R. Hamon, P. Borgnat, P. Flandrin, C. Robardet « Nonnegative matrix factorization to find features in temporal networks », IEEE-ICASSP 2014, Florence (Italie), 4-9 mai 2014.
https://hal-ens-lyon.archives-ouvertes.fr/ensl-00989760

— R. Hamon, P. Borgnat, C. Févotte, P. Flandrin, C. Robardet « Factorisation de réseaux temporels : étude des rythmes hebdomadaires du système Vélo'v », 25ème Colloque GRETSI sur le Traitement du Signal et des Images, Brest, 3-6 septembre 2013.

Communications orales à conférences sans actes

— R. Hamon, P. Borgnat, C. Robardet, P. Flandrin, « From dynamic graphs to nonstationary time series, and back», Workshop on Dynamic Graphs, Buenos Aires (Argentine), 22-23 novembre 2012
— R. Hamon, P. Borgnat, C. Robardet, P. Flandrin, « Networks as signals, with an application to bike sharing system », Workshop on Theoretical Foundations of Network Analysis, UCL, Londres (Royaume- Uni), 7-8 novembre 2013.
— G. Lozenguez, R. Hamon, J. Barnier, P. Abry, P. Borgnat, C. Robardet, M. Vogel, « Building a topology of bicycle sharing systems users : Impact of the normalization of profiles», International Symposia of Transport Simulation (ISTS) and International Workshop on Traffic Data Collection and its Standardisation (IWTDCS), Ajaccio (France), 1-4 Juin 2014.

Communications affichées à conférences sans actes

— R. Hamon, P. Borgnat, P. Flandrin, C. Robardet, « Transformation from dynamic graphs to nonstationary signals », « Network Dynamics » (Rencontres thématiques du GdR PHENIX), Montpellier, 25-26 mars 2013.
— R. Hamon, P. Borgnat, C. Févotte, P. Flandrin, C. Robardet, « Factorisation de réseaux temporels : étude des rythmes hebdomadaires du système Vélo'v », Assemblée Générale du GdR ISIS, Lyon, 30 mars-1er avril 2015.

Article grand public

— R. Hamon (entretien), « Les Vélo'v au cœur d'une thèse », Journal de bord des ARCs 2012/2013 - Région Rhône-Alpes, p. 40, 2013.

Bibliographie

[1] Annivelo'v 2015. www.annivelov.fr.

[2] ANR - Projet VEL'INNOV. http ://www.agence-nationale-recherche.fr/ ?Projet=ANR-12-SOIN-0001.

[3] IMU Intelligence des mondes urbains. http ://imu.universite-lyon.fr/.

[4] JCDecaux Developer. https ://developer.jcdecaux.com/.

[5] MoReBikeS : 2015 ECML-PKDD Challenge on "Model Reuse with Bike rental Station data", 2012.

[6] Hubway Data Visualization Challenge, 2013. http ://hubwaydatachallenge.org/.

[7] J. ADAMS : Collaborations : The rise of research networks. *Nature*, 490(7420):335–336, 2012.

[8] A. AGASKAR et Y. M. LU : A Spectral Graph Uncertainty Principle. *IEEE Transactions on Information Theory*, 59(7):4338–4356, juil. 2013.

[9] B. AGUILAR-SAN JUAN et L. Guzmán VARGAS : Earthquake magnitude time series : scaling behavior of visibility networks. *The European Physical Journal B*, 86(11):454, 2013.

[10] N. ALCOBA : Faced with legal and financial challenges, BIXI bike sharing program in a cycle of uncertainty. mai 2012.

[11] D. ALDOUS et J. FILL : *Reversible Markov chains and random walks on graphs*. Berkeley, 2002.

[12] J.-Y. AUTHIER, Y. GRAFMEYER, I. MALLON et M. VOGEL : *Sociologie de Lyon*. La Découverte, 2010.

[13] T. AYNAUD : *Détection de communautés dans les réseaux dynamiques*. Thèse de doctorat, PhD thesis, Docteur de L'université Pierre et Marie Curie, 2011.

[14] T. AYNAUD et J.-L. GUILLAUME : Long range community detection. *In LAWDN-Latin-American Workshop on Dynamic Networks*, p. 4–p, 2010.

[15] T. AYNAUD et J.-L. GUILLAUME : Static community detection algorithms for evolving networks. *In Modeling and optimization in mobile, ad hoc and wireless networks (WiOpt), 2010 proceedings of the 8th international symposium on*, p. 513–519. IEEE, 2010.

[16] J. BACHAND-MARLEAU, B. LEE et A. EL-GENEIDY : Better Understanding of Factors Influencing Likelihood of Using Shared Bicycle Systems and Frequency of Use. *Transportation Research Record : Journal of the Transportation Research Board*, 2314:66–71, déc. 2012.

[17] P. BALACHANDRAN, E. AIROLDI et E. KOLACZYK : Inference of Network Summary Statistics Through Network Denoising. *arXiv preprint arXiv :1310.0423*, 2013.

[18] R. BAR-YEHUDA, G. EVEN, J. FELDMAN et J. NAOR : Computing an optimal orientation of a balanced decomposition tree for linear arrangement problems. *Journal of Graph Algorithms and Applications*, 5(4):1–27, 2001.

[19] A. BARRAT, M. BARTHELEMY et A. VESPIGNANI : *Dynamical processes on complex networks*. Cambridge University Press, 2008.

[20] P. BASU, A. BAR-NOY, R. RAMANATHAN et M. P. JOHNSON : Modeling and analysis of time-varying graphs. *arXiv preprint arXiv :1012.0260*, 2010.

[21] A. BECK et M. TEBOULLE : A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Img. Sci.*, 2(1):183–202, mars 2009.

[22] R. BEECHAM et J. WOOD : Exploring gendered cycling behaviours within a large-scale behavioural data-set. *Transportation Planning and Technology*, 37(1):83–97, jan. 2014.

[23] M. BELKIN et P. NIYOGI : Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, 2003.

[24] B. BENBOUZID : De la prévention situationnelle au predictive policing : Sociologie d'une controverse ignorée. *Champ pénal*, (Vol. XII), juin 2015.

[25] C. D. BENCHIMOL : Introduction à l'imagerie médicale. *In 1er Colloque Image : traitement, synthèse, technologies et applications, FRA, 1984*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 1984.

[26] M. BENCHIMOL, P. BENCHIMOL, B. CHAPPERT, A. de la TAILLE, F. LAROCHE, F. MEUNIER et L. ROBINET : Balancing the stations of a self service "bike hire" system. *RAIRO-Oper. Res.*, 45(1):37–61, jan. 2011.

[27] B. BEROUD : Les expériences de vélos en libre service en Europe. *Transports urbains*, 111, 2007.

[28] A. BERTRAND et M. MOONEN : Seeing the Bigger Picture : How Nodes Can Learn Their Place Within a Complex Ad Hoc Network Topology. *Signal Processing Magazine, IEEE*, 30(3):71–82, 2013.

[29] V. D. BLONDEL, J.-L. GUILLAUME, R. LAMBIOTTE et E. LEFEBVRE : Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2008(10):P10008, oct. 2008.

[30] P. BODIK, W. HONG, C. GUESTRIN, S. MADDEN, M. PASKIN et R. THIBAUX : Intel lab data, 2004. http ://db.csail.mit.edu/labdata/labdata.html.

[31] F. BONIFAS : Vélo'v, 2005. Sous licence CC BY-SA 3.0 via Wikimedia Commons - https ://commons.wikimedia.org/wiki/File :Vélo'v.jpg.

[32] F. BONIFAS : Vélo'v station 5002 - Place des Compagnons de la chanson, 2005. Sous licence CC BY-SA 3.0 via Wikimedia Commons - https ://commons.wikimedia.org/wiki/File :Vélo'v_station_5002_-_Place_des_Compagnons_de_la_chanson.jpg.

[33] M. BORDAGARAY, A. IBEAS et L. DELL'OLIO : Modeling User Perception of Public Bicycle Services. *Procedia - Social and Behavioral Sciences*, 54:1308–1316, oct. 2012.

[34] I. BORG et P. J. F. GROENEN : *Modern Multidimensional Scaling*. Springer Series in Statistics. Springer, 2005.

[35] P. BORGNAT, P. ABRY, P. FLANDRIN, C. ROBARDET, J.-B. ROUQUIER et E. FLEURY : Shared bicycles in a city : a signal processing and data analysis perspective. *Advances in Complex Systems*, 14(03):415–438, 2011.

[36] P. BORGNAT, E. FLEURY, C. ROBARDET, A. SCHERRER *et al.* : Spatial analysis of dynamic movements of Vélo'v, Lyon's shared bicycle program. *In European Conference on Complex Systems 2009*, 2009.

[37] P. BORGNAT, C. ROBARDET, P. ABRY, P. FLANDRIN, J.-B. ROUQUIER et N. TREMBLAY : A Dynamical Network View of Lyon's Vélo'v Shared Bicycle System. *In* A. MUKHERJEE, M. CHOUDHURY, F. PERUANI, N. GANGULY et B. MITRA, éds : *Dynamics On and Of Complex Networks, Volume 2*, Modeling and Simulation in Science, Engineering and Technology, p. 267–284. Springer New York, 2013.

[38] L. BORTOLUSSI et J. HILLSTON : Efficient Checking of Individual Rewards Properties in Markov Population Models. *In roceedings for the 13th Workshop on Quantitative Aspects of Programming Languages*, London (United Kingdom), 2015.

[39] C. BOUVEYRON, E. CÔME, J. JACQUES *et al.* : The Discriminative Functional Mixture Model for the Analysis of Bike Sharing Systems. 2014.

[40] D. BOYD et K. CRAWFORD : Critical questions for Big Data : Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679, juin 2012.

[41] D. BRAHA et Y. BAR-YAM : Time-Dependent Complex Networks : Dynamic Centrality, Dynamic Motifs, and Cycles of Social Interactions. *In* T. GROSS et H. SAYAMA, éds : *Adaptive Networks*, Understanding Complex Systems, p. 39–50. Springer Berlin Heidelberg, 2009.

[42] D. BUCK, R. BUEHLER, P. HAPP, B. RAWLS, P. CHUNG et N. BORECKI : Are Bikeshare Users Different from Regular Cyclists ? *Transportation Research Record : Journal of the Transportation Research Board*, 2387:112–119, déc. 2013.

[43] I. CABANNE : Les coûts et les avantages des vélos en libre service. Rap. tech., 2010.

[44] L. CAGGIANI et M. OTTOMANELLI : A Dynamic Simulation based Model for Optimal Fleet Repositioning in Bike-sharing Systems. *Procedia - Social and Behavioral Sciences*, 87:203–210, oct. 2013.

[45] T. CALAMONERI, A. MASSINI, L. TÖRÖK et I. VRTO : Antibandwidth of Complete k-Ary Trees. *Electronic Notes in Discrete Mathematics*, 24:259–266, 2006.

[46] A. S. L. O. CAMPANHARO, M. I. SIRER, R. D. MALMGREN, F. M. RAMOS et L. A. N. AMARAL : Duality between Time Series and Networks. *PloS one*, 6(8), 2011.

[47] A. CASTEIGTS, P. FLOCCHINI, W. QUATTROCIOCCHI et N. SANTORO : Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 27(5):387–408, 2012.

[48] C. CASTELLUCCIA, S. GRUMBACH et L. OLEJNIK : Data harvesting 2.0 : from the visible to the invisible web. *In The Twelfth Workshop on the Economics of Information Security*, 2013.

[49] J. I. CASTILLO-MANZANO, M. CASTRO-NUÑO et L. López VALPUESTA : Analyzing the transition from a public bicycle system to bicycle ownership : A complex relationship. *Transportation Research Part D : Transport and Environment*, 38:15–26, juil. 2015.

[50] R. CAZABET : *Détection de communautés dynamiques dans des réseaux temporels*. Thèse de doctorat, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2013.

[51] L. CERF, T. B. N. NGUYEN et J.-F. BOULICAUT : Discovering relevant cross-graph cliques in dynamic networks. *In Foundations of Intelligent Systems*, p. 513–522. Springer, 2009.

[52] C. CHAMLEY, A. SCAGLIONE et L. LI : Models for the Diffusion of Beliefs in Social Networks : An Overview. *Signal Processing Magazine, IEEE*, 30(3):16–29, 2013.

[53] C. CHAN : Découverte de connaissances et Data Mining. *Gizmodo*, 2013. http ://gizmodo.com/the-nsa-and-fbi-have-been-spying-on-our-internet-habits-511750202.

[54] S. CHAUHAN, M. GIRVAN et E. OTT : Spectral properties of networks with community structure. *Physical Review E*, 80(5), nov. 2009.

[55] Q. CHEN et T. SUN : A model for the layout of bike stations in public bike-sharing systems : A Model for the Layout of Bike Stations. *Journal of Advanced Transportation*, p. n/a–n/a, mai 2015.

[56] Y.-D. CHEN et J.-H. YAN : A study on cyclic bandwidth sum. *Journal of Combinatorial Optimization*, 14(2):295–308, 2007.

[57] F. R. CHUNG : Labelings of graphs. *In Selected topics in graph theory*, vol. 3, p. 151–168. 1988.

[58] F. R. CHUNG : *Spectral graph theory*, vol. 92. American Mathematical Soc., 1997.

[59] J. CLARK et D. A. HOLTON : *A first look at graph theory*. World Scientific, Singapore, 1991.

[60] V. COLIZZA, A. BARRAT, M. BARTHÉLEMY et A. VESPIGNANI : The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–2020, 2006.

[61] E. CÔME et L. OUKHELLOU : Model-Based Count Series Clustering for Bike Sharing System Usage Mining : A Case Study with the Vélib' System of Paris. *ACM Trans. Intell. Syst. Technol.*, 5(3):39 :1–39 :21, juil. 2014.

[62] S. COOK, C. CONRAD, A. L. FOWLKES et M. H. MOHEBBI : Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1n1) Pandemic. *PLoS ONE*, 6(8):e23610, août 2011.

[63] L. d. F. COSTA, F. A. RODRIGUES, G. TRAVIESO et P. R. VILLAS BOAS : Characterization of complex networks : A survey of measurements. *Advances in Physics*, 56(1):167–242, jan. 2007.

[64] E. CUTHILL et J. MCKEE : Reducing the bandwidth of sparse symmetric matrices. *In Proceedings of the 1969 24th national conference*, p. 157–172, 1969.

[65] M. DANISCH, J.-L. GUILLAUME et B. LE GRAND : Towards multi-ego-centred communities : a node similarity approach. *International Journal of Web Based Communities*, 9(3):299–322, jan. 2013.

[66] K. C. DAS : Sharp bounds for the sum of the squares of the degrees of a graph. *Kragujevac J. Math*, 25:31–49, 2003.

[67] C. C. de CYBERSTRATÉGIE : Colloque « Le monde après Snowden », 2014. http ://www.cyberstrategie.org/ ?q=fr/event/monde-apres-snowden.

[68] C. DE MOL, E. DE VITO et L. ROSASCO : Elastic-net regularization in learning theory. *J. Complex.*, 25(2):201–230, avr. 2009.

[69] Y. DEKEL, J. R. LEE et N. LINIAL : Eigenvectors of random graphs : Nodal Domains. *Random Structures & Algorithms*, 39(1):39–58, 2011.

[70] L. DELL'OLIO, A. IBEAS et J. L. MOURA : Implementing bike-sharing systems. *Proceedings of the ICE-Municipal Engineer*, 164(2):89–101, 2011.

[71] P. DEMAIO : Bike-sharing : History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4):41–56, 2009.

[72] A. des SCIENCES : Colloque « La Datamasse : directions et enjeux pour les données massives », 2014. http ://www.academie-sciences.fr/video/v180214.htm.

[73] A. des SCIENCES : Colloque « à la recherche du temps ». 2015. http ://www.academie-sciences.fr/video/v190515.htm.

[74] E. DESMIER, M. PLANTEVIT, C. ROBARDET et J.-F. BOULICAUT : Granularity of Co-evolution Patterns in Dynamic Attributed Graphs. *In Advances in Intelligent Data Analysis XIII*, p. 84–95. Springer, 2014.

[75] J. DIAZ, J. PETIT et M. SERNA : A survey of graph layout problems. *ACM Comput. Surv.*, 34(3):313–356, 2002.

[76] J. DILL et N. MCNEIL : Four Types of Cyclists ? *Transportation Research Record : Journal of the Transportation Research Board*, 2387:129–138, déc. 2013.

[77] J. F. DONGES, J. HEITZIG, R. V. DONNER et J. KURTHS : Analytical framework for recurrence network analysis of time series. *Physical Review E*, 85(4), 2012.

[78] R. V. DONNER, M. SMALL, J. F. DONGES, N. MARWAN, Y. ZOU, R. XIANG et J. KURTHS : Recurrence-based time series analysis by means of complex network methods. *International Journal of Bifurcation and Chaos*, 21(04):1019–1046, 2011.

[79] R. DOORLEY, V. PAKRASHI et B. GHOSH : Quantifying the Health Impacts of Active Travel : Assessment of Methodologies. *Transport Reviews*, p. 1–24, mai 2015.

[80] J. DUCH et A. ARENAS : Scaling of Fluctuations in Traffic on Complex Networks. *Physical Review Letters*, 96(21), juin 2006.

[81] I. S. DUFF, R. G. GRIMES et J. G. LEWIS : *Users' Guide for the Harwell-Boeing Sparse Matrix Collection (Release I)*. 1992.

[82] B. EFRON, T. HASTIE, L. JOHNSTONE et R. TIBSHIRANI : Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[83] B. ESCUDIÉ et J. RAMUNNI : Une histoire du traitement du signal. *Traitement du Signal*, 6(3): 151–152, 1989.

[84] A. FAGHIH-IMANI, N. ELURU, A. M. EL-GENEIDY, M. RABBAT et U. HAQ : How land-use and urban form impact bicycle flows : evidence from the bicycle-sharing system (BIXI) in Montreal. *Journal of Transport Geography*, fév. 2014.

[85] W. FAN et A. BIFET : Mining big data : current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*, 14(2):1–5, 2013.

[86] H. FANAEE-T et J. GAMA : Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, juin 2014.

[87] A. FERREIRA : Building a reference combinatorial model for MANETs. *Network, IEEE*, 18(5): 24–29, oct. 2004.

[88] C. FÉVOTTE : Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*, p. 1980–1983. IEEE, 2011.

[89] C. FÉVOTTE, N. BERTIN et J.-L. DURRIEU : Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

[90] C. FÉVOTTE et J. IDIER : Algorithms for Nonnegative Matrix Factorization with the Beta-Divergence. *Neural Computation*, 23(9):2421–2456, juin 2011.

[91] E. FISHMAN : Bikeshare : A Review of Recent Literature. *Transport Reviews*, p. 1–22, avr. 2015.

[92] E. FISHMAN, S. WASHINGTON et N. HAWORTH : Bike Share : A Synthesis of the Literature. *Transport Reviews*, 33(2):148–165, mars 2013.

[93] E. FISHMAN, S. WASHINGTON et N. L. HAWORTH : An evaluation framework for assessing the impact of public bicycle share schemes. 2012.

[94] P. FLANDRIN : Représentations temps-fréquence des signaux non-stationnaires. *Traitement du Signal*, 6(2), 1989.

[95] P. FLANDRIN, M. SIDAHMED et D. GARREAU : Traitement du signal en mécanique. *Traitement du Signal*, p. 289–290, 1991.

[96] P. FLOCCHINI, M. KELLETT, P. MASON et N. SANTORO : Searching for Black Holes in Subways. *Theory of Computing Systems*, 50(1):158–184, jan. 2012.

[97] S. FORTUNATO : Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[98] J. FRIEDMAN, T. HASTIE et R. TIBSHIRANI : *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001.

[99] J. FROEHLICH, J. NEUMANN et N. OLIVER : Measuring the pulse of the city through shared bicycle programs. *Proc. of UrbanSense08*, p. 16–20, 2008.

[100] J. FROEHLICH, J. NEUMANN et N. OLIVER : Sensing and predicting the pulse of the city through shared bicycling. *In Proceedings of the 21st international jont conference on Artifical intelligence*, p. 1420–1426, 2009.

[101] D. FULLER, L. GAUVIN, Y. KESTENS, M. DANIEL, M. FOURNIER, P. MORENCY et L. DROUIN : Use of a New Public Bicycle Share Program in Montreal, Canada. *American Journal of Preventive Medicine*, 41(1):80–83, juil. 2011.

[102] R. GALLAGHER : ICREACH : How the NSA Built Its Own Secret Google -. 2014.

[103] R. GALLOTTI et M. BARTHELEMY : The multilayer temporal network of public transport in Great Britain. *Scientific Data*, 2:140056, jan. 2015.

[104] J. C. GARCÍA-PALOMARES, J. GUTIÉRREZ et M. LATORRE : Optimizing the location of stations in bike-sharing programs : A GIS approach. *Applied Geography*, 35(1-2):235–246, nov. 2012.

[105] A. GAUTHIER, C. HUGHES, C. KOST, S. LI et OTHERS : The Bike-share Planning Guide. ITDP Report. Rap. tech., 2013.

[106] L. GAUVIN, A. PANISSON et C. CATTUTO : Detecting the Community Structure and Activity Patterns of Temporal Networks : A Non-Negative Tensor Factorization Approach. *PLoS ONE*, 9(1):e86028, 2014.

[107] J.-F. GÉRARD : Le Vélhop plait un peu trop, la CUS dépassée et à court d'idée. *Rue89 Strasbourg*, oct. 2014.

[108] S. GEZICI, Z. TIAN, G. GIANNAKIS, H. KOBAYASHI, A. MOLISCH, H. POOR et Z. SAHINOGLU : Localization via ultra-wideband radios : a look at positioning aspects for future sensor networks. *Signal Processing Magazine, IEEE*, 22(4):70–84, juil. 2005.

[109] B. GIRAULT, P. GONÇALVES et E. FLEURY : Signaux stationnaires sur graphe : étude d'un cas réel. *In Gretsi*, 2015.

[110] B. GIRAULT, P. GONÇALVES, E. FLEURY et A. S. MOR : Semi-Supervised Learning for Graph to Signal Mapping : a Graph Signal Wiener Filter Interpretation. *In IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, p. 1115–1119, Florence, Italy, 2014.

[111] M. GIRVAN et M. E. NEWMAN : Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[112] G. GONTHIER : Formal proof–the four-color theorem. *Notices of the AMS*, 55(11):1382–1393, 2008.

[113] GOOGLE.ORG : Google Flu Trends, 2009. http ://www.google.org/flutrends/.

[114] J. C. GOWER : Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, 1985.

[115] R. M. GRAY : Toeplitz and circulant matrices : A review. *Communications and Information Theory*, 2(3):155–239, 2005.

[116] H. GUILLAUD : Police prédictive : la prédiction des banalités | InternetActu, juin 2015. http ://internetactu.blog.lemonde.fr/2015/06/27/police-predictive-la-prediction-des-banalites/.

[117] D. K. HAMMOND, P. VANDERGHEYNST et R. GRIBONVAL : Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, mars 2011.

[118] S. HANNEKE, W. FU et E. P. XING : Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4(0):585–605, 2010.

[119] Y. HARAGUCHI, Y. SHIMADA, T. IKEGUCHI et K. AIHARA : Transformation from complex networks to time series using classical multidimensional scaling. *In Artificial Neural Networks–ICANN 2009*, p. 325–334. Springer, 2009.

[120] F. HARARY et G. GUPTA : Dynamic graph models. *Mathematical and Computer Modelling*, 25(7):79–87, avr. 1997.

[121] F. HÉRAN : Vélo et politique globale de déplacements durables. *In Convention Prédit nh*, vol. 9, p. 114, 2012.

[122] R. R. HOCKING : The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32(1):1–49, 1976.

[123] N. HOE et T. KALOUSTIAN : Bike Sharing in Low-Income Communities : An Analysis of Focus Groups Findings Fall 2014. 2014.

[124] P. HOLME et J. SARAMÄKI : Temporal Networks. *Physics reports*, 519:97–125, 2012.

[125] P. HOLME et J. SARAMÄKI : Temporal Networks as a Modeling Framework. *In* P. HOLME et J. SARAMÄKI, éds : *Temporal Networks*, Understanding Complex Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[126] J. HOPCROFT, O. KHAN, B. KULIS et B. SELMAN : Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5249–5253, 2004.

[127] B. A. HUBERMAN et L. A. ADAMIC : Internet : Growth dynamics of the World-Wide Web. *Nature*, 401(6749):131–131, sept. 1999.

[128] S. HUET : La note de l'INRIA contre la loi sur le renseignement. *Libération*, 2015. http ://sciences.blogs.liberation.fr/home/2015/05/la-note-de-linria-contre-la-loi-sur-le-renseignement.html.

[129] INSEE : http ://www.insee.fr/fr/.

[130] P. JACCARD : *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[131] R. K. JAIN, J. M. MOURA et C. E. KONTOKOSTA : Big Data + Big Cities : Graph Signals of Urban Air Pollution [Exploratory SP]. *IEEE Signal Processing Magazine*, 31(5):130–136, sept. 2014.

[132] M. B. JDIDIA, C. ROBARDET et E. FLEURY : Communities detection and analysis of their dynamics in collaborative networks. *In ICDIM*, p. 744–749, 2007.

[133] M. JENSEN : Passion and heart in transport—a sociological analysis on transport behaviour. *Transport Policy*, 6(1):19–33, 1999.

[134] P. JENSEN, J.-B. ROUQUIER, N. OVTRACHT et C. ROBARDET : Characterizing the speed and paths of shared bicycle use in Lyon. *Transportation Research Part D : Transport and Environment*, 15(8):522–524, 2010.

[135] H. JIANXIU : Cyclic bandwidth sum of graphs. *Applied Mathematics-A Journal of Chinese Universities*, 16(2):115–121, 2001.

[136] A. KALTENBRUNNER, R. MEZA, J. GRIVOLLA, J. CODINA et R. BANCHS : Urban cycles and mobility patterns : Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, août 2010.

[137] S. KAR et J. M. MOURA : Consensus + innovations distributed inference over networks : cooperation and sensing in networked systems. *Signal Processing Magazine, IEEE*, 30(3):99–109, 2013.

[138] B. KARRER et M. E. J. NEWMAN : Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, jan. 2011.

[139] T. KATO : *Perturbation theory for linear operators*. Springer Science & Business Media, 1966.

[140] D. KEMPE, J. KLEINBERG et A. KUMAR : Connectivity and Inference Problems for Temporal Networks. *In Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC '00, p. 504–513, New York, NY, USA, 2000. ACM.

[141] D. KIM, H. SHIN, H. IM et J. PARK : Factors Influencing Travel Behaviors in Bikesharing. *In Transportation Research Board 91st Annual Meeting*, 2012.

[142] R. KITCHIN : Big data and human geography : Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3):262–267, nov. 2013.

[143] C. KLOIMÜLLNER, P. PAPAZEK, B. HU et G. R. RAIDL : Balancing bicycle sharing systems : An approach for the dynamic case. *In Evolutionary Computation in Combinatorial Optimisation*, p. 73–84. Springer, 2014.

[144] V. KOSTAKOS : Temporal graphs. *Physica A : Statistical Mechanics and its Applications*, 388(6): 1007–1023, mars 2009.

[145] L. KOVANEN, M. KARSAI, K. KASKI, J. KERTÉSZ et J. SARAMÄKI : Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2011(11): P11005, nov. 2011.

[146] G. KRYKEWYCZ, C. PUCHALSKY, J. ROCKS, B. BONNETTE et F. JASKIEWICZ : Defining a Primary Market and Estimating Demand for Major Bicycle-Sharing Program in Philadelphia, Pennsylvania. *Transportation Research Record : Journal of the Transportation Research Board*, 2143:117–124, déc. 2010.

[147] N. LATHIA, S. AHMED et L. CAPRA : Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research Part C : Emerging Technologies*, 22:88–102, juin 2012.

[148] LDA CONSULTING : Capital Bikeshare - Survey Report. Rap. tech., 2011.

[149] D. LE MÉTAYER et C. CASTELLUCCIA : Oui, la loi sur le renseignement prépare bien une surveillance de masse. *Le Monde.fr*, 2015. http ://www.lemonde.fr/idees/article/2015/06/09/oui-la-loi-sur-le-renseignement-prepare-bien-une-surveillance-de-masse_4650358_3232.html.

[150] J. LESKOVEC et A. KREVL : SNAP Datasets : Stanford Large Network Dataset Collection. 2014.

[151] J.-R. LIN et T.-H. YANG : Strategic design of public bicycle sharing systems with service level constraints. *Transportation Research Part E : Logistics and Transportation Review*, 47(2):284–294, mars 2011.

[152] Y. LIN : The cyclic bandwidth problem. *Systems Sci. Math. Sci*, 7:282–288, 1994.

[153] G. F. LOHRY et A. YIU : Bikeshare in China as a public service : Comparing government-run and public-private partnership operation models. *Natural Resources Forum*, 39(1):41–52, fév. 2015.

[154] M. LOZANO, A. DUARTE, F. GORTÁZAR et R. MARTÍ : A hybrid metaheuristic for the cyclic antibandwidth problem. *Knowledge-Based Systems*, 54:103–113, 2013.

[155] G. LYON : Smart City Lyon, juin 2015. http ://www.economie.grandlyon.com/smart-city-lyon-france.346.0.html.

[156] G. LYON et CYCLOCITY : Site web Vélo'v, 2015. http ://www.velov.grandlyon.com/.

[157] G. LYON et CYCLOCITY : Vélo'v bat tous ses records !, jan. 2015. http ://www.grandlyon.com/fileadmin/user_upload/media/pdf/espace-presse/cp/20150112_cp_velov.pdf.

[158] D. J. MACKAY : *Information theory, inference, and learning algorithms*, vol. 7. Citeseer, 2003.

[159] R. MAGGIORI : Norbert Wiener, cyber-héros. avr. 2014. http ://www.liberation.fr/livres/2014/04/16/norbert-wiener-cyber-heros_998859.

[160] A. G. MAHYARI et S. AVIYENTE : Fourier Transform For Signals On Dynamic Graphs. 2014.

[161] M. MAIZIA et E. DUBEDAT : Analyse quantitative d'un service de vélos en libre-service : un système de transport à part entière. *Flux*, 71(1):73–77, 2008.

[162] J. MAX : *Méthodes et techniques de traitement du signal et application aux mesures physiques. Tome 2 - appareillages, exemples d'applications, méthodes nouvelles*. Masson, 1981.

[163] R. MEDDIN et P. DEMAIO : Bike Sharing Map, mai 2015. http ://www.bikesharingmap.com/.

[164] M. MEILA et J. SHI : A Random Walks View of Spectral Segmentation. *In 8th International Workshop on Artificial Intelligence and Statistics*, 2001.

[165] L. MERCHEZ et J.-B. ROUQUIER : Les rythmes urbains au prisme du vélo'v. *In Données urbaines 6*. 2011.

[166] G. MICHAU, C. ROBARDET, L. MERCHEZ, P. JENSEN, P. ABRY, P. FLANDRIN et P. BORGNAT : Peut-on attraper les utilisateurs de vélo'v au lasso. *In Proceedings of the 23e Colloque sur le Traitement du Signal et des Images. GRETSI-201*, p. 46–50, 2011.

[167] P. MIDGLEY : The role of smart bike-sharing systems in urban mobility. *JOURNEYS*, 2:23–31, 2009.

[168] P. MIDGLEY : Bicycle-sharing schemes : enhancing sustainable mobility in urban areas. *United Nations, Department of Economic and Social Affairs*, p. 1–12, 2011.

[169] C. MORENCY, M. TRÉPANIER et F. GODEFROY : Insights into Montreal's bikesharing system. *In Transportation Research Board 90th Annual Meeting*, vol. 6, 2011.

[170] M. MORIN : Bouygues Telecom Innovations - La ville de demain se construira avec ses habitants, 2014. http ://innovations.bouyguestelecom.fr/articles/la-ville-de-demain-se-construira-avec-ses-habitants.

[171] P. J. MUCHA, T. RICHARDSON, K. MACON, M. A. PORTER et J.-P. ONNELA : Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science*, 328(5980):876–878, mai 2010.

[172] J. MUNIER : L'identification de fronts d'ondes corrélés et distordus. *Traitement du Signal*, 4(4): 281–296, 1987.

[173] H. MURPHY : Dublin bikes : An investigation in the context of multimodal transport. *Dublin : MSc Sustainable Development, Dublin Institute of Technology*, 2010.

[174] R. NAIR et E. MILLER-HOOKS : Fleet Management for Vehicle Sharing Operations. *Transportation Science*, 45(4):524–540, déc. 2010.

[175] R. NAIR, E. MILLER-HOOKS, R. C. HAMPSHIRE et A. BUŠIĆ : Large-Scale Vehicle Sharing Systems : Analysis of Vélib'. *International Journal of Sustainable Transportation*, 7(1):85–106, jan. 2013.

[176] R. NASSIF, C. RICHARD, A. FERRARI et A. H. SAYED : Performance analysis of multitask diffusion adaptation over asynchronous networks. *In Proc. Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA*, 2014.

[177] F. D. NETWORK, L. Q. du NET et Fédération des fournisseurs d'accès à INTERNET ASSOCIATIFS : Amicus curiae transmis au Conseil constitutionnel dans le cadre des saisines visant la « loi relative au renseignement », 2015. http ://www.fdn.fr/pjlr/amicus1.pdf.

[178] M. NEWMAN : *Networks : An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[179] V. NICOSIA, J. TANG, C. MASCOLO, M. MUSOLESI, G. RUSSO et V. LATORA : Graph Metrics for Temporal Networks. *In* P. HOLME et J. SARAMÄKI, éds : *Temporal Networks*, Understanding Complex Systems, p. 15–40. Springer Berlin Heidelberg, jan. 2013.

[180] R. B. NOLAND et M. M. ISHAQUE : Smart bicycles in an urban area : Evaluation of a pilot scheme in London. *Journal of Public Transportation*, 9(5):71, 2006.

[181] A. M. NUÑEZ, L. LACASA, J. P. GOMEZ et B. LUQUE : *Visibility algorithms : A short review*. INTECH Open Access Publisher, 2012.

[182] O. O'BRIEN, J. CHESHIRE et M. BATTY : Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, 34:262–273, jan. 2014.

[183] F. OGILVIE et A. GOODMAN : Inequalities in usage of a public bicycle sharing scheme : Socio-demographic predictors of uptake and usage of the London (UK) cycle hire scheme. *Preventive Medicine*, 55(1):40–45, juil. 2012.

[184] N. OTSU : A threshold selection method from gray-level histograms. *Automatica*, 11(285-296): 23–27, 1975.

[185] C. H. PAPADIMITRIOU : The NP-completeness of the bandwidth minimization problem. *Computing*, 16(3):263–270, 1976.

[186] S. D. PARKES, G. MARSDEN, S. A. SHAHEEN et A. P. COHEN : Understanding the diffusion of public bikesharing systems : evidence from Europe and North America. *Journal of Transport Geography*, 31:94–103, juil. 2013.

[187] B. PARNELL-HOPKINSON : Santander Replaces Barclays As Cycle Hire Sponsor. *Londonist*, fév. 2015.

[188] N. PAS : Images d'une révolte ludique. Le mouvement néerlandais Provo en France dans les années soixante. *Revue historique*, 634(2):343–373, 2005.

[189] S. L. PECK : Perspectives on why digital ecologies matter : Combining population genetics and ecologically informed agent-based models with GIS for managing dipteran livestock pests. *Acta Tropica*, 138:S22–S25, oct. 2014.

[190] D. PITT et L. MERCHEZ : Visualisation de flux de Vélov, 2014. http ://anr-velov.ens-lyon.fr/.

[191] T. M. PRZYTYCKA, M. SINGH et D. K. SLONIM : Toward the dynamic interactome : it's about time. *Briefings in Bioinformatics*, 11(1):15–29, jan. 2010.

[192] J. PUCHER et R. BUEHLER : *City cycling*. MIT Press, 2012.

[193] N. PUSTELNIK : *Méthodes proximales pour la résolution de problèmes inverses. Application à la Tomographie par Emission de Positrons.* Thèse de doctorat, Université Paris-Est, 2010.

[194] B. QUETELARD : Usagers et déplacements à vélo en milieu urbain. *CERTU*, 6:9, 2012.

[195] A. N. RANDRIAMANAMIHAGA, E. CÔME, L. OUKHELLOU et G. GOVAERT : Clustering the Vélib' dynamic Origin/Destination flows using a family of Poisson mixture models. *Neurocomputing*, 141:124–138, oct. 2014.

[196] E. RAVALET et Y. BUSSIÈRE : Les systèmes de vélos en libre-service expliquent-ils le retour du vélo en ville ? *Recherche Transports Sécurité*, 28(1):15–24, fév. 2012.

[197] O. RAZEMON : Deux villes, Pau et Valence, s'apprêtent à renoncer aux vélos en libre-service. *M Le magazine du Monde*, nov. 2014.

[198] O. RAZEMON : Le vélo en libre-service cherche toujours son modèle économique. *LE MONDE ECONOMIE*, mai 2015.

[199] P. RÉFRÉGIER : *Théorie du signal : signal information fluctuations*. Masson, 1993.

[200] M. RICCI : Bike sharing : Affordable convenience or unaffordable luxury ? 2013.

[201] J. RICHIARDI, S. ACHARD, H. BUNKE et D. VAN DE VILLE : Machine Learning with Brain Graphs : Predictive Modeling Approaches for Functional Imaging in Systems Neuroscience. *IEEE Signal Processing Magazine*, 30(3):58–70, mai 2013.

[202] J. RIFKIN : *The third industrial revolution : how lateral power is transforming energy, the economy, and the world*. Macmillan, 2011.

[203] E. RODRIGUEZ-TELLO, J.-K. HAO et J. TORRES-JIMENEZ : Memetic algorithms for the MinLA problem. *In Artificial Evolution*, p. 73–84. Springer, 2006.

[204] E. RODRIGUEZ-TELLO, J.-K. HAO et J. TORRES-JIMENEZ : An effective two-stage simulated annealing algorithm for the minimum linear arrangement problem. *Computers & Operations Research*, 35(10):3331–3346, oct. 2008.

[205] D. ROJAS-RUEDA, A. de NAZELLE, M. TAINIO et M. J. NIEUWENHUIJSEN : The health risks and benefits of cycling in urban environments compared with car use : health impact assessment study. *BMJ*, 343(aug04 2):d4521–d4521, août 2011.

[206] H. ROMERO-MONSIVAIS, E. RODRIGUEZ-TELLO et G. RAMIREZ : A New Branch and Bound Algorithm for the Cyclic Bandwidth Problem. *Advances in Computational Intelligence*, 2013.

[207] F. ROSSI : Visualization methods for metric studies. *In Proceedings of the International Workshop on Webometrics, Informetrics and Scientometrics*, p. 356–366, 2006.

[208] P. J. ROUSSEEUW : Silhouettes : a graphical aid to the interpretation and validation. *Journal of Computational and Applied Mathematics*, 20:50–65, 1987.

[209] G. K. D. SAHARIDIS, A. FRAGKOGIOS et E. ZYGOURI : A Multi-Periodic Optimization Modeling Approach for the Establishment of a Bike Sharing Network : a Case Study of the City of Athens. *In Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 2, 2014.

[210] A. SANDRYHAILA et J. MOURA : Discrete Signal Processing on Graphs : Frequency Analysis. *Signal Processing, IEEE Transactions on*, 62(12):3042–3054, juin 2014.

[211] A. SARKAR, N. LATHIA et C. MASCOLO : Comparing cities' cycling patterns using online shared bicycle maps. *Transportation*, avr. 2015.

[212] D. SATSANGI : *Design and development of metaheuristic techniques for some graph layout problems*. Thèse de doctorat, Dayalbag Educational Institute, 2013.

[213] D. SATSANGI, K. SRIVASTAVA et GURSARAN : General variable neighbourhood search for cyclic bandwidth sum minimization problem. *In Students Conference on Engineering and Systems (SCES)*, p. 1–6, 2012.

[214] H. SAYARSHAD, S. TAVASSOLI et F. ZHAO : A multi-periodic optimization formulation for bike planning and bike utilization. *Applied Mathematical Modelling*, 36(10):4944–4951, 2012.

[215] M. N. SCHMIDT et M. MORUP : Nonparametric Bayesian modeling of complex networks : an introduction. *IEEE Signal Processing Magazine*, 30(3):110–128, mai 2013.

[216] J. SCHUIJBROEK, R. HAMPSHIRE et W.-J. van HOEVE : Inventory rebalancing and vehicle routing in bike sharing systems. 2013.

[217] S. SHAHEEN, S. GUZMAN et H. ZHANG : Bikesharing in Europe, the Americas, and Asia. *Transportation Research Record : Journal of the Transportation Research Board*, 2143:159–167, déc. 2010.

[218] S. SHAHEEN, H. ZHANG, E. MARTIN et S. GUZMAN : China's Hangzhou Public Bicycle. *Transportation Research Record : Journal of the Transportation Research Board*, 2247:33–41, déc. 2011.

[219] Y. SHIMADA, T. IKEGUCHI et T. SHIGEHARA : From Networks to Time Series. *Phys. Rev. Lett.*, 109(15):158701, 2012.

[220] Y. SHIMADA, T. KIMURA et T. IKEGUCHI : Analysis of chaotic dynamics using measures of the complex network theory. *In Artificial Neural Networks-ICANN 2008*, p. 61–70. Springer, 2008.

[221] D. SHUMAN, B. RICAUD, P. VANDERGHEYNST *et al.* : A windowed graph Fourier transform. *In Statistical Signal Processing Workshop (SSP), 2012 IEEE*, p. 133–136. Ieee, 2012.

[222] D. I. SHUMAN, S. K. NARANG, P. FROSSARD, A. ORTEGA et P. VANDERGHEYNST : The emerging field of signal processing on graphs : Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

[223] M. SPILIOPOULOU, I. NTOUTSI, Y. THEODORIDIS et R. SCHULT : MONIC : Modeling and Monitoring Cluster Transitions. *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, p. 706–711, New York, NY, USA, 2006. ACM.

[224] J. STEHLÉ, N. VOIRIN, A. BARRAT, C. CATTUTO, L. ISELLA, J.-F. PINTON, M. QUAGGIOTTO, W. Van den BROECK, C. RÉGIS, B. LINA et P. VANHEMS : High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. *PLoS ONE*, 6(8):e23176, août 2011.

[225] SYTRAL : Transport en Commun Lyonnais (TCL). http ://www.tcl.fr/.

[226] T. TAKAGUCHI, N. MASUDA et P. HOLME : Bursty Communication Patterns Facilitate Spreading in a Threshold-Based Epidemic Dynamics. *PLoS ONE*, 8(7):e68629, juil. 2013.

[227] J. B. TENENBAUM : A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[228] R. TIBSHIRANI : Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, jan. 1996.

[229] A. N. TIKHONOV : On the stability of inverse problems. *Doklady Akademii nauk SSSR*, 39(5):195–198, 1943.

[230] W. S. TORGERSON : Multidimensional scaling : I. Theory and method. *Psychometrika*, 17(4):401–419, 1952.

[231] T. D. TRAN, N. OVTRACHT et B. F. D'ARCIER : Modeling Bike Sharing System using Built Environment Factors. *Procedia CIRP*, 30:293–298, 2015.

[232] TRANSPORT FOR LONDON : Travel in London reports. Rap. tech., 2013.

[233] N. TREMBLAY et P. BORGNAT : Graph Wavelets for Multiscale Community Mining. *IEEE Transactions on Signal Processing*, 62(20):5227–5239, 2014.

[234] N. TREMBLAY, P. BORGNAT et P. FLANDRIN : Graph empirical mode decomposition. *In Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, p. 2350–2354. IEEE, 2014.

[235] M. TRIBASTONE, A. CLARK, N. GAST, S. GILMORE et D. REIJSBERGEN : Data validation and requirements for case studies. *QUANTICOL Deliverable*, 5, 2014.

[236] D. O. TUAMA : Ripples through the city : Understanding the processes set in motion through embedding a public bike sharing scheme in a city. *Research in Transportation Business Management*, (0):–, 2015.

[237] P. UEASANGKOMSATE : Efficiency Management of Public Bike-Sharing Sytem in Bangkok. International Centre of Economics, Humanities and Management, oct. 2014.

[238] P. VAN MIEGHEM : *Graph spectra for complex networks*. Cambridge University Press, 2011.

[239] T. VIRTANEN, A. T. CEMGIL et S. GODSILL : Bayesian extensions to non-negative matrix factorisation for audio signal modelling. *In IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, p. 1825–1828. IEEE, 2008.

[240] M. VOGEL, R. HAMON, G. LOZENGUEZ, L. MERCHEZ, P. ABRY, J. BARNIER, P. BORGNAT, P. FLANDRIN, I. MALLON et C. ROBARDET : From bicycle sharing system movements to users : a typology of Vélo'v cyclists in Lyon based on large-scale behavioural dataset. *Journal of Transport Geography*, 41:280–291, déc. 2014.

[241] P. VOGEL, T. GREISER et D. C. MATTFELD : Understanding Bike-Sharing Systems using Data Mining : Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences*, 20:514–523, jan. 2011.

[242] P. VOGEL et D. C. MATTFELD : Modeling of repositioning activities in bike-sharing systems. *In Proceeding of the 12th world conference on transport research*, p. 11–15, 2010.

[243] P. VOGEL, B. NEUMANN SAAVEDRA et D. MATTFELD : A Hybrid Metaheuristic to Solve the Resource Allocation Problem in Bike Sharing Systems. *In* M. BLESA, C. BLUM et S. VOSS, éds : *Hybrid Metaheuristics*, vol. 8457 de *Lecture Notes in Computer Science*, p. 16–29. Springer International Publishing, jan. 2014.

[244] Q. WANG : *Overlapping community detection in dynamic networks*. Thèse de doctorat, Ecole normale supérieure de lyon-ENS LYON, 2012.

[245] Y. WANG, B. WU et N. DU : Community evolution of social network : feature, algorithm and model. *arXiv preprint arXiv :0804.4356*, 2008.

[246] A. WASERHOLE, V. JOST et N. BRAUNER : Pricing techniques for self regulation in Vehicle Sharing Systems. *Electronic Notes in Discrete Mathematics*, 41:149–156, juin 2013.

[247] D. J. WATTS et S. H. STROGATZ : Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[248] T. WENG, Y. ZHAO, M. SMALL et D. D. HUANG : Time-series analysis of networks : Exploring the structure with random walks. *Physical Review E*, 90(2):022804, 2014.

[249] N. WIENER : *Extrapolation, interpolation, and smoothing of stationary time series*, vol. 2. MIT press Cambridge, MA, 1949.

[250] WIKIPEDIA : List of bicycle sharing systems — Wikipedia, The Free Encyclopedia. 2015. [Online ; accessed 30-June-2015].

[251] J. WOODCOCK, M. TAINIO, J. CHESHIRE, O. O'BRIEN et A. GOODMAN : Health effects of the London bicycle sharing system : health impact modelling study. *BMJ*, 348(feb13 1):g425–g425, fév. 2014.

[252] H. XU, J. YING, H. WU et F. LIN : Public Bicycle Traffic Flow Prediction based on a Hybrid Model. *Appl. Math*, 7(2):667–674, 2013.

[253] K. S. XU et A. O. HERO III : Dynamic Stochastic Blockmodels : Statistical Models for Time-Evolving Networks. *In* A. M. GREENBERG, W. G. KENNEDY et N. D. BOS, éds : *Social Computing, Behavioral-Cultural Modeling and Prediction*, vol. 7812 de *Lecture Notes in Computer Science*, p. 201–210. Springer Berlin Heidelberg, 2013.

[254] K. S. XU, M. KLIGER et A. O. HERO : A regularized graph layout framework for dynamic network visualization. *Data Mining and Knowledge Discovery*, 27(1):84–116, juil. 2013.

[255] W. XU, R. C. WILSON et E. R. HANCOCK : Determining the cause of negative dissimilarity eigenvalues. *In Computer Analysis of Images and Patterns*, p. 589–597. Springer, 2011.

[256] J. YANG et J. LESKOVEC : Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, jan. 2015.

[257] T. YANG, P. HAIXIAO et S. QING : Bike-sharing systems in Beijing, Shanghai and Hangzhou and their impact on travel behaviour. *In Transportation Research Board Annual Meeting*, 2011.

[258] J. W. YOON, F. PINELLI et F. CALABRESE : Cityride : A Predictive Bike Sharing Journey Advisor. p. 306–311. IEEE, juil. 2012.

[259] Q. YU et L. CHEN : A New Method for Detecting Anti-community Structures in Complex Networks. *Journal of Physics : Conference Series*, 410:012103, fév. 2013.

[260] A. YUHAS : NSA reform : USA Freedom Act passes first surveillance reform in decade – as it happened | US news | The Guardian. *The Guardian*, 2015. http ://www.theguardian.com/us-news/live/2015/jun/02/senate-nsa-surveillance-usa-freedom-act-congress-live.

[261] M. ZALTZ AUSTWICK, O. O'BRIEN, E. STRANO et M. VIANA : The Structure of Spatial Networks and Communities in Bicycle Sharing Systems. *PLoS ONE*, 8(9):e74685, sept. 2013.

[262] J. ZHANG et M. SMALL : Complex Network from Pseudoperiodic Time Series : Topology versus Dynamics. *Physical Review Letters*, 96(23):238701, 2006.

[263] Z. ZHANG, C. DING, T. LI et X. ZHANG : Binary matrix factorization with applications. *In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, p. 391–400. IEEE, 2007.

[264] Q. ZHAO, Y. TIAN, Q. HE, N. OLIVER, R. JIN et W.-C. LEE : Communication Motifs : A Tool to Characterize Social Communications. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, p. 1645–1648, New York, NY, USA, 2010. ACM.

[265] H. ZOU et T. HASTIE : Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2):301–320, avr. 2005.

Table des matières

Titre –

Analyse de réseaux temporels par des méthodes de traitement du signal : Application au système de vélos en libre-service à Lyon

Résumé –

Les systèmes de vélos en libre-service sont devenus des éléments indispensables dans les offres de transport urbain des grandes villes mondiales. À partir des données que ces systèmes génèrent, il est possible d'avoir une caractérisation fine de l'utilisation du vélo en milieu urbain, tant sur des problématiques traitant du domaine des transports que des aspects socio-économiques. Comme pour de nombreux domaines profitant de la récente abondance en données permises par les technologies actuelles de communication et de stockage de l'information, les enjeux actuels résident dans le développement de méthodes d'analyse de données efficaces et adaptées aux systèmes étudiés. Cette thèse se propose de répondre à cette problématique, à la fois par des développements méthodologiques et par une application à des données réelles issues du système de vélos en libre-service Vélo'v à Lyon.

Le système Vélo'v peut se représenter sous la forme d'un réseau, décrivant un ensemble de relations entre les différentes stations. Cette représentation, valable également pour de nombreux systèmes, permet l'utilisation d'outils pour décrire la structure du réseau basés sur la théorie des graphes. Néanmoins, la prise en compte d'une dynamique temporelle dans l'évolution des systèmes nécessite d'étendre l'analyse à des réseaux temporels, dont la structure évolue au cours du temps. Le parallèle avec le domaine du traitement du signal, dont le but est l'analyse de signaux temporels, amène à considérer des connexions entre la description de l'évolution d'un réseau temporel et celle d'un signal. Ces travaux proposent de considérer une dualité entre les réseaux temporels et les signaux, de sorte que l'analyse dans le domaine des signaux, à l'aide des outils du traitement du signal, permet de caractériser le réseau temporel correspondant.

Cette méthodologie, à la frontière entre le traitement du signal et l'analyse des réseaux, est tout d'abord justifiée par l'étude du système Vélo'v, en comparant différentes approches d'analyse de données et les apports de la représentation sous la forme de réseau temporel. Une méthode d'étiquetage des nœuds d'un graphe est ensuite discutée, permettant d'ouvrir la voie vers une dualité entre réseaux et signaux. Cette dualité est étendue aux réseaux temporels, pour lesquels une méthode d'extraction automatique des structures pertinentes au cours du temps est proposée, à travers la décomposition des signaux correspondants.

Title – Analysis of temporal networks using signal processing methods : Application to the bike-sharing system in Lyon

Abstract –

Bike-sharing systems have become essential elements in urban transportation systems of many world's big cities. Thanks to the data generated by these systems, it is possible to obtain a precise characterization of urban cycling, both in terms of transportation and socio-economic aspects. Taking advantage of the recent abundance of data allowed by the current technology, the challenges lie in the development of efficient data analysis method, adapted to these systems. This PhD thesis proposes some answers to this issue, first by methodological developments and second by studying real-world data obtained from the bike-sharing system in Lyon, called Vélo'v.

The Vélo'v system can be represented as a network, describing a set of relations between the stations spread over the city. This representation, used for many systems, enables the use of tools from network theory to measure the network structure and understand the underlying mechanisms. Nevertheless, taking into account the dynamic evolution of the structure requires an extension of the classical tools to the temporal case. Parallels between this problem and the field of signal processing can be done, and opens the way to the consideration of connections between the description of the dynamics of temporal networks and those of signals. This work introduces a duality between temporal networks and signals, such that the analysis of the signals using the classical tools of signal processing helps to the characterization of the structure of the corresponding network.

This methodology, at the juncture between signal processing and network analysis, is first justified by the study of the Vélo'v network, by comparing different data analysis method and the representation of the system as a temporal network. Then, a method to relabel the vertices of the graph according to the topology of the network is discussed, opening up a duality between networks and signals. This duality is then extended to temporal networks : The analysis of the spectral properties of the signals are studied through a fully automated extraction method, enabling the decomposition of relevant network structure over time.