Habilitation à Diriger des Recherches

présentée à

L'Ecole Normale Supérieure de Lyon

par

Pierre Borgnat

Chargé de recherche CNRS

Signaux, réseaux et graphes : analyses non stationnaires ou multi-échelle

---

Section CNU : Génie informatique, automatique et traitement du signal
N° de la section CNU : 61
Discipline : **Traitement du signal**

Soutenue le 10 avril 2014 devant la commission d'examen formée de :

Rapporteurs :	
Matthieu Latapy	Directeur de recherche CNRS, LIP6, Paris
Éric Moulines	Professeur Télécom, Télécom-Paristech, Paris
Pierre Vandergheynst	Professeur associé, EPFL, Lausanne
Examinateurs :	
Marc Barthelemy	Chercheur CEA, IPhT, CEA Saclay
Patrick Flandrin	Directeur de recherche CNRS, ENS de Lyon
Président :	
Jean-Yves Tourneret	Professeur d'université, ENSEEIHT, IRIT, Toulouse

- 2014 -

Chapitre 1

Introduction

1.1 Contexte scientifique et parcours

Les travaux présentés dans ce mémoire relèvent principalement du traitement du signal, mais ils se déclinent en réalité à la croisée de plusieurs disciplines : le traitement du signal bien entendu avec un penchant vers les méthodes non stationnaires ou multi-échelle, l'analyse des systèmes complexes et des réseaux, aux rangs desquels Internet est un réseau particulier qui conduit à des signaux de télétrafic particulièrement intéressant, et l'analyse des graphes et des signaux sur graphes.

J'ai été recruté au CNRS, en section 07, depuis octobre 2004 et affecté au Laboratoire de Physique de l'ENS de Lyon (UMR 5672) pour un projet de recherche d'étude des signaux non stationnaires. Ces activités ont été menées au sein de l'équipe Sisyphe : « Signaux, systèmes et physique ». Le projet s'inscrivait alors dans un des deux thèmes forts de l'équipe et était de s'intéresser aux notions de stationnarité des signaux en un sens général, par exemple à rechercher des méthodologies concrètes pour tester le caractère stationnaire ou non d'un signal pour citer une de nos idées qui a été fructueuse. J'ai plus généralement une activité de recherche continue sur les méthodes non stationnaires en traitement du signal

En parallèle, j'ai développé de travaux portant sur l'analyse du télétrafic des réseaux d'ordinateurs à l'aide de méthodologies venues du traitement du signal – en se pliant aux enjeux propres en télétrafic informatique – en abordant les questions de modélisation, d'estimation, de détection et de classification pour les signaux associés. Ces sujets m'ont beaucoup occupé de 2005 à 2010 environ. Depuis 2008, mon intérêt a peu à peu migré vers l'étude des réseaux complexes, d'abord en marge de mes activités sur les réseaux d'ordinateurs, puis en y consacrant plus d'efforts ces dernières années pour développer une recherche en traitement du signal pour les graphes complexes qui reste originale par rapport aux approches usuellement venues des sciences physiques ou informatiques pour étudier les réseaux complexes. Ce thème, en « traitement du signal & graphes, » est devenu un axe à part entière de ma recherche et de celle de l'équipe Sisyphe.

Tous ces travaux ne sont bien entendu pas réalisés seuls et je suis reconnaissant d'avoir trouvé d'abord au sein de l'équipe Sisyphe et plus généralement au Laboratoire de Physique un cadre agréable et stimulant pour développer des approches originales sur tous ces sujets – même quand ils s'éloignent des sciences physiques ! – et des collègues, plus jeunes ou moins jeunes, toujours partants pour travailler sur ces sujets qu'on pourrait trouver divers. Bien sûr, je suis autant redevable aux collègues d'ici et d'ailleurs avec qui j'ai eu l'occasion de

travailler – tous les travaux sur le télétrafic informatique doivent beaucoup à des séjours répétés au NII à Tokyo (Japon) et aux collègues français du projet METROSEC de fin 2004 à fin 2007 grâce auxquels je suis entré dans le sujet.

1.2 Organisation du document

Plan. Le présent mémoire est organisé comme il suit. Dans le premier chapitre est donné un survol général des mes activités de recherche sur ces dix dernières années, tracé à grands traits en 1.3. On y trouvera les références aux travaux que j'ai accomplis pour servir de guide de lecture ; les références bibliographiques extérieures seront absentes dans ces quelques pages mais le lecteur les retrouvera en bonne place dans les chapitres 2 à 4 du mémoire qui décrivent plus en détail les questions abordées et les travaux effectués. Le chapitre 2 concerne des travaux en traitement du signal non stationnaire. Les activités de recherche sur l'étude du télétrafic sur les réseaux d'ordinateurs sont décrites dans le chapitre 3. Le chapitre 4 développe les contributions (certaines très récentes) sur sur le traitement du signal pour les graphes complexes. Le mémoire se termine par une conclusion en chapitre 5.

Travaux et références bibliographiques. Les références données avec les codes suivants sont celles de mes travaux : [Jxx] sont les articles parus dans de journaux, [Jsxx] des articles actuellement soumis, [Pxx] des articles dans des actes de conférences, [Psxx] des soumission et [Cxx] les chapitres d'ouvrages collectifs. Ces travaux ont été découpés en trois groupes, chacun étant lié à un des chapitres du mémoire et on les trouvera en fin de chaque chapitre. Quelques travaux supplémentaires plus anciens que ceux décrits dans ce mémoire sont juste listés avant la bibliographie : ce sont les travaux liés à ma thèse ou mon post-doc (portant sur l'étude des invariances d'échelle brisées en traitement du signal et en physique de la turbulence) ou plus tôt dans mes études (en RMN) [J1, J2]. Les travaux et des codes numériques sont disponibles sur ma page web : `http//perso.ens-lyon.fr/pierre.borgnat`. La bibliographie générale des publications que je cite dans le mémoire est donnée à la fin à partir de la page 127.

Document complémentaire pour le jury. Un deuxième document regroupe un *curriculum vitæ* détaillé et la liste complète de mes travaux, cette fois organisée par ordre chronologique. Le détail des encadrements, ma participation à l'enseignement, les conférences invitées, etc. y sont donnés. On y trouvera aussi la reproduction de quatre articles sélectionnés pour être représentatifs des travaux discutés dans les chapitres 2, 3 et 4 :

- Ch. 2 : [**J16**] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, J. Xiao, "Testing stationarity with surrogates : A time-frequency approach", *IEEE Trans. on Signal Processing*, Vol. 58 :7, p 3459-3470, July 2010.
- Ch. 3 : [**P29**] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, K. Cho, "Seven Years and One Day : Sketching the Evolution of Internet Traffic," *Proceedings of the 28th IEEE INFOCOM 2009*, pp. 711–719 Rio de Janeiro (Brazil), May 2009.
- Ch. 4.2 : [**J24**] N. Tremblay, A. Barrat, C. Forest, M. Nornberg, J.F. Pinton, P. Borgnat, "Bootstrapping under constraint for the assessment of group behavior in human contact networks", *Physical Review E*, vol. 88 :5, p. 052812, Nov. 2013.
- Ch. 4.4 : [**Js27**] N. Tremblay, P. Borgnat, "Graph Wavelets and Community Mining", submitted, July 2013 (revision, January 2014).

1.3 Grandes lignes du travail

Un fil méthodologique de fond à été le **traitement du signal non stationnaire** (au sens large). Avant d'être au CNRS, j'avais travaillé en thèse et post-doc sur l'équivalence entre stationnarité et invariance d'échelle qu'apporte l'utilisation de la transformation de Lamperti, une anamorphose logarithmique reliant le temps à l'échelle. Nous avions systématisé cette équivalence en l'employant pour construire des méthodes d'analyse, de modélisation et de représentation de signaux invariants d'échelle (ou à invariances brisées) [J3, J4, J5, C1, C2, P1, P2, P3, P4, P5, P6, P7, P8, P9, P12]. Depuis mon recrutement, le résultat principal en traitement de signaux non stationnaires a été, en collaboration avec P. Flandrin, de rendre le concept de stationnarité opérationnel (avec des tests de stationnarité) et relatif aux conditions d'observation (durée du signal, résolution de l'observation, signal aléatoire ou déterministe) [J9, J10, J11, **J16**, J21, P18, P19, P20, P27, P31, P34, P42, P45, P50, C5]. Nous avons proposé aussi des outils utiles à des opérations telles que l'extraction de tendance [J20, J22]), l'estimation spectrale [P41] ou l'extrapolation de données [P49] à base de Décomposition Modale Empirique (EMD), ou une procédure d'estimation du nombre de composantes pour une décomposition de signal en modes [P61]. Cela recouvre aussi la synthèse de signaux à propriétés prescrites [P51] grâce aux données substituts (*surrogates*). Nous avons aussi étudié les méthodes multi-fenêtres pour l'estimation cepstrale [J14] et développé des approches conduisant à des représentations temps-fréquence parcimonieuses [P23, P28, J15] en important les idées d'échantillonnage compressé dans les méthodes temps-fréquence. Un projet ANR blanc de l'appel 2007, « Stationnarité Relative et Approches Connexes » qui regroupait 5 chercheurs permanents, C. Richard ; P. Honeine, P.O. Amblard, P. Flandrin et moi. nous a permis les avancées significatives indiquées plus haut sur ces questions.

Plus récemment, depuis le recrutement au laboratoire de N. Pustelnik (CR CNRS), nous étendons la Décomposition Modale Empirique par des approches d'optimisation sous contrainte, réminiscentes de la séparation de textures et géométrie pour les images, adaptées pour l'instant à l'analyse des signaux [P52] (et un article de journal en révision).

D'autres études en non stationnaire ont été faites à l'occasion d'applications : signaux cardiaques du fœtus lors de l'accouchement [P56], signaux de capteurs environnementaux (capteurs *LiveE !* avec le NII et l'Université de Tokyo) [P30, P33], de capteurs de consommation d'énergie dans les bâtiments [P54, P55]. Grâce à ma formation initiale en physique et par ma situation au sein d'un laboratoire de physique, j'ai contribué au traitement du signal pour la physique, en particulier en mécanique des fluides. Ce fut le cas avec des travaux liés à l'extraction de tendances dans des signaux. Ils nous ont en particulier permis de proposer avec des collègues en mécanique des fluides une nouvelle approche pour la simulation numérique aux grandes échelles des fluides (LES pour *Large Eddy Simulation*) [J18, P32, P47, P48]. L'idée directrice fut de proposer une procédure d'extraction du flot moyen par filtrage de Kalman qui permet d'améliorer le calcul du modèle de turbulence sous-maille.

Ces travaux, couvrent (sans compter les travaux de thèse et post-doc) 25 publications dont 10 dans des journaux et 1 chapitre de livre.

L'activité de recherche débutée fin 2004 sur l'étude des **réseaux d'ordinateurs et le télétrafic informatique** (sur Internet principalement) a débuté par une collaboration avec P. Abry au Laboratoire de Physique et très vite avec des collègues de laboratoires français (par exemple P. Owezarski au LAAS-CNRS à Toulouse, l'équipe RESO du LIP,

à Lyon) ou internationaux (en particulier NII et IIJ à Tokyo, ou FTW à Vienne). Cette activité à l'interface entre la recherche en réseaux et en traitement du signal est partie de l'utilisation de méthodes multi-échelles pour la métrologie du trafic Internet. L'originalité de ce travail par rapport à ce qui se faisait avant (au laboratoire par P. Abry, ou ailleurs) était de combiner des modélisations statistiques avancées du trafic internet (par exemple des statistiques multi-échelles non gaussiennes) aux enjeux pratiques venus de l'ingénierie du réseau et aux mesures possibles (avec quelques expériences de métrologie du télétrafic informatique) et à la modélisation [J6, J7, P10, P11, P13, P14, P16, P22]. Après avoir développé des méthodes de détection et d'identification d'anomalies [P15, P17, P21], nous avons proposé des outils pour l'annotation automatique des anomalies dans des bases de données de trafic et la validation croisée des détecteurs d'anomalie [P38, P43, P44], des méthodes de classification de trafic et des ordinateurs connectés à internet [J17, J23] tout en renouvelant les études des caractères non gaussien et auto-similaire du trafic [J12, J13, **P29**, P40] et, bientôt, de leurs propriétés multifractales (en préparation).

Ces travaux ont été l'occasion de projets collaboratifs : une ACI (projet METROSEC avec P. Owezarski) initialement de fin 2004 à 2007, un projet ANR (OSCAR, projet pré-compétitif du RNRT) en 2006-2008 et de collaborations internationales et plus particulièrement au Japon (par des accords CNRS-JST puis JSPS) avec le consortium WIDE, l'université de Tokyo, le laboratoire de IIJ (Internet Initiative Japan) et le laboratoire NII (National Institute of Informatics) à Tokyo où j'ai fait de fréquents séjours de 1 semaine à 1 mois depuis 2006. Grâce à un financement CNRS-JSPS, cette collaboration a été activement continuée de 2010 à 2012. Cela a conduit à 10 publications en commun avec ces laboratoires depuis 2007. Globalement, j'ai contribué à 20 publications (dont 6 de journaux et 3 d'actes longs dans des conférences de réseau) dans ce domaine de l'analyse du télétrafic Internet.

Mon parcours scientifique m'a peu à peu amené, un peu depuis 2008, beaucoup depuis 2012, à m'intéresser à l'**analyse des systèmes et réseaux complexes** par la théorie du signal : au début par une étude de la dynamique de réseaux complexes tels que les réseaux de contacts entre humains (avec des enjeux de départ liés aux réseaux mobiles) [J8, C4, P26, P39], puis par l'étude des données Vélo'v [J19, P35, P36, P37, P46, C6]. Ces travaux ont mené au développement d'une nouvelle activité de recherche, que je résume sous l'intitulé « traitement du signal & graphes » et qui ne porte pas que sur l'analyse de réseaux complexes mais plus spécifiquement sur le **traitement du signal sur ou pour des graphes**. Cela s'appuie naturellement sur les deux axes méthodologiques historiquement forts de l'équipe, les méthodes non stationnaires et les approches multi-échelles, et les amène vers de nouveaux horizons. Le travail vise à apporter des développements méthodologiques originaux en traitement du signal sur ou pour des graphes. Les analyses sur le Vélo'v ont par ailleurs attiré un certain intérêt, et j'ai donné plusieurs exposés invités sur ce sujet. Ce travail continue actuellement structuré dans un projet ANR Vel'innov (Programme "Sociétés Innovantes" 2012, de 02/2013 à 02/2016), sur l'étude des systèmes Vélo'v, en collaboration entre des laboratoires de l'ENSL et les laboratoires lyonnais LET-CNRS (Laboratoire d'économie des transports) et LIRIS, et le Polytech Montreal.

Partant des études sur les systèmes complexes, le thème « traitement du signal & graphes » est devenu un axe à part entière de ma recherche et de celle de l'équipe Sisyphe. Par exemple, je participe à la direction de 4 thèses en cours dans ces thématiques. La thèse de N. Tremblay (que je dirige), commencée en septembre 2011, est partie de l'étude de réseaux de contacts entre humains et nous avons d'abord proposé une méthode de bootstraps pour

ces réseaux complexes (que nous mesurons grâce à une collaboration avec le projet Sociopatterns.org, en utilisant une plateforme à base de capteurs RFID) [**J24**, P53]. Puis, a été développée une approche multi-échelle de détection de communautés dans des réseaux par des ondelettes sur graphes [P59, P60, P63, P65] (et un article soumis [**Js27**]). Je co-encadre un doctorant (pour 1/3), R. Hamon (co-tutelle avec P. Flandrin, C. Robardet, LIRIS, INSA de Lyon) qui a commencé en septembre 2012 et étudie des modes de représentation de graphes non stationnaires et a pour objet final l'analyse des données Vélo'v [P57, P62, P64] (et 2 articles soumis). Une troisième doctorante, A. Costard (co-tutelle avec P. Abry et S. Achard, O. Michel, du GIPSA-lab, Grenoble) que j'encadre à 25%, a débuté en octobre 2011 et nous étudions comment mélanger les méthodes d'estimation bayésienne et les approches de *graphical lasso* pour l'estimation de la topologie d'un graphe de dépendance quand on a peu de points de mesure [P58]. Le contexte d'application y est ici l'analyse d'imagerie cérébrale par IRMf. Le 4e doctorant, G. Michau, (encadrement à 25%) est en co-tutelle avec le Smart Transport Research Center (QUT de Brisbane) et travaille sur l'étude des flux routiers à l'aide d'un ensemble de capteurs Bluetooth déployés en ville à Brisbane. Nous étudions les méthodes d'estimation des flux origine-destination par ces capteurs qui enregistrent le passage de tout système Bluetooth (dans les voitures) à leur proximité. Cette thèse a débuté au début de l'été 2013 et a donné lieu très récemment à une communication [P66].

Tous ces travaux s'inscrivent dans un contexte d'étude de signaux sur des réseaux (ou graphes) ou de réseaux eux-mêmes. Nos résultats récents cherchent à promouvoir le traitement de signaux sur graphes en y associant les thématiques des réseaux complexes, par exemple la détection de communautés [P59, P60, P63, P65, **Js27**] ou l'étude des sous-groupes dans un réseau [**J24**, P53]. Nous cherchons par la même à étudier des graphes qui sont les signaux (ou données) d'intérêt et pas seulement des signaux sur une topologie de graphe donnée par ailleurs. Réseaux de contacts entre humains, réseaux de mobilité, réseau de déplacement ou de transport, réseaux d'activation dans le cerveau,... sont autant d'instances de réseaux qui peuvent eux-mêmes être les signaux à modéliser, estimer ou analyser. À côté de cela, le réseau Internet ou les réseaux de capteurs sont des situations où l'on s'intéresse à des signaux sur graphes, que nous étudions par des approches multi-échelles et non stationnaires. Cet axe de recherche, des systèmes complexes au thème Signal & Graphes, a conduit à 23 publications depuis 2008, dont 3 dans des journaux et 2 chapitres de livre. Ce thème est devenu mon sujet de prédilection actuel et 10 de ces travaux ont été publiés en 2013.

Chapitre 2

Contributions au traitement du signal non stationnaire

La stationnarité des signaux est une propriété importante et souvent invoquée en traitement du signal, sur laquelle repose maintes approches habituelles telles que l'analyse spectrale ou des corrélations, le filtrage, la modélisation de signaux, ou qui simplifient des problèmes tels que l'estimation, la détection ou la classification de signaux. Dans les manuels, la propriété de stationnarité se définit pour des signaux aléatoires comme l'invariance temporelle de leurs propriétés statistiques [Loè62, Doo67, Yag87]. En pratique cependant, on souhaite accepter une notion plus large de la stationnarité qui permet par exemple de dire qu'un signal déterministe périodique est essentiellement stationnaire, ou qu'un signal peut être stationnaire ou non selon l'échelle de temps sur laquelle on l'observe. Les travaux que nous discuterons dans 2.3 portent sur des manières de proposer une approche pragmatique de la stationnarité qui s'accommode de ces variations autour de la définition stricte et permet de décider si un signal est stationnaire ou non, afin de guider l'analyse vers les outils adaptés (stationnaires ou non stationnaires). Un point novateur de ce travail a été de proposer, dans une optique générale temps-fréquence rappelée en 2.1, un cadre opérationnel pour tester la stationnarité qui englobe celle des signaux déterministes ou d'autres variations comme des notions généralisées de la stationnarité (au sens de mes travaux de thèse [C1, J5]) et se définit relativement au temps d'observation. Les solutions proposées reposent en partie sur les signaux substituts (ou *surrogates*) venant de études en physique non-linéaire [TEL$^+$92, SS96] et ils seront d'abord étudiés dans une section spécifique 2.2 [P41, P45, P51], avant de passer à une description des tests de stationnarités qui rend compte des travaux [P18, P19, P20, P24, J9, J10, P27, J11, P31, P34, J14, P42, J16, P45, P50, J21, C5]

De nombreux phénomènes ne sont pas stationnaires, même dans cette acceptation pragmatique. Un champ de recherche concerne donc a contrario l'étude des signaux non stationnaires, visant à leur description, leur représentation, leur modélisation, leur manipulation [Pri81, Fla99, Coh95, Boa02, CHT98]. Nous aborderons dans 2.4 des contributions aux techniques pour représenter ou extraire les évolutions de signaux non stationnaires ; plus spécifiquement un premier travail dans 2.4.1 a été d'améliorer l'état de l'art des représentations quadratiques temps-fréquence [Fla99, Coh95] pour construire au mieux une représentation parcimonieuse localisée [P23, P28, J15] ; le travail dans 2.4.2 s'appuie sur la décomposition modale empirique ou EMD [1] [HSL$^+$98] pour extraire et analyser des modes

1. pour *Empirical Mode Decomposition* en anglais

non stationnaires et des tendances [J20, P49, P52, J22, P61, Js29] avec des applications [P30, P33, P54, P55, P56] dans 2.4.3, en particulier à des signaux de capteurs environnementaux et de consommation d'énergie. Nous finirons cette section en 2.4.4 avec une proposition d'une méthode d'extraction de tendance qui sert cette fois pour des résolutions numériques en simulation aux grandes échelles pour les fluides [J18, P32, P47, P48, Js26].

Ce chapitre va débuter en 2.1 par un rappel du cadre que nous utiliserons fréquemment dans ce mémoire : l'analyse temps-fréquence, avant de faire un état de l'art sur les tests de stationnarité. Il finira par des perspectives en 2.5.

2.1 Cadre d'étude des signaux non stationnaires

2.1.1 L'analyse temps-fréquence

Définition d'école de la stationnarité. Un processus $\{X(t), t \in \mathbb{R}\}$ est dit stationnaire si il vérifie pour tout $\tau \in \mathbb{R}$ que $\{X(t+\tau), t \in \mathbb{R}\} \overset{d}{=} \{X(t), t \in \mathbb{R}\}$, la notation $\overset{d}{=}$ servant à désigner l'égalité de toutes les lois distributions multi-dimensionnelles finies des processus [Loè62, Doo67, Yag87]. Un processus stationnaire a en particulier une covariance $R_X(t,s) = \mathbb{E}\left\{X(t)\overline{X(s)}\right\}$ qui se réduit à une fonction de corrélation du retard entre t et s : $R_X(t,s) = \gamma_X(t-s)$ et, selon le théorème de Wiener-Khintchine, sa densité spectrale de puissance (DSP) est alors la transformée de Fourier de cette fonction : $S_X(\nu) = (\mathbb{F}\gamma_X)(\nu) = \int e^{-i2\pi\nu\tau}\gamma_X(\tau)\mathrm{d}\tau$. On parle souvent de spectre, par simplification de langage. Souvent, on se limite à considérer la stationnarité au second ordre statistique (ou stationnarité au sens large) qui contraint seulement la moyenne et la covariance du processus à être invariantes dans le temps.

Analyse spectrale des signaux non stationnaires et temps-fréquence. L'étude de ce que peut être l'analyse spectrale pour un signal qui n'est pas stationnaire a été l'objet de nombreux travaux depuis ceux de Ville en 1948 [Vil48] et on consultera avec intérêt [Pri81, Fla99, Coh95] comme monographies sur le sujet. Les processus harmonisables au second ordre, selon la définition de Loève [Loè62], (qui peuvent être stationnaires ou non) admettent un spectre dépendant du temps convenablement défini. Le spectre de Wigner-Ville (SWV) en particulier est défini au temps t et pour la fréquence f comme :

$$\mathbf{W}_X(t,f) = \int_{-\infty}^{+\infty} \mathbb{E}\left\{X\left(t+\frac{\tau}{2}\right) X^*\left(t-\frac{\tau}{2}\right)\right\} e^{-i2\pi f\tau}\, d\tau. \tag{2.1}$$

Il a pour avantage de se réduire au spectre usuel si le processus est stationnaire : $\mathbf{W}_X(t,f) = S_X(f)$ et donc de porter intrinsèquement dans ce cas l'invariance temporelle du processus.

Un deuxième avantage à utiliser le SWV est son lien avec les distributions quadratiques temps-fréquence qui visent à proposer des représentations de l'énergie du signal en fonction de t et f [Fla99, Coh95]. La distribution de Wigner-Ville (DWV) [MH97] est la version déterministe de (2.1) pour un signal x :

$$W_x(t,f) = \int_{-\infty}^{+\infty} x\left(t+\frac{\tau}{2}\right) x^*\left(t-\frac{\tau}{2}\right) e^{-i2\pi f\tau}\, d\tau. \tag{2.2}$$

et donne un estimateur du SWV, certes peu efficace si il s'applique à un signal aléatoire. D'autres distributions quadratiques temps-fréquence, comme celles de la classe de Cohen [Coh95], offrent alors de meilleurs estimateurs du SWV. Les distributions de la classe de Cohen se formulent comme étant un lissage, à travers un certain noyau, de la DWV [Fla99]. En particulier, le spectrogramme qui est le carré d'une transformée de Fourier à court-terme, est un estimateur fréquemment employé du SWV dans la classe de Cohen et il s'écrit de deux manières équivalentes :

$$S_x^{(h)}(t,f) = \left| \int_{-\infty}^{+\infty} x(s)\, h(s-t)\, e^{-i2\pi f s}\, ds \right|^2 = \int\int_{-\infty}^{+\infty} \Pi(s-t, \xi-\nu) W_x(s,\xi) ds d\xi. \quad (2.3)$$

La deuxième forme montre explicitement qu'un spectrogramme est un lissage temps-fréquence de la DWV et la forme exacte du noyau Π est contrôlée par la fonction d'ambiguïté A_h (double transformée de Fourier de la DWV) de la fenêtre h, selon les relations [Fla99]

$$A_h(\xi,\tau) = \int_{-\infty}^{+\infty} h\left(t+\frac{\tau}{2}\right)\, h^*\left(t-\frac{\tau}{2}\right)\, e^{i2\pi\xi t}\, dt \quad (2.4)$$

$$\Pi(t,\nu) = \int\int_{-\infty}^{+\infty} A_h^*(\xi,\tau) e^{i2\pi(\nu\tau+\xi t)} d\xi d\tau. \quad (2.5)$$

Il dépend donc essentiellement du choix de cette fenêtre $h(\cdot)$ qui contrôle conjointement la résolution temporelle et fréquentielle de cette estimation. Ce réglage permet de fixer la résolution de l'analyse dans l'optique de définir une stationnarité relative aux temps caractéristiques d'observation. Les distributions temps-fréquence offrent aussi une manière de concilier l'analyse de signaux déterministes pour lesquels elles apparaissent comme étant des distributions à interférence réduites, et celle des processus aléatoires pour lesquelles le lissage améliore leurs propriétés en tant qu'estimateur. Ces propriétés nous sont bien utiles pour proposer une approche opérationnelle d'un test de stationnarité relative, car l'on voudra par exemple que la périodicité déterministe soit associée à une invariance temporelle, donc à une stationnarité, si l'échelle d'observation laisse la place à suffisamment d'oscillations. La figure 2.1 montre un exemple de signal avec plusieurs composantes (certaines déterministes, d'autres aléatoires, certaines stationnaires, d'autres non) : 3 composantes déterministes (un sinus de fréquence constante, un sinus de fréquence modulée, un chirp à fréquence linéaire) et une composante stochastique stationnaire (bruit à haute fréquence). Les autres figures montrent différentes estimations du spectre temps-fréquence, codées en niveau de couleur dans le plan temps (en abscisse) et fréquence (en ordonnées, le blanc étant les valeurs faibles et le noir les valeurs élevées.

Les spectrogrammes multi-fenêtres. Afin d'améliorer les propriétés des spectrogrammes en tant qu'estimateurs, on peut employer la méthode multi-fenêtres, initialement proposée par Thomson pour l'estimation d'un spectre stationnaire [Tho82] et étendue au spectrogramme dans [BB00, XF07]. Un spectrogramme multi-fenêtres s'écrit :

$$S_{x,K}(t,f) = \frac{1}{K} \sum_{k=1}^{K} S_x^{(h_k)}(t,f) \quad (2.6)$$

mettant en jeu une famille de spectrogrammes ordinaires

$$S_x^{(h_k)}(t,f) = \left| \int_{-\infty}^{+\infty} x(s)\, h_k(s-t)\, e^{-i2\pi f s}\, ds \right|^2 \quad (2.7)$$

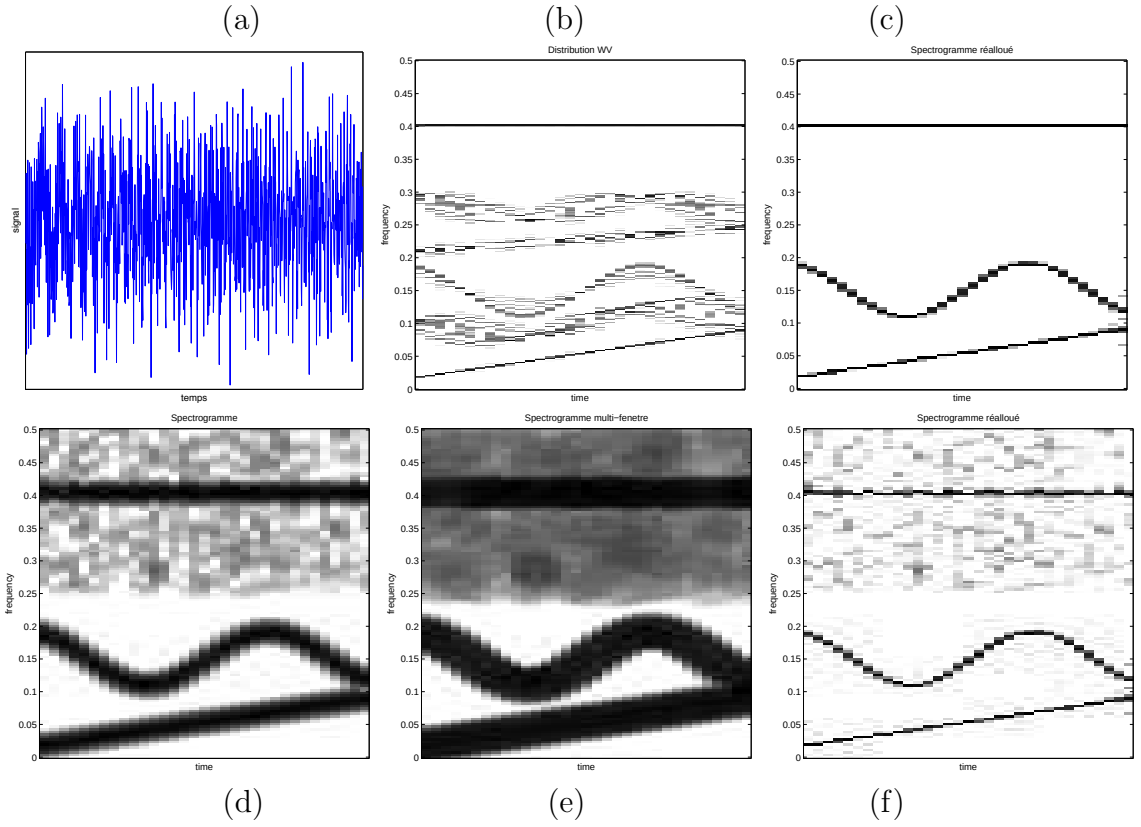


FIGURE 2.1 – Exemple de signal en (a) avec plusieurs composantes. (b) Distribution de Wigner-Ville (sans le bruit) ; elle n'est pas facilement lisible à cause d'interférences 2 à 2 entre les traces des 3 composantes ; on n'a pas mis le bruit car alors la distribution temps-fréquence obtenue ainsi est illisible. (c) Spectrogramme réalloué (sans le bruit) selon [AF95] : distribution presque exacte. En 2e ligne : avec la composante de bruit. (d) Spectrogramme classique, de résolution fréquentielle réduite par rapport à (b) et (c). (e) Spectrogramme multi-fenêtre selon [XF07], résolution fréquentielle un peu moins bonne mais un spectre estimé pour la composante stochastique meilleur (plus constant en temps). (f) Spectrogramme réalloué : assez bonne représentation, si ce n'est que le spectre bruit n'est pas lisse. Pour combiner les avantages de (c) et (e), on peut employer la méthode de [XF07].

pour lesquels les fenêtres à court-terme $h_k(t)$ sont des fonctions d'Hermite. Le choix de ces fonctions vient de ce qu'elles forment une base orthogonale de fonction qui sont concentrées de façon maximale dans le plan temps-fréquence, avec une symétrie elliptique (donc ne privilégiant pas le temps par rapport à la fréquence ni inversement). Elles sont définies par

$$h_k(t) = \left((t - \mathcal{D})^k g\right)(t)/\sqrt{\pi^{1/2} 2^k k!}, \tag{2.8}$$

avec $g(t) = \exp\{-t^2/2\}$. En pratique, elles sont calculées itérativement grâce à la formule $h_k(t) = g(t)\, H_k(t)/\sqrt{\pi^{1/2} 2^k k!}$, où les $H_k(t)$ sont les polynômes d'Hermite qui obéissent à la récurrence $H_k(t) = 2tH_{k-1}(t) - 2(k-1)H_{k-2}(t),\ k \geq 2$ avec $H_0(t) = 1$ and $H_1(t) = 2\,t$.

L'intérêt principal des spectrogrammes multi-fenêtres est de permettre une moyenne dans l'estimation du spectre temps-fréquence, là où l'on ne peut pas moyenner plus dans le temps puisqu'on regarde des signaux potentiellement non stationnaires. En combinant des spectrogrammes qui sont peu corrélés les uns aux autres, car les fenêtres d'analyse sont orthogonales, la variance de l'estimation en est réduite. La Figure 2.1 illustre cette propriété.

Ce sera très utile dans la suite car nous quantifierons numériquement les fluctuations de distributions temps-fréquence pour tester la stationnarité de signaux et l'on souhaite que les fluctuations liées aux erreurs dans l'estimation statistique ne soient pas trop grandes. Nous avons aussi employé les méthodes multi-fenêtres pour de l'estimation cepstrale [J14] dans un contexte de traitement de la parole (ce ne sera pas plus évoqué ici).

2.1.2 Stationnaires ou non stationnaire ?

Ainsi qu'indiqué au début du chapitre, la question de savoir distinguer si un signal est stationnaire, et à quelle échelle de temps, est intéressante pour appliquer les méthodes classiques, stationnaires, du traitement du signal. Cependant, l'idée de stationnarité doit pouvoir être utile autant à son cadre naturel stochastique (auquel cas il faut, pour déduire d'une variabilité temporelle constatée l'indice d'une non-stationnarité, savoir justifier que celle-ci est statistiquement significative) qu'à des signaux plus déterministes de nature. Par conséquent, sa considération pratique s'assortit le plus souvent d'aménagements heuristiques. Dire qu'un signal est « stationnaire » mais sans moyen ni de le vérifier, ni de caractériser à quel point c'est juste, est une faiblesse de construction que l'on est trop souvent obligée de suivre.

L'état de l'art des tests de stationnarité est finalement peu étendu compte tenu de l'importance de la question. Parmi les différentes approches dans la littérature, celles qui ont rencontré le plus de succès dans la communauté des séries temporelles (telles que le test de Dickey-Fuller [DF79], le test KPSS [KPSS92] et ses généralisations [HFO04]) sont explicitement basées sur des idées de modélisation, avec une réjection de l'hypothèse nulle de stationnarité liée de façon étroite à une propriété de « racine unité, » cas modèle considéré comme non stationnaire. Les non-stationnarités testées par ce genre de méthodes sont de ce fait assez spécifiques, se réduisant en général à des tendances ou des changements de moyenne. Pour dépasser cette limitation, des méthodes alternatives ont été proposées dans le domaine fréquentiel, en comparant les caractéristiques spectrales de fenêtres adjacentes et en construisant un test statistique pour décider d'une différence significative ou non entre elles [Vat98, Fue05, BI06], ou dans le domaine temps-fréquence [PS69, Mar84, MF85]. Un travail plus récent propose de s'appuyer sur des modèles paramétriques stationnaires de type ARMA [Kay08] et de tester l'adéquation au modèle. D'autres enfin sont plutôt concernés par des détections de changements, ou de ruptures de la stationnarité [LD98, DG02].

L'approche que nous avons mise en place cherche à tester la stationnarité sans modèle a priori, relativement à une échelle d'observation, et pas uniquement en tant que rupture ou changement. Elle se démarque de l'état de l'art par sa flexibilité. Le point de passage obligé est de savoir caractériser ce que peut être un signal stationnaire dans les conditions d'observation du signal étudié, et de prendre en considération les variations et fluctuations compatibles avec la stationnarité. Pour cela, nous proposons d'employer les signaux « substituts. »

2.2 Les signaux substituts pour la modélisation stationnaire non paramétrique

Les signaux substituts comme référence de stationnarité. La technique des signaux « substituts » (*surrogate data* en anglais) [TEL+92, PT94, SS96, SS00] s'inscrit,

à côté du bootstrap [Efr82, ZI04] ou du mélange aléatoire de données [HTF+01], dans le panel des outils pilotés par les données et vient au départ de l'analyse de signaux de physique non-linéaire. Cette technique repose sur une remarque très simple, à savoir que, pour un spectre marginal donné, l'idée de stationnarité évoquée correspond à la situation où la description spectrale n'est attachée à aucune structuration cohérente en temps. Or, si le poids des différentes composantes spectrales d'un signal est mesuré par le module de son spectre de Fourier, c'est dans la phase de celui-ci que sont codées les relations entre composantes pouvant conduire à des comportements temporels structurés (ou à la signature de comportements non-linéaires). Ainsi, un signal stationnaire se différenciant d'un signal non stationnaire de même spectre par une phase spectrale aléatoire, il suffit de rendre aléatoire la phase du spectre d'une observation quelconque pour la « stationnariser ».

Plus concrètement, soit $X(k) = A(k)e^{i\Psi(k)}$ la transformée de Fourier discrète du signal observé $x(n)$, supposé de longueur T. Un substitut $s(n)$ est engendré en remplaçant la phase $\Psi(k)$ par une séquence $\psi(k)$ de variables indépendantes et identiquement distribuées selon une loi uniforme sur $[-\pi, \pi[$, soit

$$s(n) = \frac{1}{T} \sum_k A(k) e^{i\psi(k)} e^{i2\pi nk/T}. \tag{2.9}$$

A notre connaissance, cette technique n'avait jusqu'ici pas été utilisée dans un autre contexte que pour les tests de non-linéarités de signaux associés à des systèmes chaotiques. Seuls les travaux [Key06, MKH07] les considéraient pour leurs propriétés temporelles, mais toujours dans un contexte d'étude de systèmes non-linéaires. Nous avons mis en avant ces signaux substituts comme pouvant rejoindre les autres techniques d'analyse statistique pilotée par les données. Les signaux substituts sont aux rangs des méthodes non paramétriques car il n'est nul besoin de prescrire de modèle ni de paramètres à estimer. On verra ci-dessous qu'ils peuvent cependant servir à la modélisation de signaux stationnaires dans une approche pilotée par les données.

Nous avons prouvé dans [J16, P42] la stationnarité des substituts $s[n]$, introduite intuitivement ci-dessus et illustrée en figure 2.2. Dans cette figure, la colonne de gauche présente un signal « non stationnaire » (en haut), son spectrogramme (au milieu) et la distribution marginale en temps de ce dernier (en bas). La deuxième colonne présente de la même façon les informations relatives à un substitut et la troisième colonne celles correspondant à une moyenne calculée sur 40 substituts du même signal. La quatrième colonne présente enfin la distribution marginale en fréquence qui, par construction, est identique pour les trois spectrogrammes. Ces différentes distributions mettent en évidence une « stationnarisation » au sens où, pour un même spectre marginal, le comportement temporel local a perdu la forte structuration du signal original. Sur la base de ce résultat, il est immédiat de créer autant de substituts (stationnarisés) que l'on opère de « randomisations » sur la phase, ce qui permet la caractérisation d'une distribution d'ensemble de l'hypothèse nulle de stationnarité pour n'importe quel descripteur de stationnarité d'un signal. Ce sera détaillé en section 2.3.

L'algorithme des substituts s'écrit aussi pour des signaux multivariés [PT94]. Soit $X(n) = [x_1(n), ..., x_M(n)]^t$ un signal multivarié de M composantes, $x_j(n)$ étant la $j$ème composante et le temps est $n = 0, ..., N-1$. L'algorithme utilise en entrée les transformées de Fourier de chaque composante, $(\mathbb{F}x_j)(f) = \sum_{n=0}^{N-1} x_j(n)e^{-i2\pi nf/N} = A_{x_j}(f)\, e^{i\Psi_{x_j}(f)}$.

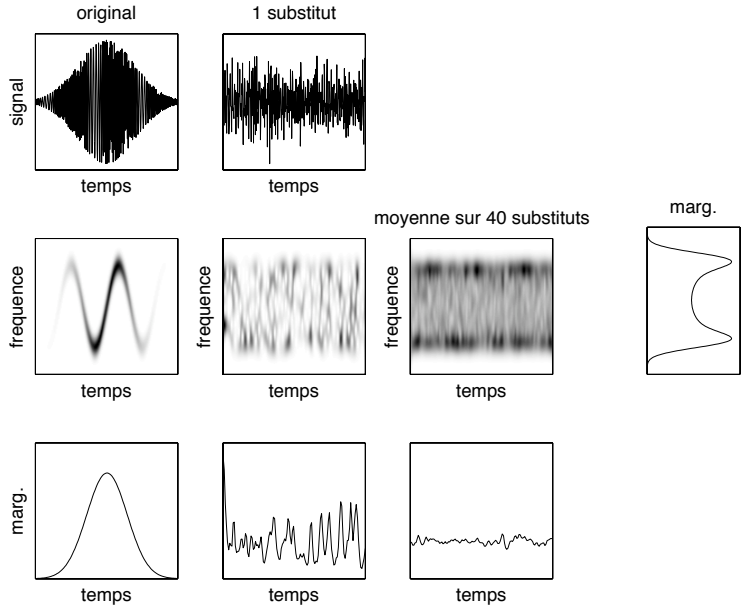


FIGURE 2.2 – Stationnarisation par substituts. On montre en 1e ligne un signal et un substitut, leurs spectrogrammes en 2e ligne et, à droite la moyenne des spectrogrammes sur 40 substituts. La dernière colonne indique la marginale en fréquence, commune à tous ces spectrogrammes par construction. La dernière ligne est la marginale en temps, non stationnaire pour le signal, stationnaire pour les substituts.

ALGORITHME des substituts multivariés :
Entrée $A_{x_j}(f)$ et $\Psi_{x_j}(f)$ pour $j = 1, ..., M$, $f = 0, ..., N-1$.
1. Tirer une phase aléatoire $\Theta(f)$, i.i.d., uniforme sur $[-pi, \pi]$.
2. Indépendamment pour chaque composante j, rendre la phase aléatoire dans le domaine de Fourier en ajoutant $\Theta(f)$ (le même pour chaque j) pour obtenir

$$s_j(n) = \frac{1}{N} \sum_{f=0}^{N-1} A_{x_j}(f) e^{i(\Psi_{x_j}(f)+\Theta(f))} e^{i2\pi nf/N}. \tag{2.10}$$

3. Former le substitut multivarié $S(n) = [s_1(n), ..., s_M(n)]^t$.
Sortie S.

Comme montré en [PT94] à l'aide du théorème de Wiener-Khintchine, les substituts S ont la même structure de covariance (y compris croisée) que le signal X car $(\mathbb{F}s_j)^*(f)(\mathbb{F}s_k)(f) = A_{x_j}(f)A_{x_k}(f)e^{i(\Psi_{x_k}(f)-\Psi_{x_j}(f))}$, qui est donc égal à $(\mathbb{F}x_j)^*(f)(\mathbb{F}x_k)(f)$. Ces signaux enfin sont gaussiens si N est assez grand car ils sont obtenus comme superpositions de modes de Fourier qui ont été rendus indépendants [SS00].

Les signaux substituts pour la synthèse de signaux modèles. Partant de ces propriétés, nous avons mis en avant les signaux substituts comme amenant en fait une méthode générale pour synthétiser des processus aléatoires multi-variés dont on veut prescrire les

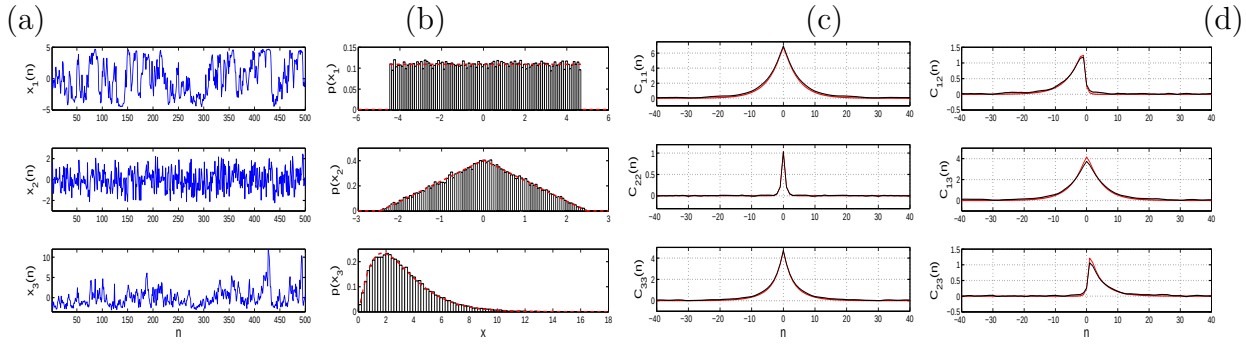


FIGURE 2.3 – Exemple de synthèse de signal multivarié avec l'algorithme IAAFT : covariance d'un processus MA et distributions marginales comme définies dans le texte. (a) extrait des séries temporelles. (b) Estimations empiriques (barres) et prescriptions théoriques (lignes rouges en pointillé) pour les distributions marginales. (c) Auto-corrélations et (d) corrélations croisées pour les substituts (en noir) et le modèle (en rouge). Toutes les estimations empiriques ont été faites avec $N = 2^{15}$ points en temps.

spectres (auto-spectres et spectres croisés) et les marginales de chaque composante, que les prescriptions soient obtenues par des modèles théoriques ou par des mesures empiriques. Utilisant l'algorithme précédent, le point qui manque est de savoir contrôler les distributions marginales et conjointes des composantes. Nous montrerons que la méthode permet même d'aller plus loin que ce que l'on sait faire avec des méthodes traditionnelles de synthèse de processus stationnaires, en prescrivant en particulier les probabilités conjointes (à un temps) entre les composantes, situation pour laquelle il n'existait pas d'outils à notre connaissance, sauf les travaux [SA09, HPA11b] inspirés des travaux de Price [Pri58] ou le travail [ABA13] qui est pour l'instant difficile à déployer sur des séries temporelles dans les cas multivariés. L'intérêt de la méthode issue des substituts est qu'elle évite l'approximation numérique nécessaire dans [HPA11b] pour calculer une transformation inverse entre la distribution gausienne et des marginales quelconques, qui n'est pas toujours faisable ou précise.

La méthode de [SS00] généralise les substituts **IAAFT** (Iteratively Amplitude Adjusted Fourier Transform) [SS96] au cas multi-varié en contrôlant les distributions marginales. Nous discuterons ensuite le travail de [P51] montrant qu'on peut améliorer le résultat en prescrivant la distribution jointe des composantes. Les substituts IAAFT contrôlent les marginales en partant des substituts classiques puis en alternant une étape de projection vers les marginales voulues avec une étape de projection vers la structure de covariance (et donc spectrale) voulue. Pour cette dernière, elle vient de la méthode des substituts, en remettant à l'itération l les amplitudes en Fourier visées A_{x_j}, tout en gardant la phase de l'itération en cours $\Psi_{r_j^{(l)}}(f)$. Cela se fait en deux étapes :

$$(\mathbb{F}r_j^{(l)})(f) = \sum_{n=0}^{N-1} r^{(l)}(n) e^{i2\pi\frac{nf}{N}} = A_{r_j^{(l)}}(f)\, e^{i\Psi_{r_j^{(l)}}(f)} \tag{2.11}$$

$$s_j^{(l)}(n) = \frac{1}{N}\sum_{f=0}^{N-1} A_{x_j}(f) e^{i\Psi_{r_j^{(l)}}(f)} e^{-i2\pi\frac{nf}{N}}. \tag{2.12}$$

Puis les marginales sont contrôlées composante par composante par un plongement ordonné des valeurs courantes du substituts $s_j^{(l)}(n)$ vers l'ensemble ordonné des valeurs $v_j = \mathrm{sort}(x_j)$

de la distribution cible $\{x_j, j = 1, ..., N\}$:

$$r_j^{(l+1)}(n) = v_j(\text{rank}(s_j^{(l)}(n))). \tag{2.13}$$

Rappelons que la fonction **rank** d'une valeur de la série s_j est defini comme $\text{rank}(s_j(n)) = k$ si $s_j(n)$ est la k-ème plus petite valeur de s_j. On stoppe les itération dès que les substituts multivariés $R = [r_1, ..., r_M]^t$ et $S = [s_1, ..., s_M]^t$ sont assez proches l'un de l'autre (ou ne se rapprochent plus si les contraintes sont infaisables). Cet algorithme utilise en fait des projections alternées sur des ensembles convexes [TC84] et on sait qu'ils convergent au mieux possible vers une solution respectant les deux contraintes si elles sont compatibles, la solution dépendant du point de départ initial (fixé ici par le tirage des phases $\Theta(f)$).

Il ne semble pas avoir été réalisé dans la littérature avant [P51] que l'on peut employer cet algorithme de deux manières :
(1) classiquement comme une **méthode de ré-échantillonnage** de données empiriques X, conservant leurs propriétés (marginales empiriques, spectres de Fourier) ;
(2) comme une **méthode de synthèse** à partir d'un modèle prescrivant les marginales (et il suffit de tirer les valeurs v_j à partir des ces marginales puis de les ordonner) et la matrice de covariance stationnaire souhaitée $C_{jk}(n)$ (pour $n = 0, ..., N-1$ et $i, j = 1...M$) pour laquelle on sait engendrer des séries multivariées gaussiennes. L'état de l'art pour cela est de s'appuyer sur la méthode de la matrice circulante [WC94a, DN93, HPA11a].

Toute combinaison de ces deux manières est possible (marginales empiriques et covariances théoriques par exemple). Un exemple utilisant les substituts IAAFT est montré en Fig. 2.3. Il simule un processus multivarié non gaussien à moyenne ajustée (MA) d'ordre 1 et de dimension $M = 3$. Il est donné par sa fonction de covariance $C(n)$, prescrite par la récurrence du MA : $X(n+1) = \Phi * X(n-1) + E(n)$ où E est un bruit blanc gaussien, i.i.d., de variance 1 et $\Phi = [[0.8\ 1.0\ 0.0]; [0.0\ 0.2\ 0.0]; [0.2\ 1.0\ 0.5]]$. Les marginales sont prises avec p_1 uniforme, p_2 selon une distribution triangulaire, p_3 selon une loi gamma avec $\alpha = 2.2$ and $\beta = 1.45$. Nous avons bien fait attention d'avoir $C_{jj}(0) = \text{Var}(p_j)$ pour assurer la compatibilité des contraintes. Les moyennes sont ajoutées à la fin pour S car les substituts IAAFT conduisent à des signaux centrés (alors que p_3 est à moyenne non nulle). On voit sur la figure que les 3 lois marginales, les 3 auto-spectres et les 3 spectres croisés (qui ne sont pas symétriques) suivent très bien les formes voulues.

Synthèse de signaux prescrits en lois conjointes à un temps et en covariances. Comme annoncé, le travail de [P51] montre qu'on peut rendre la méthode plus puissante en prescrivant aussi les distributions conjointes entre les composantes. Il suffit de remplacer le plongement unidimensionnel de (2.13) opéré sur chaque marginale, par un plongement global de $S^{(l)}$ vers la distribution conjointe souhaitée. Pour ce faire, il faut réaliser (2.13) est en fait un transport optimal unidimensionnel et qu'on dispose grâce à [RPDB11] de la construction d'une approximation d'un transport optimal en plus grande dimension.

Rappelons que le transport optimal entre deux distributions Y_k and Z_k, $k = 1, ..., N$, est une fonction $k \to \sigma^*(k)$ (où $\sigma^* \in \Sigma_N$, l'ensemble des permutations à N éléments) qui minimise une distance entre les distributions. On considère ici la distance quadratique de Wasserstein :

$$W_\sigma(Y, Z)^2 = \sum_k ||Y_k - Z_{\sigma(k)}||^2. \tag{2.14}$$

La solution à ce problème est normalement obtenue par programmation linéaire, ce qui est trop coûteux si N est grand. Une métrique alternative, appelée « distance de Wasserstein en tranches » a été proposée dans [RPDB11] :

$$\tilde{W}_\sigma(Y,Z)^2 = \int_{\theta\in\Omega} \min_{\sigma_\theta\in\Sigma_N} \sum_k ||\langle Y_k - Z_{\sigma_\theta(k)},\theta\rangle||^2 \mathrm{d}\theta, \tag{2.15}$$

où σ_θ est le transport optimal pour les points projetés sur la ligne définie par le vecteur unité $\theta \in \Omega = \{u \in \mathbb{R}^M, \text{s.t.}||u|| = 1\}$. Sur une ligne, le transport optimal est justement le plongement ordonné des valeurs de l'équation (2.13).

Il suit de cela qu'une descente de gradient stochastique rend possible de minimiser $\tilde{W}_\sigma(Y,Z)^2$ et de donner une approximation à un transport optimal multivarié. Partant de Y, une direction aléatoire θ_k est tirée à chaque itération et la descente de gradient est menée selon

$$Y^{(k+1)} = Y^{(k)} - \eta_k \left(Y^{(k)} - \langle Z_{\sigma^*_{\theta_k}}, \theta_k\rangle \right) \tag{2.16}$$

où $\sigma^*_{\theta_k}$ est le plongement ordonné optimal de $\langle Z,\theta_k\rangle$ vers $\langle Y^{(k)},\theta_k\rangle$. La convergence est discutée dans [RPDB11] et est convenable pour $\eta_k \leq 1$ en pratique. Cette procédure donne accès à la distance optimale de l'éq. (2.15), mais aussi à la transformation qui permet de passer de Y à Z pour cette distance. Nous la noterons $\tilde{\sigma}^*_{Y,Z}$ et il ne faut pas oublier qu'on l'obtient par itération d'une descente de gradient stochastique.

ALGORITHME des substituts multivariés avec lois conjointes :
Entrée : Covariance $C_{jk}(n)$ pour $n = 0,...,N-1$ et $j,k = 1,...,M$, et distribution jointe à un temps $P(v_1,...,v_M)$
1. Pour la covariance désirée $C_{jk}(n)$, engendrer une réalisation du signal gaussien X (par une méthode de matrice circulante, cf. [WC94a, DN93, HPA11a])
2. Calculer les amplitudes $A_{x_j}(f)$ et phases $\Psi_{x_j}(f)$ des transformées de Fourier de toutes les composantes, $j = 1,...,M$
3. Tirer N vecteurs indépendants $V(n)$ de la loi conjointe visée $P(v_1,...,v_M)$
4. Initialisation de l'algorithme avec $R^{(1)} = S$ de l'algorithme des substituts multivariés classiques, avec des tirages aléatoires des $\Theta(f)$; voir l'éq. (2.10)
5. Itérer sur l'indice l une procédure IAAFT modifiée :
a. Appliquer les formules. (2.11) et (2.12) pour obtenir $S^{(l)}$
b. Calculer en itérant (2.16) le transport optimal approché $\tilde{\sigma}^*_{S^{(l)},V}$ pour transformer les valeurs des $S^{(l)}$ en celles des V. Résultat : $R^{(l+1)}(n) = V(\tilde{\sigma}^*_{S^{(l)},V}(n))$
6. Itérer a. et b. et stopper si R est proche de S (ou s'ils ne changent plus)
Sortie : R et S

Cet algorithme converge car les algorithmes IAAFT et de calcul de transport optimal approché convergent individuellement et car c'est à nouveau une instance de projections alternées sur des ensembles convexes. Il est possible que les deux contraintes ne soient pas exactement satisfaites en même temps. Dans ce cas, R les suit exactement pour les lois conjointes tandis que S les suit pour la covariance.

Un exemple utilisant cette méthode est montrée en fig. 2.4. Il simule un processus gaussien multivarié de dimension $M = 2$. La covariance $C(n)$ est donnée par des fonctions exponentielles : $C_{jk} = \gamma_{jk} e^{-\alpha_{jk} n}$ de paramètres $\alpha_{11} = 0.5$, $\alpha_{22} = 1$, $\alpha_{12} = 0.7$; $\gamma_{11} = 1$, $\gamma_{22} = 1$, $\gamma_{12} = 0.7$. La distribution jointe $P(x_1,x_2)$ est choisie de marginale p_1 uniforme, p_2

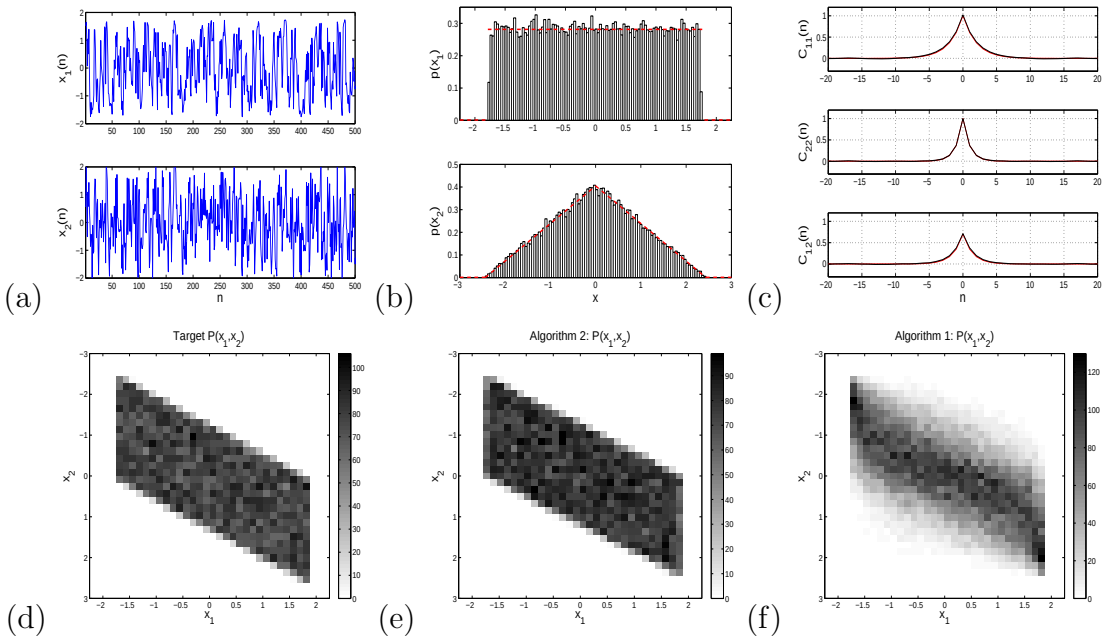


FIGURE 2.4 – Exemple de synthèse de signal bivarié avec lois conjointes prescrites : covariance et distributions marginales comme définies dans le texte. (a) extrait des séries temporelles. (b) Estimations empiriques (barres) et prescriptions théoriques (lignes rouges en pointillé) pour les distributions marginales. (c) Auto-corrélations et corrélation croisée pour les substituts (en noir) et le modèle (en rouge). (d) Distribution jointe prescrite cible (valeurs V tirées pour le point 3 de l'algorithme). (e) Distribution jointe empirique obtenue. (f) Comparaison avec la distribution jointe empirique obtenue avec un algorithme IAAFT classique. Toutes les estimations empiriques ont été faites avec $N = 2^{15}$ points en temps.

triangulaire, et à chaque point la distribution jointe est telle que $x_1(n) = x_2(n) + U(n)$ où U est une variable aléatoire de loi i.i.d. uniforme centrée (ce qui implique que p_2 soit triangulaire). Nous devons fixer $C_{jj}(0) = \mathrm{Var}(p_j)$, et $C_{12}(0) = \int x_1 x_1 P(x_1, x_1) \mathrm{d}x_1 \mathrm{d}x_2$ pour que les contraintes restent compatibles. La fig. 2.4 montre pour cet exemple des réalisations de séries temporelles, les estimations empiriques et les formes prescrites des distributions marginales, des 2 auto-covariances et de corrélation croisée. On ne distingue pas de différence entre les cibles et le résultat. Pour comparer à la méthode IAAFT, on montre la distribution jointe prescrite, celle obtenue et celle des substituts IAAFT usuels ; il est clair que ces derniers ont une distribution jointe qui n'est pas contrôlée et est différente de ce que permet de prescrire la nouvelle méthode.

Pour conclure sur ce travail, nous disposons donc d'un méthode efficace pour générer autant de réalisations que voulues de processus aléatoire stationnaires dont on peut contrôler plus de propriétés que ce que les méthodes précédentes permettaient Une boîte à outil Matlab implémentant la méthode a été mise à disposition le web.

Les substituts de transition. Dans [P45], nous proposons de nouvelles formes de *substituts de transition*, qui permettent de contrôler finement leur degré de stationnarité. L'intérêt est de pouvoir régler au plus juste un niveau de détection de déviations acceptables à la stationnarité dans l'utilisation des tests de stationnarités qui sont décrits plus loin. Le principe est toujours de modifier la phase $\Psi(k)$ du signal initial mais en conservant un peu de

corrélation en écrivant $\Psi(k) = \Psi(k) + \Theta(k)$ où $\Theta(k)$ est un bruit de phase qui peut être moins mélangeant que le bruit blanc uniforme sur $[-\pi, \pi[$ des précédents substituts. Prenant par simplicité un bruit blanc gaussien i.i.d. de variance σ^2, on montre que la distribution spectrale fréquence-fréquence (double transformée de Fourier de la covariance à 2 temps) du substitut ainsi obtenu diffère par le terme multiplicatif $\Lambda(k, k') = \exp\{\sigma^2(\delta(k - k') - 1)\}$ de celle du signal initial. Ainsi, quand σ^2 est nul, on retrouve le signal de départ et, quand σ^2 tend vers l'infini, $\Lambda(k, k')$ devient une masse de Dirac $\delta(k - k')$ qui correspond à supprimer tous les termes non stationnaires dans le substitut. Entre les deux, σ offre un contrôle fin de son degré de stationnarité.

Les substituts en tant qu'aide à l'estimation spectrale. Finalement, nous avons montré dans [P41] comment utiliser les substituts dans l'estimation du spectre temps-fréquence d'un processus aléatoire uniformément modulé. Partant du signal, nous estimons d'un côté la fonction de modulation à l'aide d'EMD (voir plus loin, en section 2.4) et de l'autre le spectre stationnaire du processus en rejouant celui-ci par ses *surrogates*. Ce ne sera pas détaillé plus avant ici.

2.3 Tester la (non) stationnarité

Stationnarité dans un cadre temps-fréquence. En s'appuyant sur le cadre temps-fréquence et les signaux substituts, nous avons proposé des méthodes pour tester la stationnarité d'un signal relativement aux paramètres d'observation, avec des tests quantitatifs non paramétriques. Nous avons décliné dans différentes communications [J9, J10, J11, J16, P18, P19, P20, P24, P27, P31, P34, P42, P45] le principe de base décrit ici et différentes variantes et adaptations à des situations spécifiques, et nous avons écrit des textes tutoriels sur cette approche [J21, C5]. Une partie de ce travail a été réalisée lors de la thèse de Jun Xiao, co-encadrée par P. Flandrin en co-tutelle avec l'ECNU de Shanghai.

Le cadre temps-fréquence est idéal pour notre objectif car il permet d'avoir une échelle de temps globale qui est le temps d'observation et une échelle de temps locale contrôlable par construction des distributions temps-fréquence. Une fenêtre h longue pour un spectrogramme ordinaire selon (2.3) représente en effet le signal avec une faible résolution temporelle (mais forte pour la résolution fréquentielle) ; inversement, une fenêtre courte mettra facilement en exergue les évolutions rapides. Il est donc possible de tester relativement à la longueur de cette fenêtre si le signal est stationnaire ou non. Le second élément permettant le test est que les distributions temps-fréquence se réduisent à une distribution invariante dans le temps, égale au spectre usuel si un signal est stationnaire ; on peut dès lors tester la stationnarité en testant si la distribution reste localement assez semblable à la distribution globale, en construisant un descripteur de contraste entre le spectre local et le spectre global. Une telle approche avait déjà été suggérée dans [MF85, Mar84] mais nous avons ajouté à cela la possibilité de formuler un test d'hypothèse.

Le test de stationnarité relative que nous proposons se fonde sur la comparaison des spectres locaux $S_{x,K}(t_n, f)$ (obtenus pour une séquence de N instants t_n répartis sur l'intervalle d'observation T avec un espacement proportionné à la taille des fenêtres à court-terme)

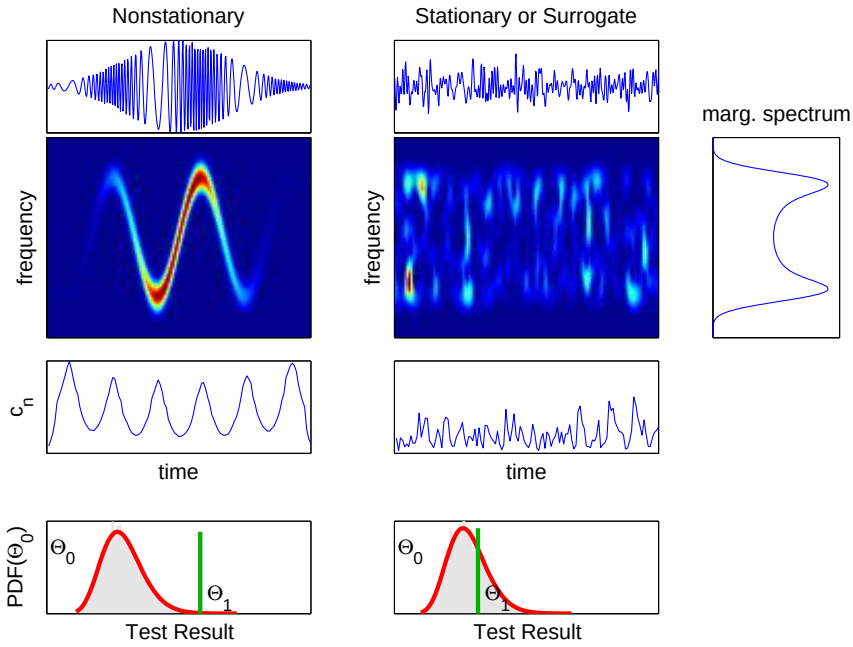


FIGURE 2.5 – Test de stationnarité. De haut en bas : le signal y, son spectrogramme multi-fenêtre, la mesure locale de contraste c_n puis la statistique du test $\Theta(y)$. À gauche : pour le signal initial (non stationnaire) ; à droite : pour un substitut (stationnaire) En dernière ligne, on figure la distribution empirique des $\Theta(j)$ (pour les substituts) en gris (ajustée par une loi Gamma en rouge). On voit que $\Theta(y)$ (indiqué par la ligne verte) est en-dehors de la distribution à gauche et dedans à droite pour un signal stationnaire.

au spectre global défini par la marginalisation :

$$\langle S_{x,K}(t_n, f)\rangle_n = \frac{1}{N}\sum_{n=1}^{N} S_{x,K}(t_n, f). \tag{2.17}$$

Sachant qu'il y aura forcément des différences entre les spectres locaux et celui global même dans un cas stationnaire, que ce soit par des fluctuations d'origine statistique dans l'estimation (dans un cadre de signal aléatoire) ou des fluctuations dues à la position du signal dans la fenêtre (dans un cadre déterministe), il faut et il suffit de proposer des manières de tester si l'ensemble des $\{S_{x,K}(t_n, f),\ n = 1, ..., N\}$ est statistiquement compatible avec le spectre global moyen $\langle S_{x,K}(t_n, f)\rangle_n$.

Pour cela, nous avons développé deux approches : la première est basée sur une notion de « distance » entre spectres et l'autre sur des idées issues de la théorie de l'apprentissage statistique. Dans les deux cas, les substituts permettent de construire un ensemble de comparaison (d'apprentissage dans le 2e cas) qui permet de caractériser ce que sont les signaux stationnaires et quelles différences entre global et local sont acceptables. Cela définit ainsi un test d'hypothèse quant à la stationnarité relative.

2.3.1 Test basé sur des distances

La littérature offre une très grande variété de mesures de dissimilarité entre spectres [Bas89]. On a pu montrer qu'un choix raisonnable pouvait être fait en considérant les mesures

les plus simples ayant déjà fait leurs preuves dans des contextes similaires. Plus précisément, la « distance » retenue entre deux spectres $G(f)$ et $H(f)$ définis sur un intervalle fréquentiel Ω est de la forme

$$\kappa(G,H) := \kappa_{\mathrm{KL}}(\tilde{G},\tilde{H}).\left(1+\kappa_{\mathrm{LSD}}(G,H)\right), \tag{2.18}$$

combinant la divergence de Kullback-Leibler symétrisée

$$\kappa_{\mathrm{KL}}(\tilde{G},\tilde{H}) := \int_{\Omega}\left(\tilde{G}(f)-\tilde{H}(f)\right)\log\frac{\tilde{G}(f)}{\tilde{H}(f)}\,df \tag{2.19}$$

appliquée aux spectres normalisés $\tilde{G}(f)$ et $\tilde{H}(f)$ issus de $G(f)$ et $H(f)$, et la déviation log-spectrale

$$\kappa_{\mathrm{LSD}}(G,H) := \int_{\Omega}\left|\log\frac{G(f)}{H(f)}\right|df. \tag{2.20}$$

L'intuition derrière ce choix est qu'une large famille de non-stationnarités peut être décrite par une modélisation de type AM-FM (modulations en amplitude et en fréquence). La divergence de Kullback-Leibler étant essentiellement une mesure de dissimilarité entre formes spectrales normalisées, elle est par nature bien adaptée à la mise en évidence de structures FM mais, du fait de la normalisation des spectres, elle est insensible à un caractère purement AM. Celui-ci est par contre pris en charge par la déviation log-spectrale, justifiant l'usage combiné des deux mesures [P20, J10].

Le test proprement dit passe alors par l'application de cette mesure de dissimilarité entre les spectres locaux et le spectre global associé, c'est-à-dire par l'évaluation de quantités

$$\{c_n^{(y)} := \kappa\left(S_{y,K}(t_n,.),\langle S_{y,K}(t_n,.)\rangle_n\right), n=1,\dots N\} \tag{2.21}$$

pour les signaux $y(t)$ correspondant tant à l'observation à tester ($y(t)=x(t)$) qu'à la collection de ses substituts ($y(t)=s_j(t); j=1,\dots,J$).

La stationnarité étant supposée correspondre à une égalité entre les spectres locaux et le spectre global, on mesure un écart éventuel à celle-ci *via* les fluctuations en temps des mesures de dissimilarité (2.21). Rapportant ces fluctuations à leur valeur moyenne définie par

$$\langle c_n^{(y)}\rangle_n = \frac{1}{N}\sum_{n=1}^{N} c_n^{(y)}, \tag{2.22}$$

le choix le plus simple consiste à faire usage de la distance quadratique, conduisant à l'évaluation de la statistique de décision

$$\Theta(y) = \frac{1}{N}\sum_{n=1}^{N}\left(c_n^{(y)}-\langle c_n^{(y)}\rangle_n\right)^2 \tag{2.23}$$

pour le signal testé $x(t)$ et les J substituts $s_j(t)$, $j=1,\dots,J$.

Dans la mesure où, comme on l'a dit précédemment, il est facile de générer autant de substituts stationnaires que l'on veut, il est alors possible d'accéder à la distribution empirique du descripteur de fluctuations (2.23) conditionnellement à l'hypothèse nulle de stationnarité, et ainsi de caractériser celle-ci. Ce faisant, pour une erreur de première espèce prescrite, on peut identifier un seuil γ à partir des $\Theta(s_j)$, et rejeter ou accepter l'hypothèse de stationnarité selon que la condition $\Theta(x)>\gamma$ est satisfaite ou non. La figure 2.5 résume

la méthode, partant d'un signal stationnaire ou non et montrant comment $\Theta(x)$ se compare à la distribution obtenue pour les substituts. Le test et la fonction de décision $d(x)$ s'écrivent finalement comme un test unilatéral :

$$d(x) = \begin{cases} 1 & \text{si} \quad \Theta(y) > \gamma \quad : \text{«non stationnaire »;} \\ 0 & \text{si} \quad \Theta(y) < \gamma \quad : \text{« stationnaire ».} \end{cases} \tag{2.24}$$

Modélisation de la loi sous hypothèse de stationnarité. Sous l'hypothèse nulle, la distribution de la statistique des fluctuations (2.23) est modélisable par une loi Gamma [P19]. Ceci peut se comprendre par la structure quadratique de la mesure choisie et le caractère fortement mélangeant des pré-traitements conduisant aux grandeurs sur lesquelles cette mesure opère. L'intérêt de ce résultat est que la charge de calcul attachée au calcul de substituts peut être significativement réduite en ramenant un problème d'évaluation empirique de densité par histogramme à une modélisation à deux paramètres pouvant être conduite, par exemple, au sens du maximum de vraisemblance. On a pu noter en ce sens, qu'à performances comparables, la seconde approche nécessite plusieurs ordres de grandeurs de moins que la première quant au nombre de substituts à utiliser [J10, P19, P20].

Reproduction de l'hypothèse nulle. Dans la mesure où le test proposé est essentiellement un test de rejet de l'hypothèse nulle de stationnarité, il convient de s'assurer d'une reproduction convenable de cette dernière dans le cas où l'observation est effectivement stationnaire. Les études conduites en ce sens ont montré que le taux d'erreur de première espèce observé était légèrement plus important que la valeur prescrite, conduisant ainsi à un test pessimiste [J10, J16]. En utilisant les *substituts de transition*, ce problème est évité [P45]. On a pu caractériser le contrôle à apporter au signal de phase des substituts pour améliorer les performances de reproduction de l'hypothèse nulle sans sacrifier celles de détection.

Caractérisation de la force des non-stationnarités. Bien que le test soit binaire, la valeur de la statistique $\Theta(x)$ apporte des informations complémentaires quant à l'importance éventuelle de la non-stationnarité détectée. Un sous-produit de la détection est en particulier la définition possible d'un *indice* de non-stationnarité en rapportant $\Theta(x)$ à sa valeur moyenne obtenue pour les substituts :

$$\text{INS} := \sqrt{\frac{\Theta(x)}{\frac{1}{J}\sum_{j=1}^{J}\Theta(s_j)}}. \tag{2.25}$$

De plus, si le test est par définition relatif à l'échelle d'observation définie par la durée T du signal analysé, il est aussi fonctionnellement dépendant de la taille T_h des fenêtres à court terme permettant de contraster les spectres locaux et le spectre global. La conséquence en est que l'on dispose d'un degré de liberté supplémentaire, le test pouvant être conduit pour plusieurs tailles de fenêtres. Ceci offre alors la possibilité de définir une *échelle* typique de non-stationnarité (ENS) selon :

$$\text{ENS} := \frac{1}{T}\,\arg\max_{T_h}\left\{\text{INS}(T_h)\right\}. \tag{2.26}$$

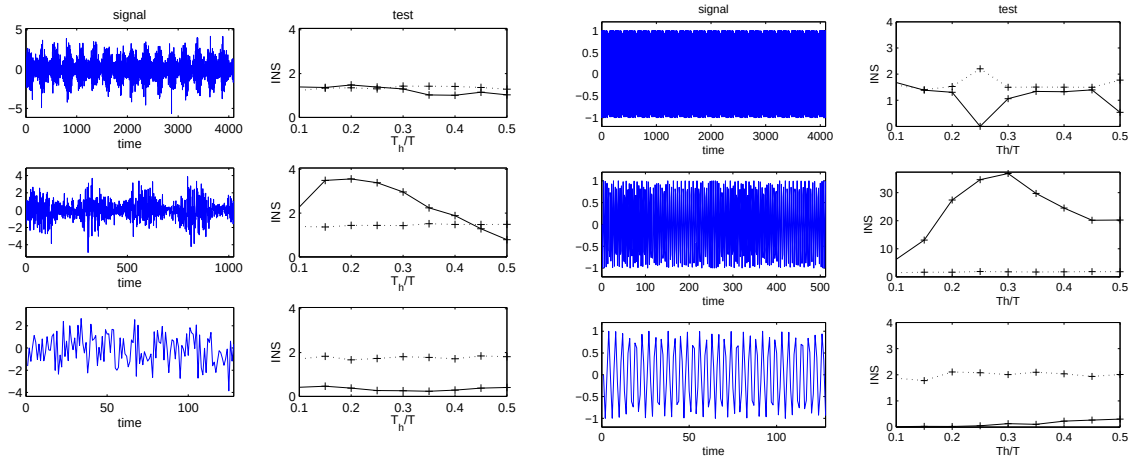


FIGURE 2.6 – Deux exemples : AM ($m = 0.5$) selon (2.27) à gauche et FM ($m = 0.02$, $f_0 = 0.25$) selon (2.28) à droite. Les indices de non-stationnarité INS (colonnes 2 et 4, ligne pleine) sont cohérents avec l'interprétation physique dépendante des échelles d'observation : macro (haut), méso (milieu) et micro (bas). À l'échelle mésoscopique, les signaux sont vus comme non stationnaires tandis qu'ils sont stationnaires aux deux autres échelles. Le seuil γ (ligne pointillée) du test de (non) stationnarité est calculé avec un seuil de vraisemblance à 95% et est représenté en tant que INS selon $\sqrt{\gamma/\langle\Theta_0(j)\rangle_j}$, avec $J = 50$. Dans le cas non stationnaire, la position du maximum de INS donne une indication de l'échelle typique de non-stationnarité.

Exemple de test relativement à l'échelle d'observation. À l'aide de signaux simples, considérons des situations qui peuvent être stationnaires ou non selon l'échelle d'observation. Les deux manières les plus simples pour un signal d'être non stationnaire sont d'avoir son amplitude modulée au cours du temps, ou d'avoir une fréquence instantanée variant dans le temps. Le premier cas sera regardé à l'aide d'une réalisation d'un signal aléatoire modulé en amplitude (AM) :

$$x(t) = (1 + m \sin 2\pi t/T_0)\, e(t), t \in T, \tag{2.27}$$

où $m \leq 1$ et $e(t)$ est un bruit blanc gaussien, T_0 est la période de la modulation et T la durée d'observation. La seconde situation est illustrée avec un modèle déterministe de signal modulé en fréquence (FM), défini ainsi :

$$x(t) = \sin 2\pi(f_0 t + m \sin 2\pi t/T_0), t \in T, \tag{2.28}$$

où $m \leq 1$, f_0 est la fréquence centrale du FM, T_0 as période et T la durée d'observation. A ce modèle on peut ajouter un bruit blanc gaussien pour obtenir différentes réalisations du signal.

Ces exemples et les résultats des tests sont montrés sur la Figure 2.6. Dans les deux cas, l'allure du signal et le résultat du test dépendent des valeurs relatives entre T_0 et T, avec trois régimes :

1. si $T \gg T_0$ (observation à une échelle macroscopique), on voit beaucoup d'oscillations ce qui engendre une quasi-périodicité qui correspond à la stationnarité opérationnelle qui nous intéresse ;
2. si $T \approx T_0$ (échelle mésoscopique), on observe principalement les évolutions locales

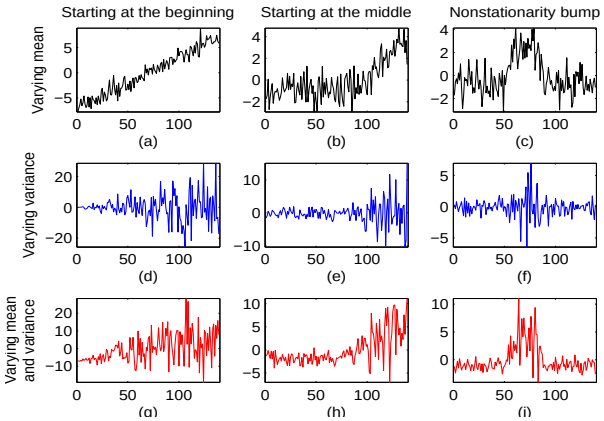


FIGURE 2.7 – Exemples de signaux avec une moyenne non stationnaire ((a),(b),(c)), une variance non stationnaire ((d),(e),(f)) ou les deux ((g),(h),(i)). De la 1e à la 3e colonne changent les moments où les évolutions se produisent : dès le début, après le milieu, milieu seulement.

Evolution dès le début			Evolution après le milieu			Nonstation. au milieu		
μ	σ	μ, σ	μ	σ	μ, σ	μ	σ	μ, σ
Distance d'Itakura-Saito (% résultats corrects)								
23.6	100	100	100	100	100	98.7	100	100
Distance combinée (% résultats corrects)								
0	100	100	0	0.2	100	0	99.4	94.4

FIGURE 2.8 – Pourcentage de résultats "non-stationnaire" (%) obtenus par le test proposé sur les signaux de Fig. 2.7 (2500 répétitions). Les résultats avec la distance d'Itakura-Saito utilisent l'approche où l'on normalise les spectres par la marginale.

dues aux modulations et le signal a l'allure d'un signal non stationnaire avec une échelle typique de non stationnarité de l'ordre de T_0 ;

3. si $T \ll T_0$ (échelle microscopique), le temps d'observation est trop court pour voir une évolution et on est à nouveau face à une portion de signal stationnaire.

La Fig. 2.6 montre que ces interprétations intuitives d'une stationnarité opérationnelle et relative coïncident justement avec ce que le test révèle. Par exemple, le test indique qu'un signal est considéré comme stationnaire que ce soit par accumulation de motifs répétés (voire même périodiques et déterministes) qui s'accumulent (cas AM ou FM macroscopique) ou par stationnarité au sens stochastique plus usuel (AM microscopique), ou bien le cas où on se retrouve avec un signal périodique à très faible bande spectrale (FM microscopique, qui tend vers un sinus simple). Cela renforce l'aspect opérationnel de la stationnarité telle que vue à travers ce test. De la même façon, le test est bien relatif car le résultat varie selon la comparaison entre T et T_0.

Quand l'échelle d'observation est mésoscopique, l'hypothèse nulle de stationnarité est bien rejetée (ligne de milieu), avec un index et une échelle de non-stationnarité définis selon les équations (2.25) et (2.26). La valeur maximum de l'INS est obtenue pour ENS $= T_h/T \approx 0.2$, en accord qualitatif avec la présence de 4 périodes de modulation à l'intérieur de la fenêtre d'observation à cette échelle. Il faut d'ailleurs noter que la longueur de fenêtre utilisée pour les spectrogrammes est bien un degré de liberté de la méthode et qu'il est utile de le varier de manière à trouver les échelles où le signal est stationnaire, ou alors où il est maximalement non stationnaire.

Finalement, on aurait pu considérer le modèle AM comme étant celui d'un signal cyclo-stationnaire et employer les outils dédiés [SPSG05]. Cependant, l'approche proposée ici ne fait aucune hypothèse sur le signal et n'a par exemple pas besoin de savoir qu'il existe une périodicité dans les corrélations. La détection de cette périodicité à une certaine échelle T_0 est en fait ici un résultat du test proposé.

Extension aux non-stationnarités de premier ordre. Dans [P50], j'ai repris le test avec des collègues du GIPSA-lab pour l'étendre aux non stationnarités du premier ordre (en moyenne) alors que le test initial supposait de s'intéresser aux (non) stationnarités au 2e ordre statistique. En effet, une évolution dans la moyenne apparaît dans une représentation temps-fréquence comme une évolution à la fréquence nulle ou très basse. Généralement, la résolution manque pour réussir à détecter une telle évolution. Pour corriger cela, il a été proposé de considérer une autre distance que celle proposée en (2.18) et de se tourner vers la distance d'Itakura-Saito :

$$D_{IS}(G,H) = \int_\Omega \left[\frac{G(f)}{H(f)} - \log\frac{G(f)}{H(f)} - 1\right] df. \tag{2.29}$$

Expérimentalement, elle améliore les résultats pour peu que l'on modifie un peu le descripteur temps-fréquence utilisé : au lieu de prendre directement les spectres locaux, on les pondère d'abord en les multipliant par la marginale en temps (distribution sommée sur toutes les fréquences) normalisée à l'unité. L'intérêt est de préserver toutes les propriétés de localisation des spectres et, pour les substituts, d'égalité au spectre global initial, tout en donnant plus d'importance aux temps où il y a plus d'énergie dans le signal. Faisant ainsi, il a été montré dans [P50] qu'on détecte alors sans difficulté les non-stationnarités en moyenne. Par exemple, sur les signaux de la Figure 2.7, les pourcentages de détection de leur évolution non stationnaire sont indiqués dans la table 2.8.

2.3.2 Test formulé par apprentissage

Apprentissage par SVM 1-classe. Une deuxième voie d'approche consiste à considérer la famille des substituts construits à partir du signal observé comme un *ensemble d'apprentissage* de la situation stationnarisée correspondante. Un des intérêts de ce point de vue est qu'il évite le choix d'une mesure de dissimilarité telle que (2.18) ou (2.29) et d'une statistique de décision associée (2.23). La méthode retenue dans [P18, J16, C5] repose sur la mise en œuvre de machines à vecteurs supports (SVM) [BGV92, Vap95, SSM98] dans le cas de la recherche d'une classe [TD04, SPST+01, STC04, Ver06]. Nous allons décrire cet algorithme de SVM 1-classe.

Considérant un ensemble d'apprentissage $\{s_1,\ldots,s_J\}$ pouvant correspondre, soit aux substituts eux-mêmes, soit à des descripteurs qui s'en déduisent, on cherche à déterminer l'hypersphère de centre a_0 qui rend compte au mieux du support de la distribution des données selon $a_0 = \arg\min_a \max_{j=1,\ldots,J} \|s_j - a\|^2$. Ceci peut se traduire par le problème d'optimisation [TD04] :

$$\begin{array}{ll} \min_{a,r,\xi} & r^2 + \frac{1}{\nu J}\sum_{j=1}^J \xi_j \\ \text{avec} & \|s_j - a\|^2 \le r^2 + \xi_j, \quad \xi_j \ge 0, \quad j = 1,\ldots,J \end{array} \tag{2.30}$$

où le paramètre $\nu \in]0,1]$ définit un compromis entre la minimisation du rayon r de l'hypersphère et le contrôle de variables de relaxation $\xi_j = [\|s_j - a\|^2 - r^2]_+$ destinées à rendre l'approche plus robuste à la présence éventuelle de données aberrantes. La résolution de ce problème, par la méthode des multiplicateurs de Lagrange, permet de déterminer numériquement le centre a_0 et le rayon r_0 de l'hypersphère recherchée. Il en résulte la statique de décision

$$\Theta(y) = \|y - a_0\|^2 - r_0^2 \tag{2.31}$$

que l'on compare à un seuil γ strictement positif, à définir en fonction de la sensibilité du test recherchée. Si $\Theta(x) \geq \gamma$, le signal testé $x(t)$ – ou ses descripteurs – figure à l'extérieur de l'hypersphère définie grâce aux substituts et est déclaré non stationnaire.

Représentation. Le caractère non paramétrique du test offre d'innombrables possibilités quant au choix de la représentation des substituts et du signal à tester, puisqu'il n'est pas nécessaire ici de modéliser ni de manipuler des densités. Selon le contexte, nous avons été amenés à extraire des attributs tels que les variances temporelles de la puissance (P) et de la fréquence (F) instantanées, comme dans [J16, P18, P34]. Pour cela, nous calculons des variances temporelles de la puissance instantanée et de la fréquence instantanée de chaque substitut, avec

$$P = \mathrm{std}(P_t)_{t=1,\ldots,N}, \quad \text{avec} \quad P_t = \int_0^{\frac{1}{2}} S(t,f)\, df$$
$$F = \mathrm{std}(F_t)_{t=1,\ldots,N}, \quad \text{avec} \quad F_t = \frac{1}{P_t} \int_0^{\frac{1}{2}} f\, S(t,f)\, df$$

où $S(t,f)$ est le spectrogramme normalisé du substitut, et $\mathrm{std}(\cdot)$ désigne l'écart type. À partir des couples [2] (P,F) de chaque substitut, on élabore un détecteur de nouveauté, à classe unique, de type machine à vecteurs supports. On obtient alors une frontière de décision et des courbes d'équiprobabilité de mauvaise attribution d'un signal stationnaire à l'ensemble des signaux stationnaires.

Dans [P31], nous avons également montré qu'on peut considérer les séquences temporelles directement, et/ou appliquer une transformation non linéaire aux données en introduisant un noyau reproduisant dans (2.30)-(2.31).

Caractérisation. Le choix du seuil γ conditionne évidemment les performances du test de stationnarité. Il a été démontré que, avec une probabilité supérieure ou égale à $1-\delta$, on peut borner la probabilité de fausse alarme que le test qualifie un substitut de non stationnaire, par la quantité suivante [J16]

$$\frac{1}{\gamma J} \sum_{j=1}^{J} \xi_j + \frac{6R^2}{\gamma\sqrt{J}} + 3\sqrt{\frac{\ln(2/\delta)}{2J}} \tag{2.32}$$

où R est le rayon de la boule centrée à l'origine contenant le support de la distribution des substituts. Une valeur approchée de cette borne, donnée par le premier terme de l'expression ci-avant puisque les deux suivants tendent vers 0 à mesure que J croît, est indiquée sur la figure 2.9 pour différentes valeurs du seuil γ. Celle-ci fournit une information intéressante sur un signal testé qui serait jugé non stationnaire puisqu'il est possible de la décliner en un indice de non-stationnarité semblable à (2.25). Pour cela, on note que $\Theta(x) = \gamma$ est la valeur seuil pour laquelle x est considéré comme non stationnaire, et que $\xi_j = [\Theta(s_j)]_+$. En prenant l'inverse de la borne approchée évoquée pour que l'indice de non-stationnarité varie inversement par rapport à la probabilité de fausse alarme et, en considérant la racine carrée

2. En pratique, les paramètres P et F sont centrés et réduits.

du résultat pour faire apparaître un rapport de distances, ou écarts types estimés comme dans (2.25), on aboutit à

$$\text{INS} := \sqrt{\frac{\Theta(x)}{\frac{1}{J}\sum_{j=1}^{J}[\Theta(s_j)]_+}}. \tag{2.33}$$

Une échelle de non-stationnarité pourrait être définie à partir de (2.33) comme dans (2.26).

Un exemple. On s'intéresse dans cet exemple à une classe de signaux d'enveloppe temporelle gaussienne et à modulation de fréquence linéaire, définis par

$$x(t) = e^{-\pi\eta t^2}(1 + \alpha\, e^{2j\pi f_0 t})\, e^{j\pi\beta t^2},$$

combinant donc modulations d'amplitude et de fréquence. Leurs proportions relatives, ainsi que le degré de non-stationnarité, sont définis par la pente de modulation β et la largeur de l'enveloppe gaussienne $\delta t = 1/\sqrt{\gamma}$ [P34]. En particulier, pour $\beta = 0$, on note que $x(t)$ se réduit à une modulation d'amplitude. Si l'on se concentre sur la composante

$$x_1(t) = e^{-\pi\eta t^2} e^{2j\pi f_0 t} e^{j\pi\beta t^2}, \tag{2.34}$$

étant entendu que les propriétés de $x(t)$ s'en déduisent directement, on montre qu'on contrôle le type de non-stationnarité de ce signal en modifiant β [P34]. En effet, le spectre de $x_1(t)$ est une fonction gaussienne, entièrement définie par sa largeur de bande en fréquence δf et on montre que $\delta f^2 = (\beta^2 + \eta^2)/\eta$, soit encore

$$\delta f^2 = \beta^2 \delta t^2 + 1/\delta t^2. \tag{2.35}$$

Il est alors possible de générer un ensemble de signaux, paramétrés par $(\delta t, \beta)$, incarnant des degrés et formes de non-stationnarités distincts, mais tous dotés du même spectre global fixé par la largeur de bande δf. Cette classe de signaux décrit ainsi une transition continue de la modulation d'amplitude à la modulation de fréquence en variant $(\delta t, \beta)$ à δf fixé. Il en résulte qu'il leur correspond à tous une même famille de substituts, l'écart à ceux-ci permettant de caractériser la nature de la non-stationnarité dont un signal testé ferait l'objet.

La figure 2.9 illustre dans le cadre du test par apprentissage, les positions dans le plan (P, F) de 100 substituts et la frontière de décision associée, ainsi que la trajectoire parcourue par les signaux $x(t)$ à largeur de bande constante $\delta f = 0.05$, paramétrée par δt et β, avec $f_0 = 0.2$ et $\alpha = 1/2$. Dans cette figure, la frontière associée au seuil $\gamma = 0$ est représentée par une ellipse en trait rouge. Les ellipses vertes illustrent les frontières de décision successives pour différentes valeurs de seuil $\gamma > 0$ associées aux probabilités 0.15, 0.10 et 0.05 de mauvaise attribution d'un signal stationnaire, selon le sens qui en est donné par les substituts, à la classe des signaux non stationnaires. En d'autres termes, elles correspondent à des iso-valeurs de l'indice de stationnarité (2.33). La trajectoire représente les lieux des signaux $x(t)$ testés. Quel que soient les paramètres adoptés pour $x(t)$, chacun de ces signaux est déclaré non stationnaire par le test. De plus, tout en attachant à chacun un degré de non-stationnarité comparable (mesuré par la distance à la zone de stationnarité définie par les substituts), il met en évidence un continuum de comportements allant d'une variance P élevée pour le cas à modulation d'amplitude dominante, à une variance F élevée lorsque la modulation de fréquence devient prépondérante. Cette observation ouvre la voie à une possibilité de caractérisation fine de types de non-stationnarité, par exemple à des fins de classification.

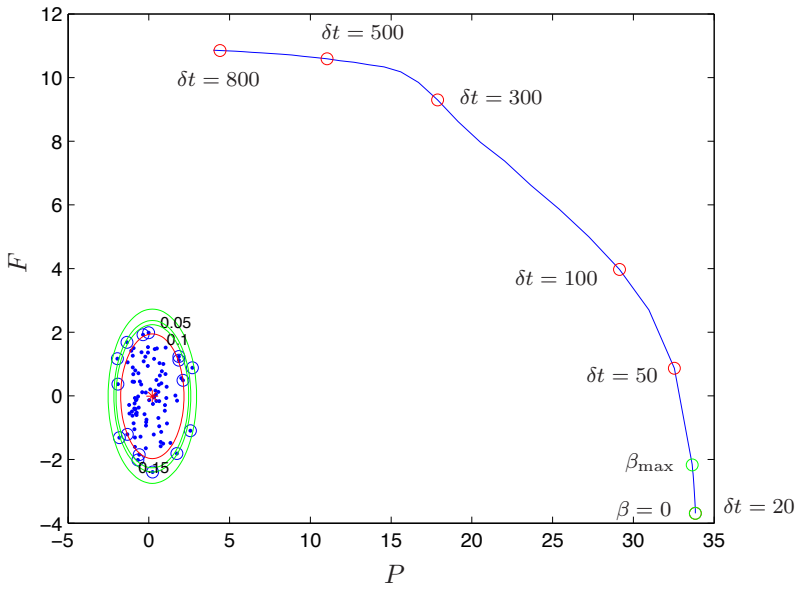


FIGURE 2.9 – Test de stationnarité par apprentissage : représentation de 100 substituts d'un signal $x(t)$ dans le plan (P, F), ceints par la frontière de décision (en rouge) obtenue par résolution du problème (2.30).

Comparaison des approches et développements. L'objectif de la série de travaux qui a été discutée ici était de proposer un cadre versatile pour tester de manière opérationnelle ce qu'est la stationnarité d'un signal et, si il est trouvé non stationnaire, être capable d'avancer vers la caractérisation d'à quel point il n'est pas stationnaire ou du fait de quel type de stationnarité. Nous avons donc proposé deux approches complémentaires qu'il semble bon de comparer. Il faut cependant dire dès le départ que les deux méthodes répondent convenablement au questionnement initial sur la stationnarité opérationnelle et relative. Cependant sur certains aspects les approches diffèrent :

— Méthode basée sur les distances : cette méthode est plus facile à mettre en oeuvre et conduit à une acception très générale de la stationnarité. De plus, étant formulée avec des expressions analytiques, il est envisageable de continuer l'étude théorique de la méthode et peut-être de prouver que la loi statistique gamma utilisé pour le test est la meilleure.

— Méthodes formulée par apprentissage : l'intérêt premier est de permettre une caractérisation plus fine, permettant d'envisager une classification des types de non stationnarités. Cependant il ne faut pas négliger qu'il y a une étape qui n'est pas forcément aisée à maîtriser concernant le choix des descripteurs à retenir. De plus, la machinerie d'apprentissage par SVM 1-classe est telle qu'il n'est pas immédiat à tous de reprendre la méthode à son compte (même si des codes sont accessible à partir de notre site web).

Nous avons développé la méthode de tests de stationnarité vers deux autres directions que nous ne discuterons pas en détail ici. Dans [J11], nous avons employé les formes générales de signaux substituts telles que les substituts multivariés. Nous avons aussi directement formulé des substituts dans le plan temps-fréquence, en prenant la transformée de Fourier

2D de distribution temps-fréquence et en rendant les phases aléatoires avant de revenir dans le domaine temps-fréquence. Grâce à cela, nous avons pu proposer un protocole de détection de forme transitoire cachée dans du bruit et une méthode pour tester l'existence d'inter-corrélations non stationnaires entre deux signaux. Dans [J9], nous avons reformulé la technique en temps-échelle (voir 3.2) de manière à pouvoir aborder la stationnarité dans les images (plutôt appelée en général homogénéité). Nous avons montre que cela permet des tests d'homogénéité d'images relatives au facteur d'échelle d'observation.

2.4 Extraction de modes non stationnaires

Décomposition en modes non stationnaires. Dans plusieurs situations comme celle de la figure 2.1 du début du chapitre, on souhaite représenter directement un signal observé en une somme de modes non stationnaires, chacun de fréquence instantanée et de modulation en amplitude évoluant dans le temps :

$$x(t) = \sum_{k=1}^{K} a_k(t)\, e^{i\varphi_k(t)}. \tag{2.36}$$

En temps-fréquence, la forme idéale du spectre dépendant du temps correspondant serait

$$\rho_x(t,f) = \sum_{k=1}^{K} a_k^2(t)\, \delta\left(f - \dot{\varphi}_k(t)/2\pi\right). \tag{2.37}$$

mais il est connu que (sauf cas particulier), il n'y a pas de méthode générale pour obtenir la décomposition (2.36) ni cette représentation (2.37) [Fla99]. Les travaux discutés ici présentent des méthodes que nous avons étudiées permettant de s'approcher de ces deux représentations, et les utilisent.

2.4.1 Représentations temps-fréquence (TF) parcimonieuses

Nous avons reconsidéré l'obtention de cette forme idéale (2.37) de localisation comme étant la recherche d'une distribution parcimonieuse dans le plan temps-fréquence. En effet, on s'attend pour un signal de longueur N en temps à avoir N^2 points en temps-fréquence alors que le modèle (2.37) nécessite seulement KN points non nuls pour décrire ces K trajectoires des modes AM-FM. Les avancées récentes en reconstruction sous contrainte de parcimonie, connues sous le nom d'échantillonnage compressé (*compressed sensing*) [Don06, BCNV08, CSR] (début d'une très longue liste de références possibles) nous ont permis de regarder dans [J15, P23, P28] ce problème sous un angle jamais utilisé jusque-là : une représentation temps-fréquence parcimonieuse peut être obtenue comme résultat d'un programme d'optimisation qui favorise la parcimonie. L'information pertinente sur le signal est gardée en conservant au mieux un sous-ensemble réduit des coefficients de Fourier de la distribution de Wigner-Ville, ceux près de l'origine dans le plan des ambiguïtés, région connue pour décrire les auto-termes qui sont les $a_k^2(t)\, \delta\left(f - \dot{\varphi}_k(t)/2\pi\right)$ dans l'équation (2.37) [HF97].

Techniquement, la parcimonie est obtenue pour une distribution temps-fréquence $\rho_x(t,f)$ en imposant qu'elle ait un nombre limité de valeurs non nuls. Plutôt que de considérer le décompte du nombre de valeur, aussi appelée "norme" ℓ_0, une série de travaux (voir par

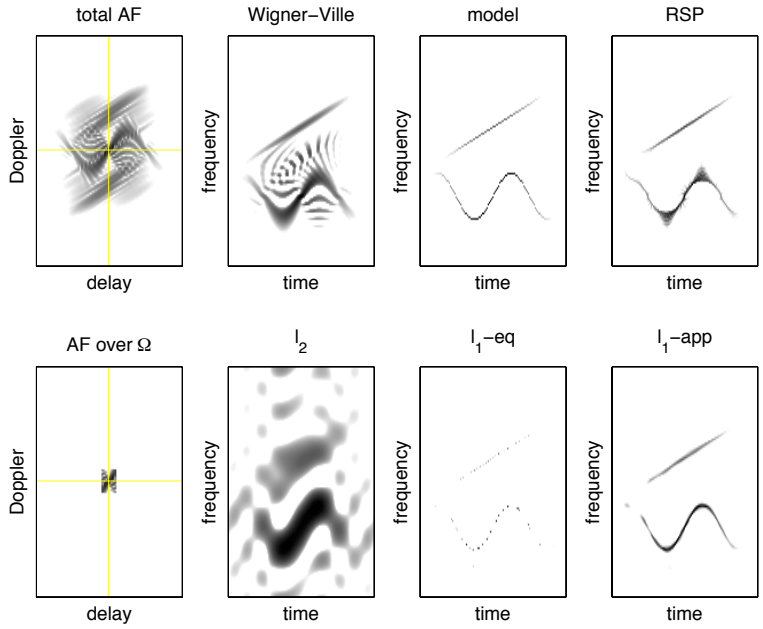


FIGURE 2.10 – Exemple de représentation temps-fréquence parcimonieuse pour un signal avec 2 composantes. Diverses distributions temps-fréquence sont données pour le signal dont la forme idéale selon (2.37) est sur la ligne du haut, entre sa distribution de Wigner-Ville et son spectrogramme réalloué. La fonction d'ambiguïté correspondante est en haut à gauche. Le domaine Ω restreint au centre proche de l'origine (de taille 13×13 ici) est en bas à gauche. On trouve ensuite en bas le résultat des estimations de ρ par optimisation en se restreignant la connaissance de la fonction d'ambiguïté sur Ω : solution si l'on minimise la norme ℓ_2, si l'on minimise la norme ℓ_1 selon (2.39) ou si l'on relaxe cela en suivant (2.40) avec $\epsilon = 0.05 \, \|x\|_2$. Toutes les amplitudes sont codées logarithmiquement en niveau de gris avec une dynamique de 18 dB.

exemple [Don06, CRT06, CT06], et [CSR] pour une liste plus complète de références sur le *compressed sensing*) a montré que des solutions presque optimales (et parfois exactes) en terme de parcimonie sont obtenues à moindre coût en minimisant une norme ℓ_1. Ce problème d'optimisation se résout par diverses techniques, par exemple des solveurs venus de la programmation linéaire [CR05], l'algorithme de Matching Pursuit [TG07], des méthodes itératives introduites initialement pour le traitement d'image [DFL08, BD08, NT09, FNW07], ou plus généralement les méthodes itératives d'optimisation convexe non-différentiable s'appuyant sur les opérateurs monotones [CP10, BACPP11, GO09].

À côté de cela, la contrainte qui permet de garder le comportement temps-fréquence du signal analysé s'exprime bien dans le domaine des ambiguïtés introduit plus haut :

$$A_x(\xi, \tau) = \int_{-\infty}^{+\infty} x\left(t + \frac{\tau}{2}\right) \, x^*\left(t - \frac{\tau}{2}\right) \, e^{i2\pi\xi t} \, dt \tag{2.38}$$

où $A_x(\xi, \tau)$ est la double transformée de Fourier de la distribution de Wigner-Ville [Fla99]. La distribution d'ambiguïté $A_x(\xi, \tau)$ s'écrit aussi comme la corrélation entre x et le signal translaté en temps et en fréquence. En tant que corrélation, son module est maximum à l'origine et les trajectoires attendues dans la forme idéale (2.37) ont alors leurs coefficients

près de l'origine (tandis que les corrélations entre 2 modes k et k' différents se retrouvent comme des termes d'interférence plus loin de l'origine). L'attache aux données que nous proposons de prendre est alors de contraindre $\rho_x(t,f)$ que l'on cherche à avoir une distribution d'ambiguïté qui soit la plus proche possible de $A_x(\xi,\tau)$, celle du signal analyse, dans un domaine Ω pas trop grand autour de l'origine. Ainsi, la construction de ρ_x vient de la résolution du problème :

$$\rho_x = \arg\min_{\rho} \|\rho\|_1 \,;\ \text{t.q.}\ \ \mathcal{F}\{\rho\} - A_x = 0|_{(\xi,\tau)\in\Omega}\,. \tag{2.39}$$

Suivant [Don06, CT06], on peut relaxer l'égalité stricte en écrivant plutôt le problème :

$$\rho_x = \arg\min_{\rho} \|\rho\|_1 \,;\ \text{t.q.}\ \ \|\mathcal{F}\{\rho\} - A_x\|_2 \leq \epsilon|_{(\xi,\tau)\in\Omega}\,, \tag{2.40}$$

où ϵ est une borne maximale à spécifier par l'utilisateur. Nous avons étudié les deux possibilités. La méthode d'optimisation employée est celle proposée par la Toolbox Matlab ℓ_1-Magic [CR05] mais on disposerait maintenant d'algorithmes plus adaptés, en particulier ceux basés sur les opérateurs monotones [CP10, BACPP11, GO09].

La figure 2.10 illustre la méthode, en montrant les distributions de Wigner-Ville et d'ambiguïté d'un signal (défini par la trace de son modèle en temps-fréquence) et le résultat de la méthode que l'on compare à la résolution du même problème inverse avec une norme ℓ_2 et à la méthode de réallocation dans le domaine temps-fréquence [AF95, CMDAF97, FACM03, ACMF12]. Remarquons qu'on aurait pu imposer aussi que $\rho \geq 0$ (pour l'interprétation en temps de densité de puissance) dans les programmes d'optimisation des équations (2.39) et (2.40) ; cependant il a été trouvé qu'en pratique les distributions obtenues sont positives. Nous avons montré que le concept fonctionne et que la représentation obtenue a des performances très bonnes même en comparaison d'une techniques de pointe pour la localisation TF telle que la réallocation dans le domaine temps-fréquence. Le prix à payer est celui d'un coût algorithmique bien plus élevé cependant. Suite à ces travaux, le même problème a été considéré par des collègues pour cette fois l'estimation parcimonieuse de spectre de signaux non stationnaires, dans un contexte stochastique donc plutôt qu'à des fins de localisation [JTH09b, JTH09a, JTH13]. Le principe de base de se servir d'un échantillonnage compressé dans le plan des ambiguïtés reste le même.

2.4.2 EMD et extraction de modes et de tendances

La décomposition modale empirique. Cette méthode, référencée comme EMD (initiales en anglais), vise à écrire un signal comme une somme de modes selon l'équation (2.36). L'idée est de chercher à décomposer le signal en modes d'oscillations non stationnaires. Ces modes sont extraits l'un après l'autre tels que, localement autour d'un instant, chaque mode est à une fréquence instantanée plus petite que celles des modes déjà extraits. L'objectif est d'obtenir une décomposition en modes que l'on nomme des fonctions de modes intrinsèques (IMF pour *Intrinsec Mode Functions*) et qui doivent vérifier deux conditions :
(1) une IMF contient le même nombre (à un près) d'extrema que de passage à zéro,
(2) les oscillations sont symétriques par rapport à zéro et ont donc une amplitude bien définie.

Mises ensemble, ces conditions disent qu'une IMF est un signal qui s'écrit bien comme un mode modulé en amplitude et en fréquence comme dans (2.36). Pour cela, l'algorithme appelé décomposition modale empirique a été proposé initialement par Huang et al. [HSL+98],

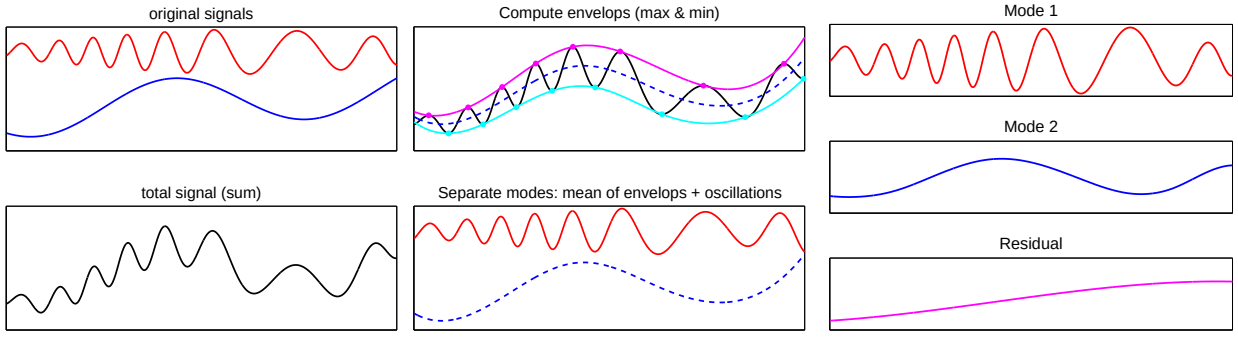


FIGURE 2.11 – Illustration de l'extraction de modes par EMD : (a) signaux de départ, (b) première étape d'extraction de la moyenne des enveloppes (tamisage), (c) résultat avec les trois modes.

qui a connu depuis de nombreuses applications (voir [HS05] et ses références). L'algorithme extrait de manière itérative chaque mode en partant par celui de plus haute fréquence. Il s'écrit comme il suit.

ALGORITHME DE L'EMD

- Considérer un signal $x(t)$ à l'entrée ; initialiation $x_1 = x$ et $k = 1$.
- Itérativement sur k, calculer $c_k(t)$ en utilisant la procédure de tamisage (*sifting*) sur ρ_m, avec $\rho_1 = x_k$ procédure décomposée ci-dessous :
  1. Identifier tous les maxima et minima locaux de $\rho_m(t)$.
  2. Calculer l'enveloppe maximale (respectivement minimale) autour de ρ_m par une interpolation entre les maxima (resp. minima) locaux (note : on utilise souvent une interpolation par des splines).
  3. Obtenir une tendance locale $Q_k(t)$ en calculant la moyenne des enveloppes maximale et minimale.
  4. Extraire les oscillations locales et mettre à jour $\rho_m = \rho_m(t) - Q_m(t)$.
  5. Si ρ_m n'est pas une IMF (c'est-à-dire qu'elle ne satisfait pas les deux conditions indiquées plus haut), ré-itérer le tamisage à partir de l'étape 1 pour ρ_{m+1}, jusqu'à obtenir une IMF valide.
     Si $\rho_m(t)$ est une IMF, poser $c_k(t) = \rho_m(t)$ et extraire le résidu $x_{k+1}(t) = x_k(t) - c_k(t)$.
- Augmenter k à $k + 1$ et revenir au début en reprenant le tamisage sur x_k.
- Fin de l'algorithme à l'itération K quand le résidu $x_{K+1}(t)$ ne contient plus assez d'oscillations. Fixer le résidu final à $r_K(t) = x_{K+1}(t)$.

Le résultat final de cet algorithme est la décomposition additive du signal x en une somme d'IMF et un résidu :

$$x(t) = \sum_{k=1}^{K} c_k(t) + r_K(t). \tag{2.41}$$

La figure 2.11 illustre les étapes principales de l'algorithme EMD, partant d'un signal simple. Il est fait par la somme de deux modes, l'un d'oscillations lentes et l'autre d'oscillations rapides. La figure centrale montre comment la construction des enveloppes lors de l'algorithme de tamisage permet d'extraire la composant lente du signal et d'estimer ainsi le mode le plus

rapide. En itérant, on obtient sur la colonne de droite la séparation de l'oscillation rapide, de l'oscillation lente et du résidu qui n'est plus qu'une tendance non constante sans oscillation. En quelque sorte, l'algorithme vise à résoudre à chaque itération en k un problème de séparation d'un signal en "tendance + fluctuations" [ABD+08], où la fluctuation est le mode rapide extrait $c_k(t)$ et la tendance est $x_{k+1}(t)$, ce qui reste.

Extraction de la tendance d'un signal. Partant de cette équivalence de l'EMD avec un problème d'extraction de tendance, nous avons montré dans [J20, J22] en quel sens l'EMD est une bonne méthode, pilotée par les données et avec peu de paramètres, pour réaliser une telle décomposition "tendance + fluctuations". Il suffit de savoir à quel mode s'arrêter dans l'algorithme EMD, mode après lequel ce qui reste a les caractéristiques d'une tendance. Pour cela, nous nous appuyons dans [J22] sur les propriétés statistiques des modes de fluctuations de signaux large bande (usuellement l'énergie des modes décroît et le nombre de passages à zéro des modes diminuent d'un facteur 2 d'un mode à l'autre) pour proposer un test détectant quels modes de l'EMD sont à mettre dans la tendance. La comparaison de cette méthode avec les extractions de tendance par optimisation ℓ_2 (filtre de Hodrick-Prescott) [HP97] ou ℓ_1 [KKBG09], montre qu'elle a de bonnes performances [J20].

Développements autour de l'EMD. Nous avons proposé dans [P49] une manière de faire du remplissage de données manquantes (*gap-filling*) qui fonctionne convenablement par interpolation individuelle de chaque mode. Dans [P61], nous nous intéressons cette fois à une méthode de décompte du nombre de modes dans une représentation temps-fréquence, à l'aide d'un critère entropique. Cette méthode est utile à l'EMD pour savoir a priori à combien de modes on s'attend, mais plus encore à d'autres méthodes de caractérisation de modes telle que les représentations temps-fréquence de la section 2.4.1, ou une autre approche d'extraction de modes telle que le *synchrosqueezing* [DM96, DLW11].

EMD comme un problème optimisation. Puisque l'EMD vient de la résolution d'un problème de séparation d'un signal en "tendance + fluctuations", nous pouvons pousser plus loin cette analogie et la reconsidérer à la lumière des travaux de séparation "texture + géométrie" en traitement d'images [AGCO06]. Ces travaux montrent qu'on peut formuler le problème à travers un critère convexe à minimiser. Faisant pareil, nous proposons dans [P52] le critère adapté et montrons comment les méthodes proximales [BACPP11, CP11] permettent de proposer un algorithme Proximal-EMD qui donne des résultats tout à fait prometteurs.

2.4.3 Utilisations de l'extraction de modes pour des réseaux de capteurs.

Signaux de capteurs environnementaux. Un domaine d'application intéressant de l'extraction de modes dans des signaux non stationnaires est celui des mesures par réseaux de capteurs. En effet, on s'attend à ce qu'à long terme, ces signaux ne restent pas stationnaires et portent la signature des cycles dans les activités humaines (journée ou semaine par exemple pour des capteurs de consommation d'énergie) ou des phénomènes naturels (alternance jour/nuit et saison pour des capteurs de température). Dans [P30, P33], nous avons regardé comment l'extraction de modes aide à caractériser les signaux obtenus de capteurs

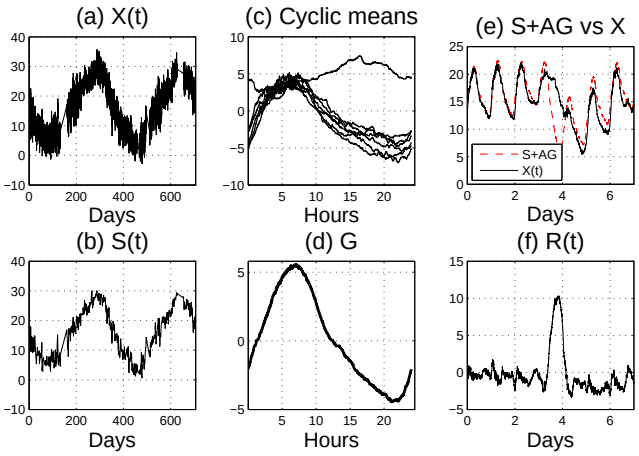


FIGURE 2.12 – Décomposition de la série des températures en modes : **a)** série initiale ; **b)** tendance sur l'année ; **c)** superposition cyclique des séries sur la journée (pour une semaine de données) ; **d)** tendance cyclique journalière estimée ; **e)** somme des deux premiers modes comparée à $X(t)$ sur la semaine ; **f)** résidu $R(t)$.

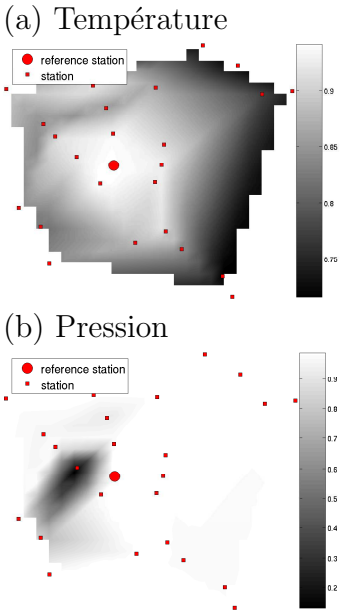


FIGURE 2.13 – Carte des corrélations dans l'espace des séries des $R(t)$ entre chaque capteur et le capteur de référence au centre.

environnementaux, en s'appuyant sur les données du **projet Live E !** de collaborateurs de l'université de Tokyo [MIO+07, liv09]. Live E ! est un système de capteurs qui collectent des données de température, de pression, de vent, etc. (et les transmettent pour stockage en temps réel sur Internet) en différents points du Japon et d'Asie. Nous avons montré qu'un simple modèle avec 3 modes convient bien :

$$X(t) = S(t) + \sum_{n=1}^{N_d} [A_J(d) \times G_J(t - nD)] + R(t) \tag{2.42}$$

représentant la tendance annuelle, $S(t)$, un cycle journalier $A_J(d) \times G_J(t)$ et des fluctuations à court terme $R(t)$. Des exemples de signaux et leurs modes sont montrés en Figure 2.12 et ils permettent d'extraire des cartes de corrélation spatiale entre les mesures des capteurs tels qu'en 2.13 pour le mode de fluctuations. L'intérêt est que les tendances et les cycles sont par définition très corrélés et, si on souhaite détecter des anomalies ou des caractéristiques plus fines, il faut être capable de considérer les fluctuation seules. Ici par exemple, on voit que les capteurs de température montrent une corrélation qui décroît simplement avec la distance tandis que les capteurs de pression dans la zone en gris de la figure 2.13 ont une corrélation anormale avec les autres (où la corrélation reste près de 1 à cette distance qui est petite pour les fluctuations de pression) : ces capteurs sont en fait défectueux ou mal calibrés et l'analyse nous le révèle.

Signaux de capteurs de consommation d'énergie. Pour étudier des données venant des capteurs de consommation d'énergie dans les bâtiments de l'Université de Tokyo [P53, P54], collectées dans le cadre du **"Green University of Tokyo Project"** [GUT], nous

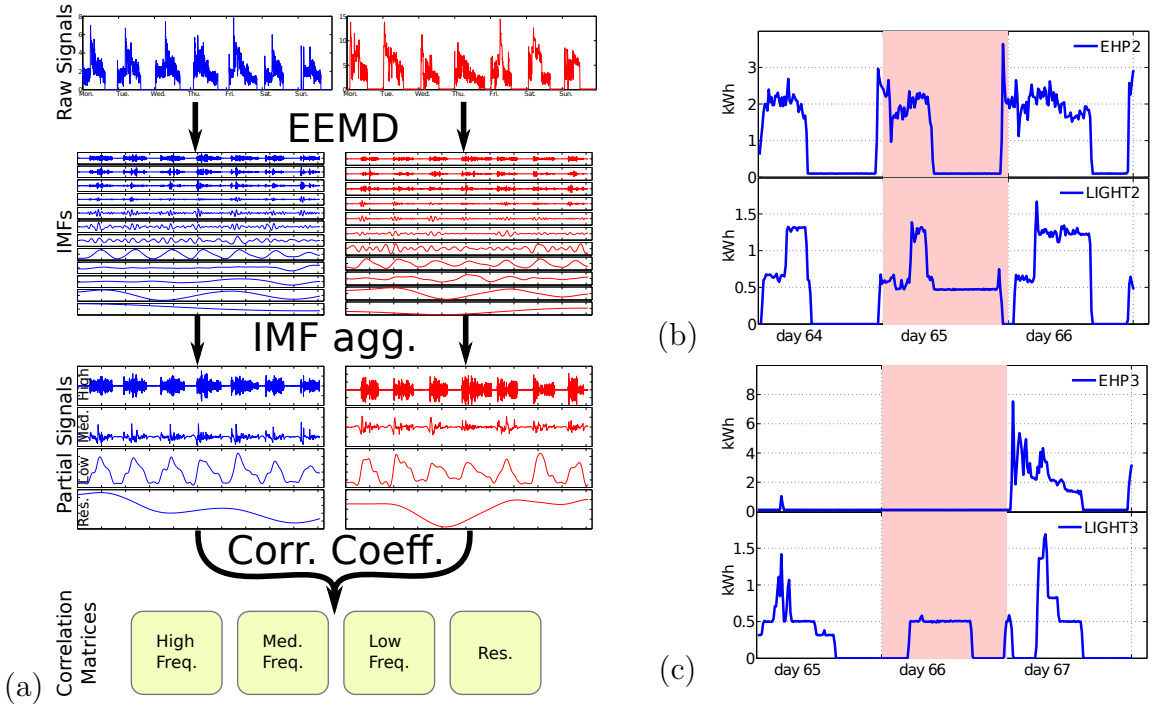


FIGURE 2.14 – Détection d'anomalies par des capteurs de consommation d'énergie. (a) Représentation schématique des étapes de la méthode pour estimer les matrices de corrélation $C_{i,j}^{B,u}$ entre toutes les paires de capteurs. À droite, exemples d'anomalies détectées : (b) Anomalie utilisant une grande puissance, avec une lumière laissée allumée toute la nuit. (c) Anomalie de faible impact en puissance (cas où le système de climatisation est inhabituellement inutilisé en journée).

avons dû nous appuyer sur l'EMD. Plusieurs milliers de capteurs mesurent en temps réel la consommation électrique de toutes les lampes et climatiseurs d'un bâtiment de l'université. Le but initial était de sensibiliser à la surconsommation électrique, et il est maintenant complété par la volonté de développer des méthodes pour automatiquement repérer les anomalies de consommation. Des méthodes comme [BMA+11, WEFW12, CC11] font l'hypothèse de stationnarité en passant par le spectre de Fourier des signaux ; d'autres emploient des outils de détection de fautes [KB05a, KB05b, See07] pas toujours simples à calibrer.

Dans cette situation, ni les signaux ni les périodes ne sont en réalité de durée prédéterminée ; on retrouve des jours, parfois avec des heures supplémentaires, des semaines mais aussi des alternances entre semaine et week-end, ou des jours fériés. Ni le simple spectre, ni un modèle a priori tel que (2.42) ne conviennent. L'adaptabilité de l'EMD et sa possibilité de varier localement en temps les fréquences des oscillations sont alors importantes. La méthodologie adoptée est tracée sur la figure 2.14 (a). Les signaux sont d'abord décomposés par EMD (en utilisant la variante de l'ensemble-EMD [WH09, TCSF11]) avant de calculer des matrices de corrélation $C_{i,j}^{B,u}$ entre toutes les paires de capteurs (i,j), autour du temps u sur une fenêtre de durée assez longue pour capturer les évolutions des signaux (ici une semaine) et pour les modes de différents domaines en fréquence B. En se limitant par exemple aux fréquences hautes (périodes moins de 20 minutes), moyennes (de 20 minutes à 6 heures), basses (plus de 6 heures) et très lentes (les résidus des EMD), on arrive déjà a détecter des anomalies. Pour cela, une référence est estimée en prenant la médiane des

matrices de corrélations :

$$R^B_{i,j} = \text{median}_u(C^{B,u}_{i,j}) = \text{median}(C^{B,1}_{i,j}, ..., C^{B,n}_{i,j}). \tag{2.43}$$

Ensuite les déviations locales sont calculées par comparaison entre le local et le global (ici caractérisé par cette médiane mais cela nous rappelle le cadre proposé pour les tests de stationnarité) :

$$l^{B,t}_i = \left(\sum_{j=1}^{d} w_{ij} |C^{B,t}_{i,j} - R^B_{i,j}|^p \right)^{1/p}. \tag{2.44}$$

Les poids $w_{ij} = R^B_{i,j} / \sum_{k=1}^{d} R^B_{i,k}$ servent à donner plus d'importance aux événements qui sont usuellement corrélés (de la même façon qu'on pondérait les spectres dans l'extension aux non-stationnarités de premier ordre en 2.3.1). La détection des anomalies est possible en regardant la Déviation Moyenne absolue (MAD) [HR09], définie comme : $\text{MAD}_i = b\,\text{median}(|l_i - \text{median}(l_i)|)$. La constante b est en général fixée à 1.4826 pour se conformer au paramètre usuel si l'on test des données gaussiennes (ce qui n'est pas le cas ici). Un capteur i révèle une anomalie au temps t si $l^{B,t}_i > \text{median}_t(l^B_i) + \tau\,\text{MAD}(l_i)$ où τ ajuste la sensibilité du détecteur. Empiriquement, $p = 4$ conduit à de bons résultats, retenant les anomalies importantes sur $C^{B,t}_{i,j}$ sans lever des alarmes en permanence. Deux exemples d'anomalies détectées ainsi sont montrées en figure 2.14 (b) et (c). Sur des données couvrant 10 semaines (du 27/06/2011 au 5/09/2011), pour 135 capteurs de consommation des lumières et systèmes de chauffage, aération et climatisation, dans un bâtiment de 12 étages, nous avons repéré ainsi 9 alarmes importantes comme la (b), causant des pertes d'énergie injustifiées (la plus grande pour 165 kWh) et ce dans un contexte post-Fukushima au Japon où la consommation d'énergie était particulièrement surveillée.

Avec très peu d'a priori sur les signaux, la méthode combinant EMD et détection d'anomalie s'est révélée être intéressante pour détecter, sans supervision, des anomalies.

2.4.4 Extraction de tendance pour les LES en physique

La simulation aux grandes échelles pour les fluides. Un problème de simulation en mécanique des fluides nous a amené à élargir nos approches sur les extractions de tendances non stationnaires. En collaboration avec E. Lévêque (Laboratoire de physique), J. Boudet (LMFA, EC Lyon) et le doctorant A. Cahuzac, nous avons étudié une nouvelle de méthode de simulation numérique d'écoulement de type LES (*Large Eddy Simulation*) qui s'appuie sur le modèle de *Shear-Improved Smagorinsky* [LTSB07].

La simulation des grandes échelles en mécanique des fluides [LMC05] permet de calculer l'évolution d'un fluide en écoulement turbulent, au prix de n'avoir qu'une version filtrée en espace. Pour ce faire, les équations de la mécanique des fluides sont réécrites pour un champ de vitesse $\bar{\mathbf{u}}(\mathbf{x})$ moyenné à l'échelle de la grille autour de la position $\mathbf{x}$. Les équations à résoudre sont celles pour le du champ de vitesse $\bar{\mathbf{u}}(\mathbf{x})$ moyenné à l'échelle de la grille autour de la position $\mathbf{x}$, de la densité du fluide ρ et de son énergie par masse ρe_t. En introduisant

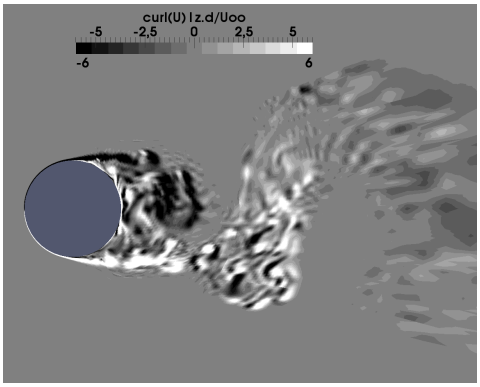


FIGURE 2.15 – LES : visualisation d'une coupe 2D de la vorticité verticale (parallèle au cylindre) instantanée ; on voit les tourbillons qui se développent derrière le cylindre (disque gris).

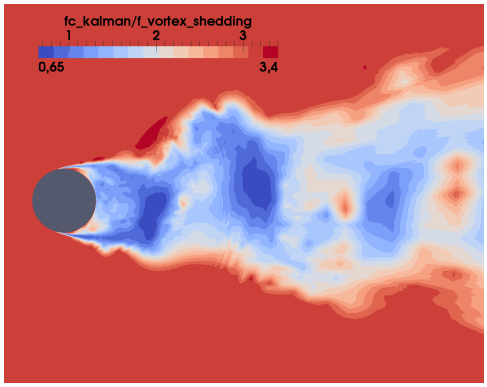


FIGURE 2.16 – Carte instantanée de la fréquence de coupure adaptative estimée par le filtre de Kalman (normalisée par la fréquence d'émission des tourbillons).

la notation de l'opérateur de Favre $\widetilde{q} = \overline{\rho q}/\overline{\rho}$, les équations filtrées sont :

$$\frac{\partial \overline{\rho}}{\partial t} + \frac{\partial \left(\overline{\rho}\widetilde{u_j}\right)}{\partial x_j} = 0 \quad ; \quad \frac{\partial \left(\overline{\rho}\widetilde{u_i}\right)}{\partial t} + \frac{\partial \left(\overline{\rho}\widetilde{u_i}\widetilde{u_j}\right)}{\partial x_j} = -\frac{\partial \overline{P}}{\partial x_i} + \frac{\partial \overline{\tau}_{ij}}{\partial x_j} + \frac{\partial \Pi_{ij}^{\text{sgs}}}{\partial x_j} \tag{2.45}$$

$$\text{et} \quad \frac{\partial \left(\overline{\rho}\widetilde{e_t}\right)}{\partial t} + \frac{\partial \left(\left(\overline{\rho}\widetilde{e_t} + \overline{P}\right)\widetilde{u_j}\right)}{\partial x_j} = \frac{\partial \left(\widetilde{u_i}\left(\overline{\tau}_{ij} + \Pi_{ij}^{\text{sgs}}\right)\right)}{\partial x_j} + \frac{\partial}{\partial x_j}\left(\overline{\lambda}\frac{\partial \widetilde{T}}{\partial x_j}\right) - \frac{\partial \Theta_j^{\text{sgs}}}{\partial x_j} \tag{2.46}$$

Pour la LES, on modélise les termes de pression $\overline{P}$ (par exemple : gaz parfait), de chaleur sous-grille Θ^{sgs} (par exemple : proportionnel au gradient de température) et de tenseur de déformation sous-grille Π_{ij}^{sgs}.

Un enjeu des LES est de modéliser le comportement des fluctuations turbulentes à plus petite échelle que la grille, en particulier leur contribution μ_{SGS} à la viscosité effective, pour retrouver correctement la vitesse à grande échelle. Il a été proposé dans [LTSB07] qu'une viscosité de Smagorinsky ajustée par l'étirement (SISM pour *Shear-improved Smagorinsky's Model*) est un bon modèle pour cela :

$$\mu_{SGS} = \bar{\rho}(C_s\Delta)^2\left(|\bar{S}| - |\langle\bar{S}\rangle|\right), \tag{2.47}$$

où $C_s = 0.18$ est la constante de Smagorinsky, Δ l'espacement de la grille et $\bar{S} = (\partial\bar{\mathbf{u}}_i/\partial\mathbf{x}_j + \partial\bar{\mathbf{u}}_j/\partial\mathbf{x}_i)/2$, le tenseur d'étirement résolu en temps et en espace. La notation $\langle\cdot\rangle$ indique la *moyenne d'ensemble* sur l'écoulement. Le problème est alors d'estimer cette moyenne en toute situation, même sans direction d'homogénéité (pour prendre des moyennes spatiales), ni stationnarité de l'écoulement (qui permettrait des moyennes dans le temps), et ceci que l'écoulement soit laminaire ou turbulent.

Extraction de tendance pour les LES. Notre travail dans [J18, P32, P47, P48] a été d'aborder la question d'extraction du flot moyen comme un problème d'estimation de la moyenne temporelle de l'écoulement, point à point sur la grille. L'extraction de ce flot moyen résolu en temps et en espace a été d'abord considéré par une méthode de lissage exponentiel dans le temps. Du fait des contraintes d'une simulation numérique où l'on ne

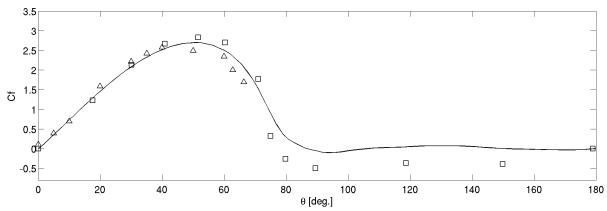


FIGURE 2.17 – Coefficient moyen de friction autour du cylindre, fonction de l'angle (direction d'arrivée du flot à $\theta = 0°$). — : LES à $\mathrm{Re} = 4.7 \times 10^4$; △ : points experimentaux à $\mathrm{Re} = 9.1 \times 10^4$ [YR90] ; □ : points experimentaux à $\mathrm{Re} = 10^5$ [Ach68].

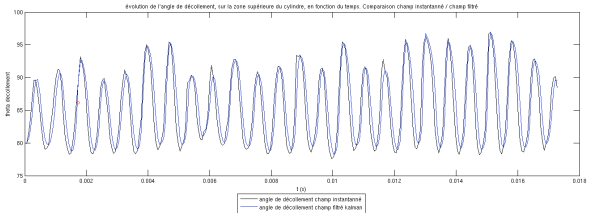


FIGURE 2.18 – Angle de décollement de la couche limite sur le côté haut du cylindre en fonction du temps, pour le l'écoulement complet et l'écoulement moyen extrait. On note un retard de ce dernier qui est de 15% environ.

retient pas plus que le pas de temps précédent pour intégrer l'équation d'évolution de $\bar{\mathbf{u}}$, et pour garder une méthode locale (aisément parallélisable), il a été proposé d'estimer la vitesse moyenne locale $\langle \bar{u} \rangle$ au temps $n+1$ (et à une position $\mathbf{x}$) par :

$$\langle \bar{\mathbf{u}} \rangle_{n+1}(\mathbf{x}) = (1-\alpha) \cdot \langle \bar{\mathbf{u}} \rangle_n(\mathbf{x}) + \alpha \cdot \bar{\mathbf{u}}_n(\mathbf{x}). \tag{2.48}$$

Le paramètre de mémoire constant α donne la fréquence de coupure f_c du filtre passe-bas correspondant par l'équation :

$$\alpha = 2\pi f_c \Delta t / \sqrt{3}, \tag{2.49}$$

où Δt est le pas de temps de la discrétisation numérique [J18]. La limite de cette approche est que α, et par conséquent la fréquence de coupure, sont fixés une fois pour toute. Pour un écoulement qui peut avoir des zones laminaires (peu de fluctuations) ou turbulentes (avec des fluctuations importantes), et des transitions entre les deux, ce filtrage conduit à des retards aux transitions si le paramètre de mémoire est trop petit, ou à trop suivre les fluctuations instantanées si α est trop grand.

Filtre de Kalman adaptatif pour l'extraction du flot moyen. Pour remédier à ces difficultés, nous introduisons une manière de rendre adaptatif dans le temps (et l'espace) le lissage à opérer sur l'écoulement pour en extraire le flot moyen. Pour cela, nous remplaçons la moyenne exponentielle de l'eq. (2.48) par un filtre de Kalman adaptatif. On sait (voir par exemple [Har89]) que dans son régime stationnaire, le filtre de Kalman appliqué à un modèle d'observation bruitée d'une moyenne locale qui évolue en marche aléatoire, est équivalent à un lissage exponentiel. La forme état-espace de ce modèle très simple est :

$$\langle \bar{\mathbf{u}} \rangle_n = \langle \bar{\mathbf{u}} \rangle_{n-1} + \eta_n \quad \text{et} \quad \bar{\mathbf{u}}_n = \langle \bar{\mathbf{u}} \rangle_n + \epsilon_n \tag{2.50}$$

où ces équations sont à prendre indépendamment en chaque point $\mathbf{x}$ et où ϵ et η sont respectivement les bruits d'observation et d'évolution de la moyenne (supposés indépendants). La structure classique de filtre de Kalman se réécrit dans ce cas sous la forme du lissage exponentiel (2.48) avec cette fois α_n qui est le gain de Kalman et évolue dans le temps comme il suit :

$$\alpha_n = P_n / \left(P_n + \sigma^2(\epsilon)_n \right) \tag{2.51}$$

Ensuite l'étape de correction et de prédiction de la variance de l'erreur P s'écrit :

$$P_{n+1} = (1 - \alpha_n) \cdot P_n + \sigma^2(\eta)_n, \tag{2.52}$$

où P_n est la prédiction de la variance de l'erreur de la composante moyennée, et σ^2 sont les variances de ϵ et η (noter que, par rapport à l'écriture classique du filtre de Kalman, sont condensés ici l'étape de correction et celle de prédiction pour la variance de l'erreur). Dans un état stationnaire, ce filtre de Kalman conduit à un paramètre de mémoire [3] α_n égal à $\sigma(\eta)/\sigma(\epsilon)$. Supposant que l'écoulement moyen évolue en fonction des temps caractéristiques de l'écoulement, on fixe $\sigma(\eta) = 2\pi f_c \Delta t \cdot u_c/\sqrt{3}$ indépendant de l'instant, où f_c and u_c sont des références fixes en fréquence et vitesse données par les paramètres généraux de l'écoulement. L'équation d'observation a un sens un peu particulier ici : la fluctuation ϵ_n rend compte du caractère turbulent ou non de l'écoulement. En se contraignant à des estimations locales en espace et à un pas de temps, nous ne disposons que des estimateurs grossiers de la variance. Nous proposons de prendre :

$$\widehat{\sigma^2(\epsilon)}_n = \max\left(u_c \cdot |\langle \bar{\mathbf{u}} \rangle_n - \bar{\mathbf{u}}_n|, 0.1 \cdot {u_c}^2\right) \tag{2.53}$$

comme estimateur représentatif des fluctuations turbulentes. Il est dimensionnellement correct et impose une borne minimum aux fluctuations supposées (ce qui est nécessaire pour que le filtre puisse toujours réagir). Le rapport typique entre $\sigma(\eta)$ et $\widehat{\sigma(\epsilon)}$ est alors de l'ordre de $2\pi f_c \Delta t/\sqrt{3}$, sauf fluctuations importantes. Cela revient à supposer un paramètre de mémoire typique qui est cohérent avec celui que l'on donnait pour la moyenne exponentielle.

Mise en œuvre et résultats. Cette méthode d'extraction du flot moyen a été comparée au lissage exponentiel sur deux cas, l'écoulement dans un canal plan et l'écoulement autour d'un cylindre dans [J18]. Pour un cylindre, dans un régime turbulent sous-critique ($Re_D = 4.7 \cdot 10^4$), l'écoulement présente des zones turbulentes et laminaires, un lâcher tourbillonnaire et des couches de cisaillement instationnaires. La simulation est réalisée avec Turb'flow, un solveur standard de LES (plus de détails sont donnés dans [BCSL07]) sur un maillage en accord avec les pratiques usuelles en LES [Sag06]. Les valeurs de références sont prises égales à $f_c = \mathrm{St} \cdot U_\infty/D = 1400$ Hz (avec un nombre de Strouhal $\mathrm{St} = 0.2$ associé au lâcher tourbillonnaire) et une vitesse de fluide en amont $u_c = U_\infty = 70$ m.s^{-1}. Le diamètre du cylindre est $D = 10$ mm et le fluide est de l'air. Il s'agit d'une configuration standard en aérodynamique.

En illustration de l'écoulement obtenu, la Fig. 2.15 montre la structure de vorticité résultante ; on y voit une couche limite laminaire qui décolle en haut et en bas du cylindre puis transitionne pour conduire à un lâcher tourbillonnaire. La Fig. 2.16 est un résultat physique donné par le filtre de Kalman : elle montre l'estimation de la fréquence de coupure instantanée liée au gain de Kalman instantané α_n par l'équation (2.49), en chaque point. Cette information est physiquement pertinente pour l'écoulement : on y reconnaît le sillage turbulent où la fréquence de coupure est de l'ordre de la fréquence des tourbillons lâchés, et la zone hors du sillage où la fréquence de coupure est haute (environ 3 f_c). Dans cette dernière zone, il y a très peu de turbulence et toute évolution est principalement celle du flot moyen. L'adaptation du filtre se comporte donc correctement.

3. dans le cas physique pertinent où $\alpha_n \ll 1$

En Fig. 2.17, on compare le profil du coefficient moyen de friction autour du cylindre à des résultats expérimentaux. L'accord est bon. On trouve en particulier que l'angle de décollement de la couche limite, où le frottement s'annule, est en moyenne à $\theta_s = 86°$ dans les simulations, proche la valeur de $\theta_s \approx 83°$ données dans la littérature [Zdr02] pour les nombres de Reynolds $4.0\times10^4 \leq \mathrm{Re}_D \leq 4.5\times10^4$. D'autres grandeurs physiques sont comparées dans [J18, Js26] à des valeurs expérimentales ou numériques de la littérature. Cette méthode est maintenant implémentée dans le code numérique de développement industriel (Turb'Flow) pour des calculs réalistes d'éléments de turbine.

Perspectives. Bien que donnant de bons résultats, l'extraction du flot moyen avec ce filtre de Kalman ne résout cependant pas tous les problèmes du lissage exponentiel : il reste un retard à l'évolution de la moyenne extraite. En Fig. 2.18, est représenté l'angle de décollement instantanée en haut du cylindre, en fonction du temps, pour le flot complet et pour le flot moyen extrait. Le flot moyen extrait est toujours en retard sur le flot total, en particulier à ce point mais pas seulement. Le retard ici est de 15% environ par rapport à la période du lâcher tourbillonnaire. Ceci induit une erreur sur le calcul de la viscosité sous-maille instantanée, et donc sur le champ de vitesse prédit.

Avec le simple modèle de l'éq. (2.50), il n'est pas possible de contrôler ce retard sans changer la réponse aux fluctuations. Il faut pour cela augmenter l'ordre du modèle et du filtre correspondant, en passant par exemple au modèle à pente locale suivant :

$$\left\{\begin{array}{rcl} \langle\bar{\mathbf{u}}\rangle_n &=& \langle\bar{\mathbf{u}}\rangle_{n-1} + \bar{\mathbf{b}}_{n-1} + \eta_n \\ \bar{\mathbf{b}}_n &=& \bar{\mathbf{b}}_{n-1} \;+\; \zeta_n \end{array}\right. \quad \text{et} \quad \bar{\mathbf{u}}_n = \langle\bar{\mathbf{u}}\rangle_n + \epsilon_n \tag{2.54}$$

La structure du filtre de Kalman pour la prédiction de ce modèle est classiquement connue. Posant $\mathbf{A} = [[1\,1];[0\,1]]$ la matrice de transition du vecteur $[\langle\bar{\mathbf{u}}\rangle_n, \bar{\mathbf{b}}_n]^{\mathrm{t}}$, et $\mathbf{H} = [1\;0]$ le vecteur d'observation de la moyenne, le filtre de Kalman pour ce modèle à pente locale s'écrit comme il suit (en chaque point, pour chaque composante de la vitesse) :

$$\begin{bmatrix}\langle\bar{\mathbf{u}}\rangle_{n+1}\\ \bar{\mathbf{b}}_{n+1}\end{bmatrix} = \mathbf{A}\begin{bmatrix}\langle\bar{\mathbf{u}}\rangle_{n}\\ \bar{\mathbf{b}}_{n}\end{bmatrix} + \mathbf{K}_n\left(u_n - \langle\bar{\mathbf{u}}\rangle_n - \bar{\mathbf{b}}_n\right) \tag{2.55}$$

avec

$$\mathbf{K}_n = \frac{\left[p_n^{11},\; p_n^{12}\right]^{\mathrm{t}}}{p_n^{11} + \sigma^2(\epsilon)_n} \text{ et } \mathbf{P}_{n+1} = (1 - \mathbf{K}_n\mathbf{H})(\mathbf{A}\mathbf{P}_n\mathbf{A}^{\mathrm{t}} + \mathbf{Q}_n). \tag{2.56}$$

$\mathbf{K}_n$ est le gain de Kalman, $\mathbf{P}_n = [p_n^{ij}]$ la matrice de covariance de prédiction de l'erreur ; $\sigma^2(\epsilon)_n$ est la variance des fluctuations turbulentes dont un estimateur $\widehat{\sigma^2(\epsilon)}_n$ a déjà été proposé en équation (2.53). La matrice de variance de l'état est $\mathbf{Q}_n = \sigma^2(\eta)[[1\;0];[0\;s_n]]$ où $\sigma^2(\eta)$ est maintenu constant comme précemment et s_n permet de régler la réactivité de changement du niveau moyen. Il reste à proposer une estimation de s_n et donc de $\sigma^2(\zeta)$, la variance supposée de $\bar{\mathbf{b}}_n$. Le premier objectif étant d'avoir un filtre plus réactif lors des transitions instationnaires, on fait dépendre l'estimateur des fluctuations turbulentes observées, donc proportionnel à $\widehat{\sigma^2(\epsilon)}_n$. Reste à fixer une éventuelle constante dans l'estimateur $\widehat{s}_n$.

Pour cela, on calcule l'état stationnaire de ce filtre qui est un filtre de Holt-Winters avec pente locale, dont le double lissage exponentiel est un cas particulier :

$$\left\{\begin{array}{rcl} \langle\bar{\mathbf{u}}\rangle_n &=& \bar{\mathbf{b}}_{n-1} + (1-\lambda_0)\langle\bar{\mathbf{u}}\rangle_{n-1} + \lambda_0\mathbf{u}_n \\ \bar{\mathbf{b}}_n &=& (1-\lambda_0\lambda_1)\bar{\mathbf{b}}_{n-1} + \lambda_0\lambda_1(\mathbf{u}_n - \langle\bar{\mathbf{u}}\rangle_{n-1}) \end{array}\right. \tag{2.57}$$

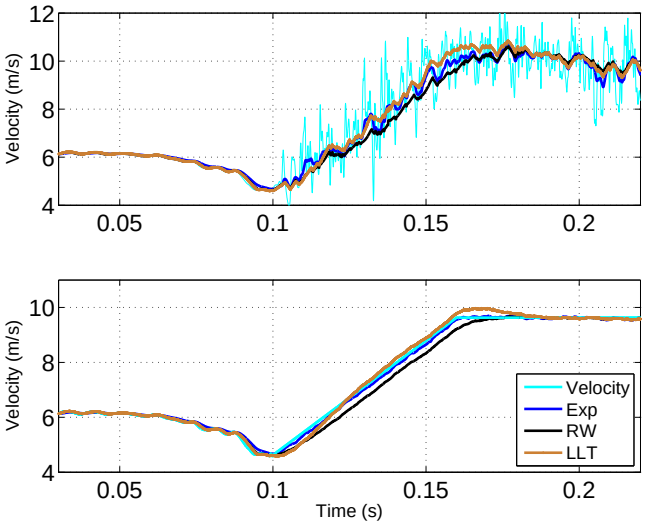


FIGURE 2.19 – Comparaison entre vitesse (fluctuante) réelle (cyan) et vitesses moyennes extraites par lissage exponentiel (`Exp`, en bleu), Kalman pour marche aléatoire (`RW`, en noir) ou pour modèle avec pente locale (`LLT`, en brun). En haut : sur une réalisation. En bas : sur une moyenne de 50 réalisations. Les données viennent de la LES dans un canal de [J18].

Suivant le calcul de [Har89], les relations entre les coefficients (fixes) de ce filtre et les variances (variant dans le temps et s'adaptant si la situation n'est pas stationnaire) du filtre de Kalman sont $\sigma^2(\zeta) = \sigma^2(\epsilon)\lambda_0^2\lambda_1^2/(1-\lambda_0)$ et $\sigma^2(\eta) = \sigma^2(\epsilon)(\lambda_0^2(1+\lambda_1) - 2\lambda_0\lambda_1)/(1-\lambda_0)$. Le régime du filtre est d'avoir des variances des états petites par rapport à celles des fluctuations. À l'ordre le plus bas il reste $\lambda_0 \approx \sigma(\eta)/\sigma(\epsilon)$ et $\lambda_1 \approx \sigma(\zeta)/(\sigma(\epsilon)\lambda_0) \approx \sqrt{s_n}$ compte tenu des estimations proposées. Pour finir de fixer λ_1 (et donc $\widehat{s}_n$), on rappelle que pour un double lissage exponentiel, le coefficient λ_1 vérifie $\lambda_1 \approx \lambda_0/4 \approx \alpha/4$ (si petits devant 1). La deuxième règle proposée est de retrouver ce comportement en état stationnaire et cela impose plus précisément comme estimateur : $\widehat{s}_n = \alpha^2 \ \widehat{\sigma^2(\epsilon)}_n/(16\ u_c^2)$, où α est le lissage moyen prescrit de l'eq. (2.49) et u_c la référence de vitesse des paramètres globaux de l'écoulement.

En Fig. 2.19 on montre comment se comportent alors, lors d'une transition temporelle d'écoulement laminaire à turbulent, en un point, les trois méthodes considérées d'extraction de flot moyen. L'intérêt du filtre avec pente locale est que l'écoulement moyen dans la partie laminaire (avant 0.1 s) est estimé égal à l'écoulement instantané (alors que le lissage exponentiel calibré sur f_c lisse les oscillations), tandis que lors de la transition vers la turbulence (entre 0.1 s et 0.16 s), la vitesse moyenne extraite augmente plus rapidement qu'avec filtre de Kalman du modèle de l'eq. (2.52). Visuellement, le modèle de l'éq. (2.56) semble améliorer la réactivité du filtre, sans perdre la possibilité de suivre les évolutions en régime laminaire. Pour quantifier cela, on mesure une moyenne d'ensemble sur plusieurs réalisations, montrée en Fig. 2.19, en bas. Le modèle sans pente est en moyenne en retard de 0.06s par rapport au comportement réel lors de la transition, le modèle avec pente est décalé de 0.01s seulement (comme le lissage exponentiel). Ces tests ont pour l'instant été réalisés en traitement a posteriori et la perspective est d'intégrer ce modèle dans le calcul LES afin de voir si il corrige effectivement le déphasage entre la vitesse et sa moyenne, et ainsi l'erreur sur l'angle de décollement sur le cylindre.

2.5 Perspectives : des approches non stationnaires en action

Le fil conducteur des travaux en traitement non stationnaire des signaux a été de proposer des approches pragmatiques à l'aide de l'analyse temps-fréquence ou en modes non stationnaires. Nous avons ainsi proposé une méthode efficace et versatile de synthèse de réalisation de processus aléatoire à l'aide des signaux substituts, des tests opérationnels de la stationnarité relative à l'échelle de temps d'observation qui autorisent plusieurs variantes, et des méthodes pour extraire des modes non stationnaires qui ont été adaptées en fonction du contexte et des propriétés des signaux regardés. L'usage d'algorithmes pilotés par les données (comme l'EMD ou les substituts) est aussi un point saillant de ce travail.

Les travaux sur les tests de stationnarité ont à mon sens atteint une bonne maturité et permettent de disposer d'un cadre et d'une batterie de variantes pour estimer si un signal est stationnaire ou non, trouver ses échelles typiques d'évolution non stationnaire le cas échéant et même de caractériser plus précisément une classe non stationnarité.

Concernant les représentations en modes non stationnaires, je continue à contribuer dans ce domaine, en mettant plus particulièrement avant quelques approches innovantes et leur utilisation sur des applications concrètes :

- Nous avons proposé une nouvelle méthode d'EMD qui s'appuie sur les méthodes d'optimisation pour extraire les modes [P52], dans un esprit d'extraction "tendance + fluctuations" proche de ce qui est fait en "géométrie + texture". Nous continuons en particulier en abordant l'EMD en deux dimensions, pour les images, qui n'a pas encore de solution idéale, via ce cadre variationnel d'extraction des modes. Ceci sera en partie fait dans le projet ANR ASTRES (Analyse, Synthèse et Transformations par Réallocation, EMD et Synchrosqueezing) qui vient de débuter.
- Les méthodes d'extraction de tendances sont utiles dans diverses applications. Je regarde actuellement comment les utiliser pour le suivi, lors de l'accouchement, de l'état du fœtus par ECG. Ceci est l'objet du projet ANR blanche FETUSES portant sur la "Non Stationary and Multifractal Statistical Analyses of Per Partum Fetal Heart Rate for Asphyxia Diagnosis", en cours, et j'ai déjà participé au travail de [P56]. Ma contribution est de s'interroger sur comment on pourrait améliorer l'extraction de la tendance tout en détectant les zones de décélérations et d'accélérations du rythme cardiaque, importantes pour diagnostiquer l'état de santé du fœtus.
- Avec l'équipe « Matière molle et cristaux liquides », je travaille sur l'analyse de signaux de diffusion de lumière dans des cellules vivantes (expériences d'E. Freyssingeas, MC ENSL) [SPGF08, SPGF09]. Ces expériences permettent de mesurer l'activité des noyaux cellulaires et on obtient des signaux non stationnaires. Il faut extraire les caractéristiques pertinentes concernant l'évolution des régions chromatiques. Une doctorante dirigée par E. Freyssingeas travaillant depuis un an et demi sur ce sujet, nous avons de nouvelles données expérimentales sur lesquelles nous pouvons employons les outils non stationnaires (et multi-échelles) que nous adaptons à ces signaux. L'objectif sera clairement d'identifier les différents temps caractéristiques d'évolution non stationnaire de la dynamique interne des noyaux de cellules in vivo et cela passera par du traitement du signal non stationnaire.

Les perspectives de nouveaux développements méthodologiques ainsi que de nouvelles applications des outils déjà développés, sont donc riches et nombreuses.

Travaux liés au chapitre 2

Journaux à comité de lecture

[Js29] N. Pustelnik, P. Flandrin, P. Borgnat, "A Multicomponent proximal algorithm for Empirical Mode Decomposition" submitted, 07/2013.

[Js26] A. Cahuzac, E. Lévêque, P. Borgnat, J. Boudet, M.C. Jacob, "A Kalman Filter adapted to mean gradient extraction in Large-Eddy Simulation of Unsteady Turbulent Flows", submitted, 2013.

[**J22**] A. Moghtaderi, P. Flandrin, P. Borgnat, "Trend Filtering via Empirical Mode Decompositions", *Computational Statistics & Data Analysis*, Vol. 58, p. 114-126, February 2013.

[**J21**] P. Flandrin, C. Richard, P.-O. Amblard, P. Borgnat, P. Honeine, H. Amoud, A. Ferrari, J. Xiao, « Stationnarité Relative et Approches Connexes », *Traitement du signal* (numéro spécial ANR) Vol. 28, No. 6, pp. 691-716, 2012.

[**J20**] A. Moghtaderi, P. Borgnat, P. Flandrin, "Trend Filtering : Empirical Mode Decompositions vs. ℓ_1 and Hodrick-Prescott" *Advances in Adaptive Data Analysis*, Vol. 3, No. 1-2, p. 41-61, 2011.

[**J18**] A. Cahuzac, J. Boudet, P. Borgnat, E. Lévêque, "Smoothing Algorithms for Mean-Flow Extraction in Large-Eddy Simulation of Complex Turbulent Flows", *Physics of Fluids*, Vol. 22, P. 12510408, 14 December 2010.

[**J16**] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, J. Xiao, "Testing stationarity with surrogates : A time-frequency approach", *IEEE Trans. on Signal Processing*, Vol. 58 :7, p 3459-3470, July 2010.

[**J15**] P. Flandrin, P. Borgnat, "Time-Frequency Energy Distributions Meet Compressed Sensing", *IEEE Trans. on Signal Processing*, Vol. 58 :6, p 2974-2982, June 2010.

[**J14**] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, P. Borgnat, "Multitaper estimation of frequency-warped cepstra, with application to speaker verification", *IEEE Signal Processing Letters*, Vol. 17, N° 4, p. 343 - 346, April 2010.

[**J11**] P. Borgnat, P. Flandrin, "Stationarization via surrogates," *Journal of Statistical Mechanics : Theory and Experiment : Special issue UPoN 2008*, P01001, January 2009.

[**J10**] J. Xiao, P. Borgnat, P. Flandrin, « Sur un test temps-fréquence de stationnarité », *Traitement du Signal*, Vol. 25, N° 4, p. 357-366, 2008

[**J9**] P. Flandrin, P. Borgnat, "Revisiting and testing stationarity," *J. Phys : Conf. Series "2008 Euro American Workshop on Information Optics WIO'08"*, Vol. 139, p. 012004, IOP Publishing, Annecy, June 2008.

Actes publiés dans des colloques avec actes à comité de lecture

[**P61**] N. Saulig, N. Pustelnik, P. Borgnat, P. Flandrin, V. Sucic, "Instantaneous counting of components in nonstationary signals", *21th European Signal Processing Conf. EUSIPCO-13*, Bucharest (RO), September 2013.

[**P56**] P. Abry, S. Roux, V. Chudacek, P. Borgnat, P. Gonçalves, M. Doret, "Hurst Ex-

ponent and IntraPartum Fetal Heart Rate : Impact of Decelerations", *IEEE CBMS 2013 (International Symposium on Computer-Based Medical Systems)*, Porto (Portugal), 22-24 June 2013. Best Paper Award of the conference.

[P55] R. Fontugne, N. Tremblay, P. Borgnat, P. Flandrin, H. Esaki, "Mining Anomalous Electricity Consumption Using Ensemble Empirical Mode Decomposition", *IEEE Int. Conf. on Acoust., Speech and Signal Proc., ICASSP-2013*, Vancouver (Canada), May 2013.

[P54] R. Fontugne, J. Ortiz, N. Tremblay, P. Borgnat, P. Flandrin, K. Fukuda, D. Culler, H. Esaki, "Strip, Bind, and Search : A Method for Identifying Abnormal Energy Consumption in Buildings", *IEEE/ACM Information Processing in Sensor Networks, ISPN'13*, Philadelphia (PN, USA), 8-11 April 2013

[P52] N. Pustelnik, P. Borgnat, P. Flandrin, "A Multicomponent Proximal Algorithm for Empirical Mode Decomposition", *20th European Signal Processing Conf. EUSIPCO-12*, Bucharest (RO), August 2012.

[P51] P. Borgnat, P. Abry, P. Flandrin, "Using Surrogates and Optimal Transport for Synthesis of Stationary Multivariate Series with Prescribed Covariance Function and non-Gaussian Joint-Distribution" *IEEE Int. Conf. on Acoust., Speech and Signal Proc., ICASSP-2012*, Kyoto (Japon), March 2012.

[P50] D. Baptista de Souza, J. Chanussot, A.-C. Favre, P. Borgnat, "A Modified Time-Frequency Method for Testing Wide-Sense Stationarity", *IEEE Int. Conf. on Acoust., Speech and Signal Proc., ICASSP-2012*, Kyoto (Japon), Mars 2012.

[P49] A. Moghtaderi, P. Borgnat, P. Flandrin, "Gap-Filling by the Empirical Mode Decomposition" *IEEE Int. Conf. on Acoust., Speech and Signal Proc., ICASSP-2012*, Kyoto (Japon), Mars 2012.

[P48] A. Cahuzac, J. Boudet, P. Borgnat, E. Lévêque,, "Dynamic Kalman filtering to separate low-frequency instabilities from turbulent fluctuations : Application to the Large-Eddy Simulation of unsteady turbulent flows", *Colloque ETC-13*, Warsaw (Poland), 12-15 septembre 2011.

[P47] A. Cahuzac, J. Boudet, E. Lévêque, P. Borgnat,, "Extraction de flot moyen dans des simulations numériques à grande échelle de fluides par filtre de Kalman adaptatif", *23e Colloque sur le Traitement du Signal et des Images. GRETSI-2011*, id. 191, Bordeaux (France), 5-8 septembre 2011.

[P45] P. Borgnat, P. Flandrin, A. Ferrari, C. Richard, "Transitional Surrogates", *IEEE Int. Conf. on Acoust., Speech and Signal Proc. ICASSP-11*, Prague (CZ), 23-27 mai 2011.

[P42] C. Richard, A. Ferrari, H. Amoud, P. Honeine, P. Flandrin, P. Borgnat, "Statistical hypothesis testing with time-frequency surrogates to check signal stationarity", *IEEE Int. Conf. on Acoust., Speech and Signal Proc. ICASSP-10*, Philadelphia (PA), Dallas (TX), 14-19 mars 2010.

[P41] A. Moghtaderi, P. Flandrin, P. Borgnat, "Time-varying spectrum estimation of uniformly modulated processes by means of surrogate data and Empirical Mode Decomposition", *IEEE Int. Conf. on Acoust., Speech and Signal Proc. ICASSP-10*, Philadelphia (PA), Dallas (TX), 14-19 mars 2010.

[P34] H. Amoud, C. Richard, P. Honeine, P. Flandrin, P. Borgnat, "Sur la caractérisation

de non-stationnaritié par la méthode des substituts", *22e Colloque sur le Traitement du Signal et des Images. GRETSI-2009*, Dijon (France), 8-11 septembre 2009.

[**P33**] E. Ben Hamida, P. Borgnat, H. Esaki, P. Abry, E. Fleury, "Live E ! Sensor Network : Correlations in Time and Space", *22e Colloque sur le Traitement du Signal et des Images. GRETSI-2009*, Dijon (France), 8-11 septembre 2009.

[**P32**] E. Leveque, J. Boudet, A. Cahuzac, P. Borgnat, F. Toschi, "Towards practical large-eddy simulation of complex turbulent flows", *Advances in Turbulence XII : Proceedings of the 12th EUROMECH European Turbulence Conference*, Marburg (Germany), Sept. 7 - 10, 2009.

[**P31**] H. Amoud, P. Honeine, C. Richard, P. Borgnat, P. Flandrin, "Time-Frequency Learning Machines For NonStationarity Detection Using Surrogates", *IEEE Workshop on Statistical Signal Processing SSP-2009*, pp. 565–568, Cardiff (UK), 31 August-2 September 2009.

[**P30**] E. Ben Hamida, H. Ochiai, H. Esaki, P. Borgnat, P. Abry, E. Fleury, "Measurement Analysis of the Live E ! Sensor Network : Spatial-Temporal Correlations and Efficient Data Aggregation", *IEEE 3rd International Workshop on Practical Applications of Sensor Networking*, workshop in Conference *SAINT 2009*, Seattle (USA), 20-24 July 2009.

[**P28**] P. Flandrin, P. Borgnat, "Sparse time-frequency distributions of chirps from a compressed sensing perspective" *8th IMA International Conference on Mathematics in Signal Processing*, Cirencester (UK) , 16-18 December 2008.

[**P27**] P. Borgnat, P. Flandrin, "Time-Frequency Surrogates for Nonstationary Signal Analysis" *8th IMA International Conference on Mathematics in Signal Processing*, Cirencester (UK), 16-18 December 2008.

[**P24**] P. Borgnat, P. Flandrin, "Revisiting and testing stationarity," *J. Phys : Conf. Series "2008 Euro American Workshop on Information Optics WIO'08"*, Annecy, June 2008.

[**P23**] P. Borgnat, P. Flandrin, "Time-frequency localization from sparsity constraints," *IEEE Int. Conf. on Acoust., Speech and Signal Proc. ICASSP-08*, Las Vegas (NV), April 2008.

[**P20**] J. Xiao, P. Borgnat, P. Flandrin, « Sur un test temps-fréquence de stationnarité », *21e Colloque sur le Traitement du Signal et des Images. GRETSI-2007*, Troyes (France), 11-14 septembre 2007.

[**P19**] J. Xiao, P. Borgnat, P. Flandrin, "Testing Stationarity with Time-Frequency Surrogates", *15th European Signal Processing Conference EUSIPCO-2007*, Poznan (Pologne), 3-7 septembre 2007.

[**P18**] J. Xiao, P. Borgnat, P. Flandrin, C. Richard, "Testing Stationarity with Surrogates – A One-Class SVM Approach," *IEEE Workshop on Statistical Signal Processing SSP-2007*, Madison (Wisconsin, USA), 26-29 août 2007.

Chapitre dans des ouvrages collectifs

[**C5**] P. Borgnat, P. Flandrin, C. Richard, A. Ferrari, H. Amoud, P. Honeine "Time-Frequency Learning Machines For NonStationarity Detection Using Surrogates", in *Data Mining and Machine Learning for Astronomical Applications*, K. Ali, A. Srivastava, J.D. Scargle, M.J. Way eds., Chapman & Hall/CRC Press, 2012.

Chapitre 3

Réseaux d'ordinateurs et signaux de télétrafic informatique

Les signaux de télétrafic portant la communication par réseaux entre les ordinateurs, au premier plan desquels le réseau Internet, sont scrutés de tous côtés depuis bien vingt ans maintenant. Il a très tôt été réalisé que les modèles poissoniens de communication des téléphones ne convenait pas aux communications sur les réseaux d'ordinateurs [LTWW93, LTWW94a, PF95], voir aussi [PKC96], qu'elles soient filaires ou sans fil. Plusieurs facteurs entrent en ligne de cause : aux temps courts, la grande variété des protocoles et des informations échangées, créent une variabilité très forte des propriétés de paquets ou flots de données passant par les réseaux ; aux temps long, ce sont des effets de corrélations à temps long [LTWW94a], ou longue mémoire [ENW96], parfois caractérisés en terme d'autosimilarité, qui dominent [WgTSW97, AV98, PWg00], ou même de multifractalité [TTW97, FGW98], avant même que des cycles journaliers ou hebdomadaires s'ajoutent en rendant le trafic non stationnaire. À toutes les échelles de temps, les propriétés statistiques auxquelles s'attendre dans du trafic d'ordinateurs sont, dans les fait, inconnues. En étant devenu le réseau de communication universel pour tous les types d'informations, du transfert simple de fichiers binaires à la transmission de la voix, de la vidéo, d'informations interactives en temps réel,... Internet est en effet rançon de son succès et victime d'anomalies de trafic (pannes, congestions, augmentations soudaines de trafic comme les phénomènes de *Flash Crowd* [JKR02] ou des propagations de virus, attaque du réseau comme des attaques de déni de service distribué (DDoS), par exemple par SYN-flooding, [Bru00, MVS01] ou les scans à large échelle [APT07]) qui réduisent l'espoir de modéliser complètement les signaux de télétrafic. Et cela est sans compter qu'on n'a pas de cartes précise des réseaux, puisqu'il sont à la fois toujours changeant [CAI, LM08] et incorporent beaucoup d'éléments privés d'opérateurs de télécommunications, qui transportent ces signaux.

Mes travaux étudiant le télétrafic Internet ont démarré fin 2004 alors qu'il était déjà clair que des approches de traitement du signal – en particulier les outils temps-échelle et les transformations en ondelettes que nous définissons en 3.2 – permettaient de bien caractériser certains aspects du télétrafic en terme de longue mémoire et d'autosimilarité, et par des modèles avec lois de probabilités à queue lourde, voir [PW00, ABF$^+$02, CMP$^+$02, DOT03a, HVA03, HVA05].

L'angle de travail qui a alors été le nôtre a été principalement de montrer comment aller plus loin que la simple caractérisation du trafic pour contribuer à l'ingénierie du trafic et l'administration des réseaux, en proposant d'un côté des modèles effectifs (à défaut

d'être exacts) des comportements attendus pour du trafic normal, et de l'autre des outils de détection et de classification pour le trafic et les ordinateurs connectés au réseau. Ce travail s'appuie beaucoup sur la métrologie des réseaux d'ordinateurs que nous discutons ci-dessous. Pouvoir mesurer le télétrafic Internet permet une réelle approche expérimentale pour l'analyser, le modéliser et proposer des outils utiles à l'administration des réseaux.

Ce travail a été une part significative de mon activité de recherche de 2005 à 2010, avec quelques prolongements plus récents. Sur ces sujets, j'ai donné des exposés d'introduction à la métrologie et la modélisation statistique du télétrafic informatique (et l'apport des outils de traitement du signal non stationnaire ou à invariance d'échelle dans le domaine) à Algotel 2007, ResCom 2008 et au workshop Tera-Net d'Icalp en 2010. Le cadre de ces travaux a été initialement une collaboration dans des projets ACI/ANR multi-partenaires français (Metrosec puis projet RNRT pré-compétitif OSCAR sur les réseaux *overlay*), et avec le NII, IIJ et l'Université de Tokyo (tous au Japon) dans un partenariat entre le CNRS et WIDE au Japon.

Ce chapitre commence par quelques rappels sur la métrologie des réseaux pour les aspects nécessaires à la compréhension de la suite en 3.1 et par une brève introduction aux ondelettes en 3.2 avant de passer à nos travaux. Ceux-ci se déclinent autour d'enjeux très familiers pour qui vient du traitement du signal tout en restant pertinents pour l'analyse du télétrafic informatique :

- *modéliser* les signaux de trafic en 3.3 [P11, J6, J7, P22, J12, P40, J13] ;
- *estimer* les propriétés du trafic 3.4 [P29] et *détecter* les anomalies présentes dans le trafic [P10, P13, P14, P15, P16, P17, P21] ainsi qu'évaluer les détecteurs [P38, P43, P44] en 3.5 ;
- *classifier* les comportements de différents ordinateurs ou des flots en 3.6 [J17, J23, Js30].

Nous dressons finalement un bilan de cette activité et quelques perspectives en 3.7.

3.1 Métrologie pour les réseaux d'ordinateurs

Internet étant devenu un réseau complexe par excellence, puisqu'on ne connaît ni exactement sa constitution ni le télétrafic qu'il transporte, la métrologie des réseaux s'est développée comme une approche expérimentale pour analyser les réseaux d'ordinateurs, voir par exemple [OLB+07, CAI].

Sans faire un cours sur les réseaux d'ordinateurs et leur métrologie, rappelons en simplifiant que la communication dans des réseaux est structurée en couches (de la couche physique à celles des applications) et que, sur l'Internet, le "quantum" d'information le plus étudié est au niveau IP (*Internet Protocol*) : le paquet IP encapsule les informations venues de l'application qui doit transférer quelque chose par le réseau, et les données liées à la couche de transport (TCP ou UDP en général) avant de les transmettre à la couche des liens (ethernet en général pour des réseaux filaires). La donnée de base en métrologie est ainsi constituée de ces paquets d'informations IP qui sont des textes décrivant les informations de routage (type de protocole IP, destination et origine données avec le numéro IP et le numéro du port), les informations nécessaires pour contrôler la réception et la vitesse de transfert (couches TCP ou UDP) suivies enfin par ce qui concerne l'application (web, mail, commande shell, etc.). Ces paquets sont envoyés sur des liens de communications (éventuellement sans

fil sur un canal radio) et il s'ajoute donc une estampille de temps qui fait que la donnée étudiée est un suite de paquets dans le temps. Les paquets ayant les mêmes origines et destinations et protocole (selon les données IP) sont regroupées au sein de ce qu'on appelle un flot. Il est donc possible de voir les données IP comme un processus ponctuel marqué, par exemple non stationnaires [KMFB04], markovien modulé stationnaire [AN98], ou même de grappe [HVA03]. Mais le très grand nombre de paquets impliqués et les marques possibles (comme l'identité du flot) ayant une dimension quasi infinie, la manipulation de ces processus et des données nécessaires est difficile. Rien que pour l'identité IP du paquet, dans la version actuelle d'IPv4 il y a 2^{32} numéros IP différents et 1024 numéros de ports possible. En IPv6 qui se déploie peu à peu pour augmenter le nombre (déjà saturé) d'adresses IP, on passe à 2^{128} adresses possibles, soit dans les 3.4×10^{38} ce qui, élevé au carré (origine et destination), fait une dimension très grande d'espace...

Par conséquent, une approche offrant plus de souplesse pour les données IP est de considérer les séries décomptant le nombre d'octets ou de paquets du trafic agrégé, notés $W_\Delta(k)$ et $X_\Delta(k)$. Elles correspondent au nombre d'octets (resp. paquets) qui transitent au cours de la k-ème fenêtre de taille $\Delta > 0$, i.e., dont les estampilles se situent entre $k\Delta \leq t_l < (k+1)\Delta$. D'autres analyses reposent sur le processus d'arrivée des flots, comme par exemple dans [BTI+02]. Dans nos travaux, nous restons la plupart du temps au niveau paquet et étudierons des statistiques des processus $X_\Delta(k)$ ou $W_\Delta(k)$. Un exemple illustre une telle série dans la prochaine section, figure 3.1.

En collectant dans une trace de trafic les paquets (tout ou partie comme l'entête IP) qui passent en un ou des différent(s) point(s) du réseau, on peut partir des mesures et des données pour caractériser le trafic, le modéliser et développer des méthodes pour détecter des événements ou classifier du trafic. Pour cela, il faut bien sûr disposer des traces fiables, documentées ou pour lesquelles on a des idées du contenu. Les traces de trafic utilisées dans nos travaux viennent essentiellement de quatre sources :

- des traces collectées avec notre participation dans le cadre des projets METROSEC et OSCAR à l'aide de cartes DAG [CDG+00] déployées sur le réseau RENATER (l'opérateur de réseau public qui connecte les universités et établissements de recherche français), en particulier à l'ENS de Lyon, et le réseau d'Orange. Les cartes DAG sont des sondes passives qui capturent en ligne des paquets, éventuellement en ne conservant que les octets du début des protocoles IP et TCP/UDP pour alléger la charge, et les enregistrent après anonymisation pour traitement ultérieur. Nous avons ainsi constitué une base de données décrite dans [P16, J6, J7] contenant des traces de trafic normal et des traces avec des anomalies (DDoS, Flash crowds) documentées.
- des traces collectées dans un travail commun avec le LIP (ENS de Lyon) sur le réseau de Grid5000 (porté par RENATER) dédiées à un usage spécifique : tester le mécanisme d'apparition de l'auto-similarité dans le trafic, voir [J13] et 3.3.
- des traces du groupe japonais MAWI [CMK00] de WIDE, avec qui nous travaillons depuis 2007. Ces traces sont publiquement disponibles sur le web [MAW] et consistent en des traces courtes quotidiennes depuis 1999, et des traces longues (24h voire 72h) régulières. Elles sont de plus en plus employées pour la recherche, en particulier grâce à nos travaux [P17, P29] validés sur ces traces et grâce aux annotations indiquant les anomalies que l'on a pu obtenir par nos travaux [P38, P43, P44].
- des traces collectées par d'autres collègues, depuis les traces « historiques » du Bell-lab [LTWW94b] jusqu'à des traces plus récentes comme celles de l'Univer-

sité d'Auckland déjà utilisée dans [HVA03] ou de l'université de Keio apparaissant dans [KcF+08]. pour plus de comparaison.

Différents problèmes d'ingénierie du télétrafic Internet peuvent alors être abordées par la métrologie des réseaux, en s'appuyant sur des collectes de traces de trafic et leurs analyses :

- comment caractériser le trafic, par exemple à des fins de planification du déploiement des réseaux et de définition des besoins adaptés,
- étudier le contenu du trafic : est-ce du trafic légitime ou non, normal, inattendu, des attaques,... et donc potentiellement intervenir sur les communications qui passent pour éviter de dysfonctionnements ou mettre en place des systèmes de détection d'intrusions (IDS) ou d'autres défenses contre les anomalies,
- classifier les profils d'activités des ordinateurs connectés, classifier les types de flots ou d'applications.

Notons qu'au moment des débuts de nos travaux les outils préférés des administrateurs réseaux pour détecter des anomalies s'appuyaient sur des détections par signature (c'est-à-dire qu'on tente de reconnaître la signature pré-définie d'attaques, de virus, etc., souvent par fouille exhaustive dans les paquets). L'approche par détection profil statistique qui sera discutée ici a pour précurseur le travail [BKPR02] en 2002. Les méthodes alternatives aux nôtres ont été proposées majoritairement entre 2005 et 2010.

3.2 Un outil d'analyse privilégié : les ondelettes

La transformation en ondelettes [Mey90, Dau92, Mal99] a été mise en avant comme un très bon estimateur spectral dans le contexte des traces de télétrafic informatique [AV98, VA99, AFTV00, VA01]. Pour ce mémoire, il est intéressant de rappeler (en suivant [Fla99]) que la transformation en ondelettes peut être vue comme un analogue de la transformée de Fourier à court terme (qui mise au carré conduit au spectrogramme (2.3)) en remplaçant l'opérateur de translation en fréquence ($\times e^{-i2\pi fs}$) par l'opérateur de translation en échelle, ou dilatation d'un facteur $a > 0$:

$$T_x(a,t) = \int_{-\infty}^{+\infty} x(u)\, \frac{1}{\sqrt{a}} \psi_0 \left(\frac{u-t}{a} \right) ds \tag{3.1}$$

Le lecteur trouvera bien plus de détails sur la transformée en ondelettes dans des livres (très nombreux) sur le sujet, par exemple [Mey90, Dau92, Mal99]. Le terme temps-échelle [Fla99] a été proposé pour cette transformation puisqu'elle réalise l'équivalent des méthodes temps-fréquence du chapitre précédent en remplaçant la fréquence par cette notion d'échelle. Pour la suite de ce travail, nous avons besoin de deux points principaux :

1. la fonction ψ_0, appelée ondelette mère, doit être une fonction oscillante dont l'énergie est plutôt concentrée en temps et en fréquence. Pour inverser la formule (3.1), il faut qu'elle vérifie la condition d'admissibilité [Dau92, Mal99] qui implique qu'elle est à moyenne nulle. Un paramètre utile caractérisant l'ondelette est alors son nombre de moments nuls $N_\psi \geq 1$ tel que $\forall k = 0, ..., N_\psi - 1$, $\int t^k \psi_0(t) dt = 0$ et $\int t^{N_\psi} \psi_0(t) dt \neq 0$. L'équation (3.1) définit ainsi des coefficients $T_x(t,a)$ d'une transformée en ondelettes continue (ou CWT pour *Continuous Wavelet Transform*) dans le paramètre en échelle a qui joue le rôle d'un inverse de la fréquence.
2. pour passer à une base discrète, il faut choisir ψ_0 tel que $\{\psi_{j,k}(t) = 2^{-j/2}\psi_0(2^{-j}t - k), k \in \mathbb{Z}, j \in \mathbb{Z}^{*,+}\}$ forme une base de $L^2(\mathbb{R})$; la correspondance avec le cas continue

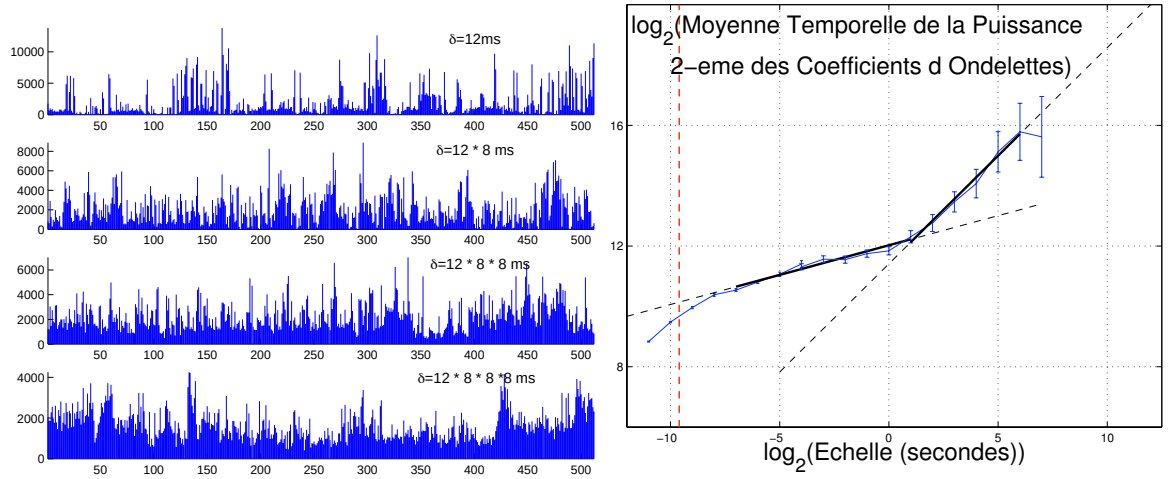


FIGURE 3.1 – Exemple de séries agrégées de télétrafic (mesures d'Auckland), pour différents niveaux d'agrégation Δ (de 12 ms à 12*8*8*8 = 6144 ms). La comparaison visuelle de ces séries – et leur similarité – nous renseigne déjà sur l'existence de propriétés (au moins partielles) d'invariance d'échelle. À droite, le diagramme log-échelle (*LD plot*) montre le graphe $\log_2 \bar{S}(j)$ vs. $\log_2 2^j = j$. On y voit deux zones bien distinctes caractéristiques, dont celle de droite porte la longue mémoire de la série et est étudiée plus avant ici.

se fait avec $a = 2^{-j}$. Cela se fait en construisant une analyse multirésolution [Dau92, Mal99]. On obtient alors les coefficients $d_x(j,k)$ d'une transformée en ondelettes discrète (ou DWT pour *Discrete Wavelet Transform*) par

$$d_x(j,k) = \langle \psi_{j,k}, x \rangle. \tag{3.2}$$

Pour en savoir plus sur les ondelettes discrètes et de leurs utilisations, on consultera par exemple [Mal99]. La correspondance avec le cas continu se fait par $d_x(j,k) = T_x(a = 2^{-j}, t = k2^j)$.

Regarder les coefficients d'ondelettes continues $T_x(a,t)$ pour plusieurs échelles a, ou ceux des ondelettes discrètes $d_x(j,k)$ à plusieurs j, constitue une représentation multi-échelle, aussi dite multi-résolution, des données étudiées.

Pour l'analyse du télétrafic, une propriété importante des ondelettes est qu'on obtient un bon estimateur spectral grâce aux ondelettes [AV98, AGF95]. On peut montrer que pour un processus stationnaire (au second ordre) Y, on a :

$$\mathbb{E}\,|d_Y(j,k)|^2 = \int_{-\infty}^{+\infty} S_Y(\nu) 2^j |\tilde{\Psi}_0|^2 d\nu, \tag{3.3}$$

où $\tilde{\Psi}_0$ est la transformée de Fourier de ψ_0 et S_Y le spectre de Y.

Cela est particulièrement intéressant pour des processus où le spectre a un comportement proche d'une loi de puissance, en particulier pour ceux dits à mémoire longue [Ber94] et il se trouve que le trafic présente beaucoup d'indications d'une telle propriété [LTWW93, PWg00]. Rappelons que la propriété de mémoire longue se définit par une divergence en loi de puissance du spectre à l'origine [PWg00, Ber94, DOT03b]). Pour un tel processus Y, il vérifie pour des constantes $D > 0$ et $c \in]0,1[$:

$$S_Y(\nu) \sim D|\nu|^{-c}, \text{ pour } |\nu| \to 0. \tag{3.4}$$

De manière équivalente, cela s'écrit dans le domaine temporel pour la fonction de corrélation $\gamma_Y(\tau) \sim D'|\tau|^{-(1-c)}$, pour $|\tau| \to +\infty$. La correspondance avec les processus auto-similaires [Ber94, ST94] conduit à poser $c = 2H - 1$ où H est l'exposant de Hurst. L'interprétation est qu'on ne peut pas singulariser de temps caractéristique pour un tel processus. Une conséquence majeur est la dégradation des procédures usuelles d'estimation spectrale [Ber94]. Passer par les ondelettes conduit à

$$\mathbb{E}|d_Y(j,k)|^2 \sim C2^{j(2H-1)} \text{ pour } 2^j \to +\infty. \tag{3.5}$$

Cette équation suggère de regarder le spectre moyen dans le temps, $\bar{S}(j) = \frac{1}{n_j}\sum_k |d_Y(j,k)|^2$, en log-log. Le graphe $\log_2 \bar{S}(j)$ vs. $\log_2 2^j = j$ est appelé diagramme log-échelle (*LD plot*) [AV98, AGF95] et permet une estimation de H grâce à une régression linéaire pondérée [VA99, AFTV00] :

$$\hat{H} = \frac{1}{2}\left(1 + \sum_{j=j_1}^{j_2} w_j \log_2 S(j)\right), \tag{3.6}$$

où les poids w_j vérifient les contraintes de la régression linéaire [AV98, AGF95]. Une trace de trafic agrégée à différentes échelles de temps est montrée en figure 3.1 avec le diagramme log-échelle correspondant, pour le trafic Auckland-IV (2001) [HVA03]. Il suit une forme assez usuelle où l'on voit une pente linéaire à grande échelle caractéristique de la longue mémoire. Une zone différente à petite échelle semble ici linéaire mais n'est génériquement pas associé à un processus invariant d'échelle [HVA05]. Cette forme coudée, avec un coude vers 1s, est fréquente dans les analyses de trafic IP [HVA03, HVA05] [J6].

3.3 Modèles de trafic : validation par la mesure

Muni de traces de trafic IP et d'un outil d'analyse, les ondelettes, nous avons pu contribuer à l'étude du mécanisme de génération de la propriété de longue mémoire dans le trafic théoriquement [J12] et expérimentalement [J13, P40]. Nous avons proposé et utilisé dans plusieurs communications [J6, J7, P10, P11, P13, P14, P16] un modèle effectif de trafic, rendant compte de cette propriété mais plus versatile que les modèles théoriques pour qu'il soit utilisable pour caractériser et détecter des anomalies.

3.3.1 Étude de la longue mémoire dans le trafic

Dès les années 90 et les premières mesures de trafic IP, une propriété de longue mémoire était apparent et un mécanisme a été propose dans [LTWW93, LTWW94a] pour en rendre compte. L'idée est de relier la longue mémoire des séries agrégées à une distribution large (à queue lourde) des tailles des flots. C'est le théorème de Taqqu pour le trafic Internet [LTWW94a, PKC96, TWS97, PWg00]. Il est établi pour un modèle dit "On/Off" fluide qui représente chaque flot par un signal d'activité constante (typiquement 1) quand il est actif et zéro quand il est silencieux ; c'est l'aspect "fluide" du modèle : on a gommé la constitution en paquets pour la remplacer par une constante. On écrit la série de trafic agrégé comme $X_N(t) = \sum_{i=1}^N Z_i(t)$ où les $\{Z_i(t), t \in\}_{i=1,\ldots,N}$ sont un ensemble de processus de renouvellement binaire (prenant des valeurs 0 ou 1) avec des périodes d'activation indépendantes.

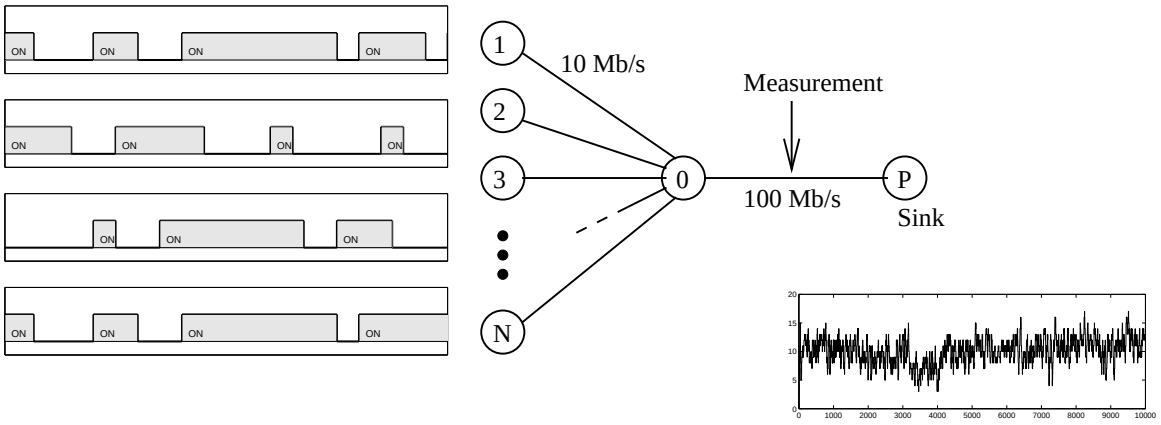


FIGURE 3.2 – Figuration du modèle On/Off fluide : c'est l'équivalent pour les réseaux d'un processus de renouvellement M/G/N On/Off. Ce modèle est utilisé pour des simulations Matlab et NS-2 pour voir comment obtenir expérimentalement les conditions du théorème de Taqqu.

Supposons que les période "On", notées τ_{on}, sont distribuées comme une variable aléatoire i.i.d. Supposons aussi que la distribution a une queue lourde de paramètre α, c'est-à-dire qu'elle vérifie que $P(\tau_{\text{on}} > w) = 1 - F_W(w) \sim cw^{-\alpha}$ pour $w \to +\infty$, avec $\alpha < 2$ (variable à variance infinie, et même moyenne infinie si $\alpha < 1$) [AFT98, ST94]. De la même manière, les périodes "Off" sont supposées à queue lourde avec un exposant β. Une représentation d'une série obtenue selon ce modèle est donnée en figure 3.2, avec N sources qui s'agrègent sur un lien unique entre 0 et P, le point de mesure étant sur le lien entre ces deux points. On regardant la série de trafic agrégée $X_N(t)$ à travers sa somme cumulée $Y_N(t)$:

$$Y_N(tT) = \int_0^{Tt} X_N(u)\mathrm{d}u = \int_0^{Tt} \left(\sum_{i=1}^{N} Z_i(u) \right) \mathrm{d}u. \tag{3.7}$$

Le théorème de Taqqu [LTWW94a, TWS97] prouve qu'il existe une constante positive C telle que :

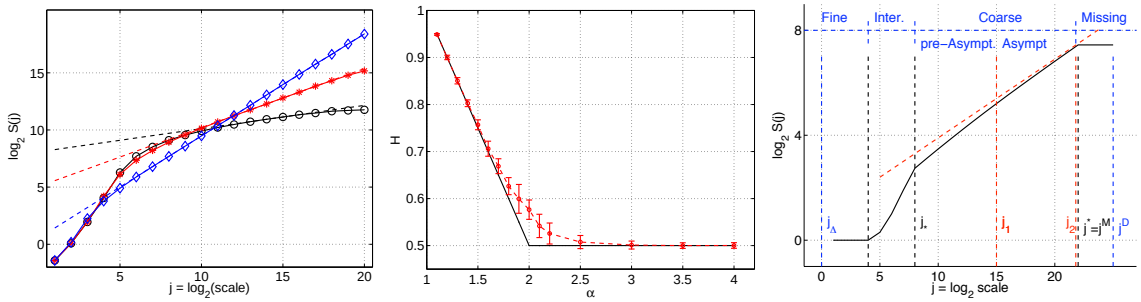
$$\lim_{T\to+\infty} \lim_{N\to+\infty} \frac{Y_N(tT) - \dfrac{\mathbb{E}\,\tau_{\text{on}}}{\mathbb{E}\,\tau_{\text{on}} + \mathbb{E}\,\tau_{\text{off}}} NTt}{C\sqrt{N}T^H} = B_H(t), \tag{3.8}$$

où B_H est un mouvement brownien fractionnaire de paramètre de Hurst H satisfaisant :

$$H = \frac{3-\alpha^*}{2}, \text{ avec } \alpha^* = \min(\alpha, \beta, 2). \tag{3.9}$$

Le point essentiel de ce théorème est qu'une fois enlevée la tendance linéaire (qui est la somme cumulée de la moyenne stationnaire de (3.7)), les fluctuations qui restent sont celles d'un mouvement brownien, stationnaire dans la limite d'un nombre infini de flots possibles (N) et d'un temps d'observation (T) infini. Ce théorème propose un mécanisme pour la longue mémoire dans le trafic puisque l'on sait qu'en dérivant B_H, on obtient un processus stationnaire à longue mémoire d'exposant $c = 2H - 1$ si α^* est plus grand que 2 (le cas égal à 2 correspondant à $H = 1/2$, soit un processus de fluctuation sans mémoire) [Ber94, ST94].

Ce résultat est obtenu asymptotiquement pour le modèle (3.7) mais il reste que c'est un modèle génératif (certes très simple) de trafic. Pour du trafic, on s'attend selon les



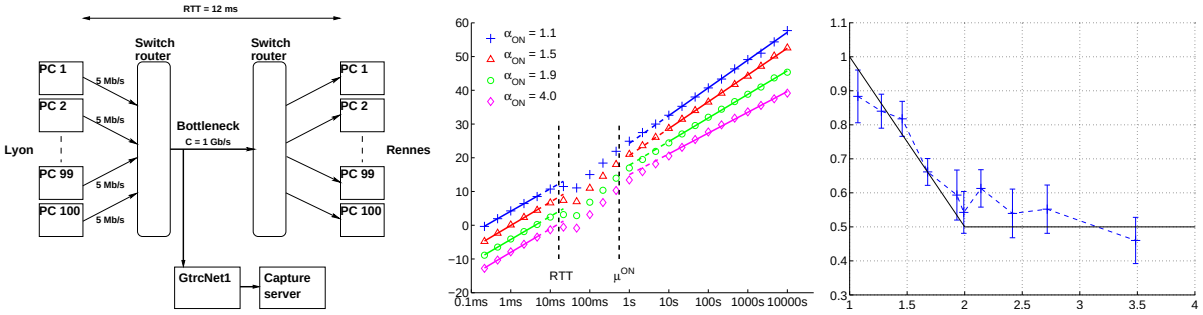
(a) (b) (c)

FIGURE 3.3 – Pour le modèle de la figure 3.2, on montre : (a) les LDs pour $\alpha = 1.1$ ($\diamond$), $\alpha = 1.5$ ($*$), $\alpha = 1.9$ (o). Le comportement asymptotique en loi d'échelle commence à une échelle j_1 qui augmente quand α augmente dans $[1, 2]$. (b) Relation $H(\alpha)$ estimée sur des simulations de trafic des intervalles de confiance à 95%, comparée à la prédiction théorique de Taqqu. Bon accord, sauf pour α proche de 2. (c) Diagramme log-échelle schématique reprenant les trois grandes zones en échelle trouvées pour le trafic (voir détails dans [J12]) avec des zones intermédiaires (ou à très long temps les échelles absentes) qui peuvent être larges et compliquent l'estimation des propriétés asymptotiques (ici à temps long pour H). L'existence de ces zones explique la difficulté d'estimer la relation de Taqqu pour α qui s'approche de 2.

mesures de trafic en condition normale à avoir un H entre 1/2 et 1, soit de la longue mémoire [Ber94, ST94]. Du point du vue de l'étude des réseaux d'ordinateurs, ce mécanisme se comprend assez bien car les flots sont en général actifs en fonction soit de l'activité humaine, soit en fonction de la taille des données à transférer. Comme dans beaucoup de situations liées à l'activités humaine, ce sont des quantités qui peuvent varier sur des ordres de grandeurs, avec des lois de distribution larges voire en lois de puissance. On peut consulter [PKC96, CB96] à ce sujet sauf que, lors de ces travaux, aucune mesure statistiquement pertinente simultanée des paramètres H, α (et éventuellement β) n'était possible à l'époque, par manque d'enregistrement assez longs et par manque d'outils adaptés (les ondelettes en particulier).

L'analyse théorique de [J12] consistait à montrer que la nature asymptotique du théorème de Taqqu avait des conséquences sur la manière dont on peut mesurer la longue mémoire du trafic à l'aide des outils basés sur une transformée en ondelettes. Suivant cette analyse, nous avons été les premiers à montrer que la correspondance théorique est effectivement mesurable numériquement sur des séries à longue mémoire. Les graphes de la figure 3.3 résument ce que nous avons obtenu par simulation en Matlab (et des simulations d'un réseau d'ordinateurs par NS-2 ont aussi été effectuées qui confortent le résultat) selon la topologie de 3.2 : le diagramme log-échelle des processus fluides "On/Off" voient la pente à grande échelle varier en fonction de l'exposant α et la meilleure estimation possible de H (en sélectionnant convenablement les échelles selon la discussion de [J12]) conduit à des estimées compatibles avec la relation (3.9), sauf pour α vers 2 qui est le cas limite où l'on a le moins d'échelles disponibles pour estimer H. Selon 3.3 (c), on voit qu'il y a plusieurs zones attendues en fonction de l'échelle j et qu'il faut impérativement se mettre dans la zone à grande échelle ou le résultat asympotique de (3.8) est valide.

Après avoir validé que la relation (3.8) est observable numériquement malgré son caractère asymptotique, nous avons pu la tester expérimentalement par des mesures sur un réseau d'ordinateurs [J13]. Nous avons réalisé avec l'équipe RESO (projet INRIA, au LIP à



(a) (b) (c)

FIGURE 3.4 – Expériences menées sur Grid5000 pour vérifier la relation de Taqqu. (a) topologie expérimentale utilisée. (b) diagrammes log-échelle en fonction du paramètre α_{ON} qui contrôle les queues des flots émis. On note en comparant à 3.3 que les courbes sont bien différentes ici à petite échelle : le mécanisme TCP (avec les temps caractéristiques du RTT et μ_{ON}) pilote ce qui s'y passe là où la précédente figure avait une pente contrôlé par la caractère fluide du modèle à petite échelle. (c) Relation $H(\alpha)$ estimée sur le vrai trafic collecté, avec les écarts types figurés.

l'ENSL) des expériences de métrologie sur Grid 5000, que l'on pouvait contrôler pour générer du trafic selon le modèle 'On/Off" mais avec une structure en paquets TCP ou UDP plutôt que fluide comme dans les précédentes simulations. Le résultat est qu'on reproduit exactement les propriétés déduites du théorème de Taqqu [J13]. En figure 3.4, on montre en (a) la topologie expérimentale, en forme de "papillon", réalisée sur Grid 5000, qui correspond à la situation modèle de la figure 3.2 avec plusieurs ordinateurs comme destination au lieu d'une seule car c'est un cas réaliste réalisable (tandis que c'etait plus lourd à réaliser par NS-2), en (b) les diagrammes log-échelle obtenus (dont on voit qu'ils diffèrent beaucoup à petite échelle des mêmes courbes dans le cas fluide (voir 3.3 (a)) et en (c) la relation expérimentale, statistiquement validée, entre $\hat{H}$ et α pour du trafic Internet.

Continuant ensuite sur ce résultat pour l'étude des traces de trafic informatiques, nous avons reposé des questions sur les rôles relatifs des flots et des sessions dans l'approche usuelle des modèles de trafic [P40] (nous ne détaillerons pas plus cela ici).

3.3.2 Proposition d'un modèle effectif de trafic : les processus Gamma-fARIMA

Par les travaux précédents, nous disposons d'un mécanisme pour modéliser ce qui se passe dans le trafic à grande échelle. Pour les petites échelles, le mécanisme à l'œuvre dans le réseau est bien plus controversé ; faut-il chercher de la multifractalité [TTW97, FGW98, HVA05] ? des lois d'échelles [RZMD05] ? étudier les propriétés des flots ? [BTI+02] Plutôt qu'une approche théorique, nous avons privilégié dans une série de communications [J6, P10, P11, P13, P14, P16], une approche d'ingénierie pour proposer un modèle qui s'appuie sur les statistiques d'ordre 1 et 2 du processus, sans proposer de mécanisme « réseau » de génération du trafic. L'objectif pragmatique est d'avoir un modèle qui ressemble à du trafic habituel, sans idée sur comment il se connecte aux protocoles, au routage, etc. dans le réseau. Pour ce faire, notre proposition est de modéliser les séries temporelles de trafic agrégé (en octets ou paquets) comme des processus stochastiques non gaussiens à longue mémoire, que l'on estime en tant que processus stationnaires Gamma-fARIMA [Ber94] [J6, J7].

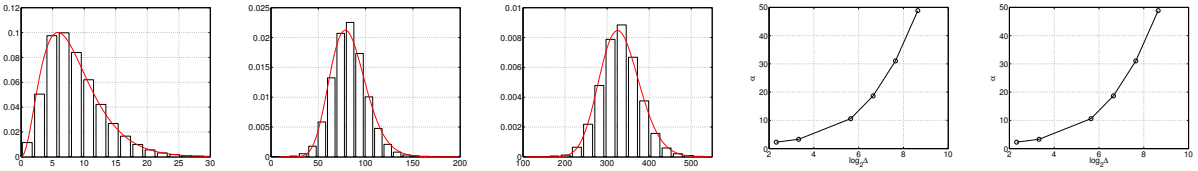


FIGURE 3.5 – Pour le trafic d'Auckland-IV, ajustement des marginales par une loi $\Gamma_{\alpha,\beta}$, pour $\Delta = 10, 100, 400$ ms (de gauche à droite) et évolution de α et β estimés avec Δ.

Un processus Gamma-fARIMA est décrit par ses propriétés statistiques d'ordre 1 et 2. À l'ordre 1, la distribution marginale est prise comme une loi Gamma [Mel93] :

$$\Gamma_{\alpha,\beta}(x) = \frac{1}{\beta}\Gamma(\alpha)\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left(-\frac{x}{\beta}\right), \tag{3.10}$$

où $\Gamma(u)$ est la fonction Gamma standard (voir [EHP00]). Elle dépend de deux paramètres : la forme α et l'échelle β. Sa moyenne est $\mu = \alpha\beta$ et sa variance $\sigma^2 = \alpha\beta^2$. La figure 3.5 montre sur des données d'Auckland à différentes échelles d'agrégation Δ, l'ajustement de la loi marginale par une distribution Gamma en $(\alpha_\Delta, \beta_\Delta)$. On voit qu'en agrégeant sur un temps plus long, la distribution s'éloigne de plus en plus d'une loi qui serait exponentielle ($\alpha = 1$) et s'approche (lentement) d'une gaussienne. À noter que l'inverse du paramètre de forme, $1/\alpha$, agit comme un indicateur de la distance avec une loi gaussienne. Cependant, un modèle gaussien ne tient pas vraiment sauf aux temps très longs (largement plus que quelques minutes) et serait peu opérant par exemple pour la détection rapide d'anomalies (quelques secondes), voir 3.5. Le modèle en loi Γ fournit, depuis les temps d'agrégation les plus courts jusqu'aux plus longs, une évolution douce des exponentielles aux gaussiennes (cela exploite fondamentalement la stabilité sous addition (facteur de forme) des lois gamma). On voit sur la figure 3.5 qu'il est pertinent de regarder les courbes de α et β en fonction de Δ car elles ne suivent pas les lois qu'on obtiendrait par agrégation si il n'y avait pas de dépendance dans la série X_Δ. Sans dépendance, l'agrégation de données s'écrivant $X_{2\Delta}(k) = X_\Delta(2k) + X_\Delta(2k+1)$, devrait donner des lois α_Δ augmentant de façon linéaire avec Δ alors que β_Δ reste constant. Ce n'est clairement pas le cas ici. Garder donc les courbes expérimentales nous renseigne sur une signature de l'existence de corrélations à courts temps et nous l'utiliserons en pratique pour la détection d'anomalies.

Au deuxième ordre statistique, on postule un modèle fARIMA (modèle auto-régressif à moyenne ajustée avec intégration fractionnaire) [Ber94]. Un tel modèle est défini par son spectre. Un coefficient $d \in [-1/2, 1/2]$ est associé à l'intégration fractionnaire et deux polynômes d'ordre P et Q ajustent les corrélations à temps courts de telle sorte que son spectre est

$$S_X(\nu) = \sigma_\epsilon^2 |1 - e^{-i2\pi\nu}|^{-2d} \frac{|1 - \sum_{q=1}^{Q} \theta_q e^{-iq2\pi\nu}|^2}{|1 - \sum_{p=1}^{P} \phi_p e^{-ip2\pi\nu}|^2}, \tag{3.11}$$

pour les fréquences $-1/2 < \nu < 1/2$. Une conséquence immédiate est que, pour $d \in (0, 1/2)$, ce processus est à mémoire longue, avec $c = 2d$ ou $H = d + 1/2$ en terme d'exposant de Hurst équivalent. Les polynômes d'ordre P et Q permettent de reproduire le spectre aux hautes fréquences (i.e. les petites échelles), alors que d représente l'intensité de la mémoire longue (i.e. les grandes échelles). Ici, il sera suffisant de recourir à des polynômes P et Q de degrés au

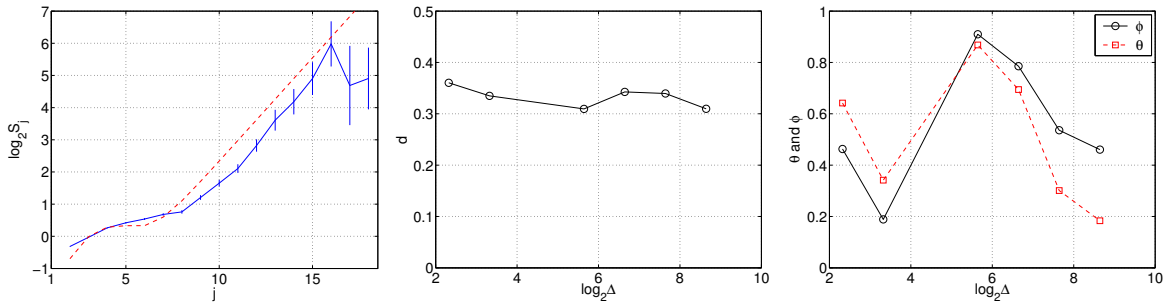


FIGURE 3.6 – Pour le trafic d'Auckland-IV, ajustement du diagramme log-échelle de la covariance empirique par un fARIMA(1,1) de paramètres (ϕ, d, θ), pour $\Delta = 10$ms ; $j = 1$ correspond donc à 10 ms. À droite : évolution de d, et des ϕ et θ avec Δ.

plus égal à 1, de coefficient respectivement ϕ et θ, pour limiter à 3 paramètres l'ajustement du spectre. Plutôt que s'appuyer sur un maximum de vraisemblance fondée sur la forme analytique du spectre (équation (3.11)), approche lourde en temps de traitement, nous avons proposé une procédure en deux étapes : estimation du paramètre de longue mémoire d en utilisant la méthode présentée avec les ondelettes et le diagramme log-échelle [AV98, VA99] ; faire alors une dérivation fractionnaire d'ordre d de X_Δ qui élimine ainsi la longue mémoire du processus, de sorte qu'il ne reste plus que les composants ARMA (corrélations à temps courts) à estimer, à l'aide d'une procédure itérative très classique [Lju99]. La figure 3.6 montre le résultat de la procédure à gauche et on voit que le diagramme log-échelle en coude est correctement ajusté par le modèle fARIMA. À droite, on a reporté les paramètres d et (ϕ, θ) en fonction de Δ : on voit que d reste quasi contant, ce qui est attendu car il décrit les échelles de temps long qui ne sont pas affectées par l'agrégation ; à l'inverse, (ϕ, θ) évoluent sans cesse signe que l'agrégation affecte bien la corrélation à court temps qu'il ne faut pas non plus négliger.

En plus de rendre compte pragmatiquement de la longue mémoire (corrélation aux temps longs) et des dépendances aux temps courts via α_Δ et β_Δ ou (ϕ, θ) (paramètres redondants mais moins précis), pour le trafic IP, ce modèle accommode sans problème les fluctuations en volume du trafic qui surviennent naturellement du fait des rythmes journaliers (la stabilité sous multiplication par une constante (facteur d'échelle) des lois gammas est utile). Notons que nous avons étendu la modélisation statistique pour étudier conjointement les statistiques à plusieurs échelles et leur corrélation grâce à des lois gamma bivariées et des lois de Bessel [P22]. Enfin, on peut employer la technique de *circulant embedded matrix* [WC94b] ou les techniques de signaux substituts [P51] pour générer numériquement des trajectoires de ces processus aléatoires, y compris le cas bivarié. Le modèle a été employé par des collègues dans [Janowski et. al 2007, 2009] pour la simulation d'effets de files d'attente dans les réseaux.

Cependant, en confrontant le modèle Gamma-fARIMA à du trafic en situation anormale, les caractéristiques ne suivent pas toujours ce à quoi l'on s'attend. En figure 3.7, le modèle tient plutôt bien autant pour la série agrégée de paquets que d'octets, avec par exemple avec un exposant de Hurst pour la longue mémoire légèrement plus grand que 0.9. Dans des situations réelles, le trafic peut être affecté par de la congestion sur un lien (saturation de la bande passante) comme en figure 3.8, qui supprime alors la variabilité du trafic totale pour la série d'octets à temps d'agrégation long (et le H effectif obtenu n'est plus de la longue mémoire selon le résultat de Taqqu, il est plus petit que 0.5) ; le trafic peut être fortement

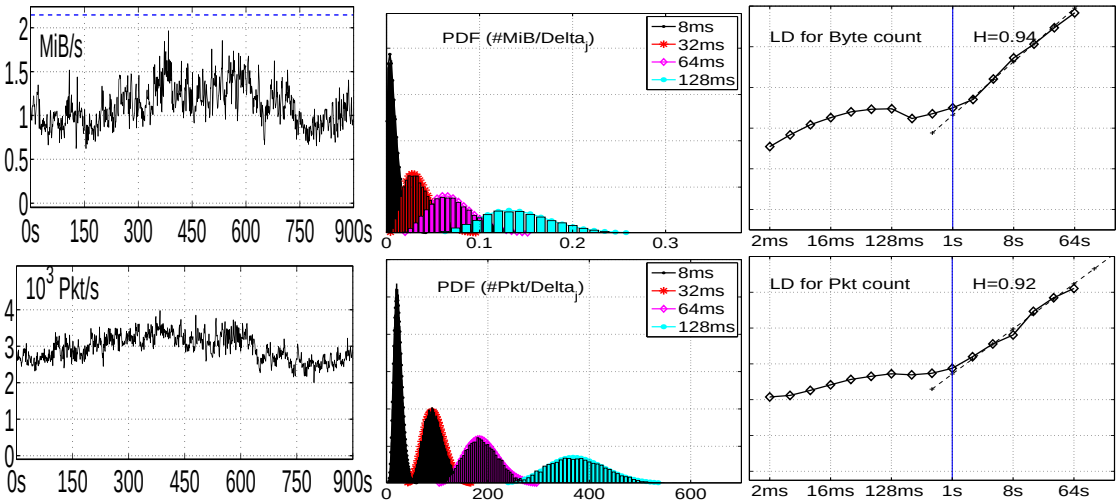


FIGURE 3.7 – Exemple de trafic MAWI sur un lien à 18 Mpbs CAR, point de mesure **B**, US vers Japon, 2005/07/11. État de trafic normal.

dominé par une ou plusieurs anomalies comme en 3.9 et ici le paramètre de Hurst estimé n'est plus le même pour les séries X et W et a pour valeur 1 ce qui montre que le mécanisme n'est plus celui de Taqqu. On a même l'impression de voir un pic d'énergie autour de l'échelle 1s. On constate cependant que la forme en coude du diagramme log-échelle reste valable, même si la longue mémoire n'est pas gouvernée par ce mécanisme, et que les marginales sont toujours ajustables par des lois Gamma.

3.4 Les *sketches* pour estimation-détection robuste de signaux

Analyser le trafic sur Internet par les signaux agrégés IP pose en fait un problème de fond : l'agrégation oublie complètement les marques (l'entête IP des paquets par exemple) qui structurent en réalité le trafic et on jette donc la plus grande majorité de l'information des paquets en faisant cela. Une manière de garder cette information serait de regarder des séries décomposées par flots, ce qui a par exemple été regardé dans les modèles de *Cluster Point Process* proposés pour le trafic dans [HVA03]. Le problème est alors qu'on garde beaucoup de données et qu'il est difficile de développer des méthodes d'estimation ou de détection pour les signaux IP sur cette base. Notre contribution a été de se mettre à mi-chemin : ne pas jeter les entêtes IP mais les employer pour agréger les données avec une stratégie d'échantillonnage aléatoire (mais inversible) dans l'espace des entêtes IP.

3.4.1 Méthodes à *sketches* pour des signaux IP

La structure retenue pour l'agrégation alétatoire, est celle des *sketches* et vient du domaine du *data streaming* [Mut03, KSZC03, CM05]. Elle sépare le trafic reçu en une collection de sous-traces en appliquant une opération de hachage aléatoire sur un (ou plusieurs) attribut(s) choisi(s) a priori (adresse IP source, adresse IP destination,...). Soit h_n, $n \in \{1, ..., N\}$, des fonctions de hachage $k-$universelles, générées de manière indépendantes en partant de N

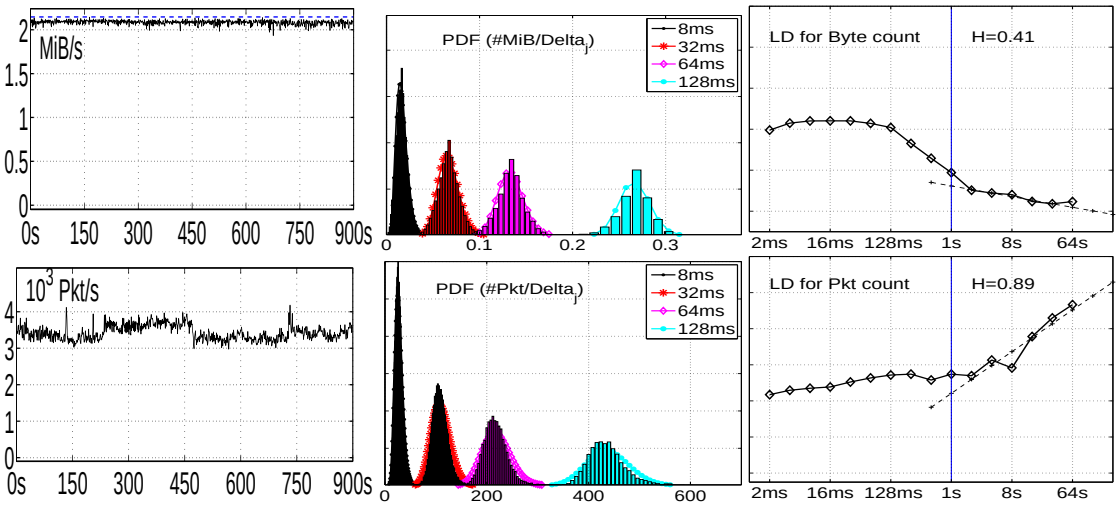


FIGURE 3.8 – Exemple de trafic MAWI sur un lien à 18 Mpbs CAR, point de mesure **B**, US vers Japon, 2003/06/03. Etat de trafic avec congestion du lien.

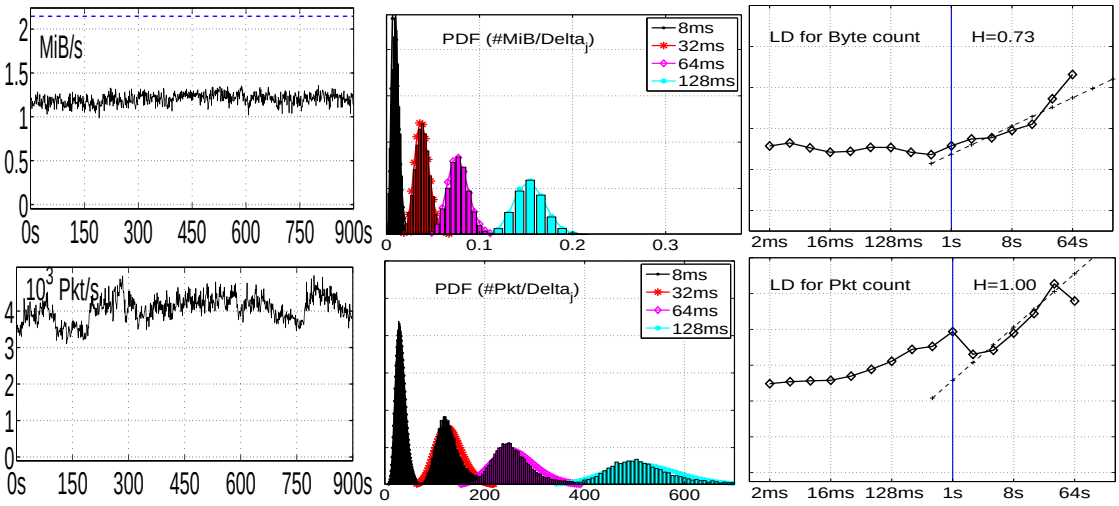


FIGURE 3.9 – Exemple de trafic MAWI sur un lien à 18 Mpbs CAR, point de mesure **B**, Japon vers US, 2004/09/21. Trafic avec anomalies : scan réseau ; spoofed flooding ; attaque sur Realserver.

graines aléatoires différentes. Ces h_n sont construites à l'aide de la méthode de tabulation rapide proposée dans [TZ04], qui construit facilement des fonctions de hachages 4-universelles. Appelons M la taille des tables de hachages (la même pour tout n).

Notons la suite d'arrivée des paquets $\{t_i, \{x_{i,l}, l = 1, ...5\}\}$ avec l'estampille de temps t_i et le 5−uplet IP usuel (protocole IP, IPsrc, IPdst, sPort, dPort) pour chaque paquet $i = 1, ..., I$. En choisissant A_i une donnée de hachage (par exemple, $A_i = x_{i,2} = \text{IPsrc}_i$, mais on peut combiner des informations du 5-uplet IP), on créé M sous-traces pour chaque fonction h_n, à partir de la trace complète $\{t_i, \{x_{i,l}, l = 1, ...4\}, i = 1, ..., I\}$ en groupant les paquets en fonction de la clef de hachage :

$$\{t_i, m_{n,i} = h_n(A_i) = m, i = 1, ..., I\}_{n,m}. \tag{3.12}$$

L'idée qui guide cette construction est celle d'un sous-échantillonnage aléatoire de la trace, qui permet d'avoir M réalisations partielles du trafic, et l'on peut reproduire N fois de

manière indépendante de manière à avoir $N \times M$ sous-traces. Avoir plusieurs réalisations permet, pour l'estimation, de faire des moyennes ou, mieux, des médianes pour obtenir des estimateurs robustes, sur ces sous-traces ; cela permet pour de la détection d'événement de mieux définir une référence de trafic normal – sans a priori sur le trafic –, par comparaison entre les sous-traces. La puissance de la méthode est que ce classement en sous-trace est possible en temps réel grâce aux techniques de hachage rapide [TZ04]. Un dernier avantage est qu'en disposant de plusieurs h_n, on peut inverser la table de hachage pour remonter exactement à la valeur de hachage incriminée (adresse IP source ou destination par exemple) dans le cas d'une détection d'anomalie (ce sera décrit en 3.5).

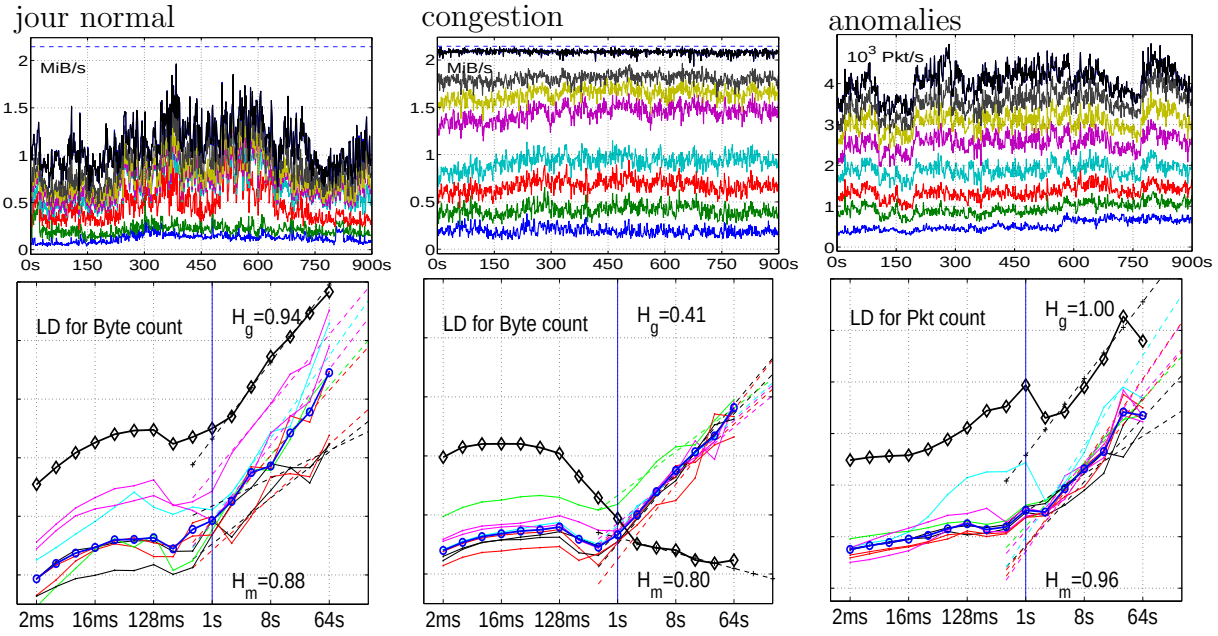
Il est important de noter que la procédure d'analyse par sketches et les outils de détection et de classification du trafic qui s'en déduisent ont été implantés de manière à pouvoir fonctionner en temps réel. Les programmes constituent même un ensemble d'outils de dépouillement et d'analyse de traces (au niveau paquet ou flot), appelé IPTools, qui a été déposé à l'APP par l'ENSL et le CNRS [L1] et est mis à la disposition de la communauté de recherche.

Cette technique a été employée pour une analyse longitudinale robuste des paramètres du trafic [P29] et pour la détection d'anomalies dans le trafic IP au niveau des paquets [P15, P17] et des flots [P21] ; nous l'avons aussi utilisée pour obtenir des caractéristiques utiles à la classification de trafic dans [J17, J23]. Décrivons les résultats des ces différents travaux.

3.4.2 Estimation robuste pour les séries agrégées IP

Appliquant l'échantillonnage par sketches, nous construisons ensuite $N \times M$ séries temporelles agrégées à une échelle de temps fine Δ_0 (typiquement 10 ms) : $X_0^{n,m}(t)$ est alors le nombre de paquets pendant Δ_0 au temps t, pour la sortie m de la table de hachage n. Nous pouvons alors utiliser la modélisation fARIMA sur chaque série agrégée. Si l'on souhaite faire une estimation robuste, au sens statistique du terme qui est robuste à l'existence d'anomalies ou de comportements extrêmes, nous prendrons la médiane sur m (puisque d'éventuelles anomalies seront pour certains ordinateurs, donc certains m, mais pas tous).

En figure 3.10, on voit l'effet de transformer des traces de trafic en sous-traces par des sketches avec $M = 8$ (et $N = 1$) avant d'agréger ces sous-traces. L'analyse de la série agrégée totale se révèle très sensible aux conditions de trafic (deux des figures précédentes : une cas où le lien est saturé par congestion et un cas avec des anomalies) et présente un diagramme log-échelle atypique avec un paramètre de Hurst H_g en dehors des valeurs attendues selon le modèle de Taqqu et les mesures correspondantes. La situation est différente pour chaque sous-trace et surtout le diagramme log-échelle de leur médiane. Celui de chaque sous-trace présente une certaine variabilité, avec des volumes (hauteur de la courbe) différentes, des petits pics locaux parfois représentatifs de temps caractéristiques ; prendre la médiane lisse tout cela et conduit à un diagramme log-échelle coudé attendu pour du trafic. On peut alors mieux estimer l'exposant de Hurst médian H_m lié à la longue mémoire du trafic qui reste entre 0.8 et 0.9 que l'on soit dans une situation de trafic normal ou une situation de trafic avec congestion. Ce résultat, discuté plus longuement dans [P29], montre que la propriété de longue mémoire du trafic ne disparaît pas quand le volume de trafic augmente sur un lien (ici, il occupe même dans les 95% de la capacité nominale du lien en situation de congestion), contrairement à ce que certains travaux pensaient mesurer [KMF04].



B-US2Jp, 2005/07/11 **B**-US2Jp, 2003/06/03 **B**-Jp2US, 2004/09/21

FIGURE 3.10 – Estimation robuste avec les sketches. Dans les trois cas (normal, congestion, attaque), on montre en haut les séries agrégées à 1s, figurées cumulativement par sortie de sketch. En dessous sont reportés les diagrammes log-échelle du trafic global (ligne noire épaisse, ◇), des sketches (lignes fines) et de la médiane sur les sketches (ligne bleue épaisse, o).

3.4.3 Analyse longitudinale robuste du trafic internet

Il est possible d'exploiter cet outil sur les très riches données MAWI [MAW] dont on dispose sur une longue période, afin de caractériser sur un temps long l'évolution des propriétés statistiques et de la composition du trafic. Les questions sont celles de l'évolution du trafic Internet lors de ces dernières années : évolutions des attributs de paquets ? Tendances dans le P2P ? Trafic plus ou moins gaussien ? Quelle évolution pour la longue mémoire du trafic ? L'estimation robuste fournie par les outils à *sketch* nous permet de suivre les caractéristiques du trafic en s'affranchissant des anomalies, des accidents, des non-stationnarités. Assez peu d'articles de l'époque ont pu faire une même étude. Certains se focalisaient sur la dernière application en vogue, web [CB97], P2P [KBB+04, AC07], video streaming [CKR+07], ou des anomalies comme des scans [APT07]. Quelques rares travaux précurseurs apportaient des éléments sur le passé du télétrafic [cPB94, FKMc04] sans avoir eu de remise à jour liées aux augmentations des capacités des liens Internet ou à l'apparition incessante de nouvelles applications.

Nous avons réalisé une étude longitudinale poussée du trafic capturé sur un lien pendant 7 ans consécutifs de 2001 à début 2008 [P29]. La figure 3.11 résume le contenu du trafic (en haut) et l'évolution du paramètre H de Hurst de longue mémoire du trafic (en bas) : cette figure prouve que la longue mémoire est plutôt stable au cours du temps, entre 0.8 et 1, et ne montre en tout cas que peu d'indication de ce qu'elle serait en train de disparaitre. Les seuls endroits où H varie beaucoup sont en présence d'anomalies très fortes car parfois certaines dominent très largement le trafic (nous discuterons plus les anomalies de trafic en 3.5) ou quand la capacité du lien change (au moment du trou dans les données avant 2007), H passe de environ 0.8 à un peu plus de 0.9, ce qui reste de la longue mémoire – plus

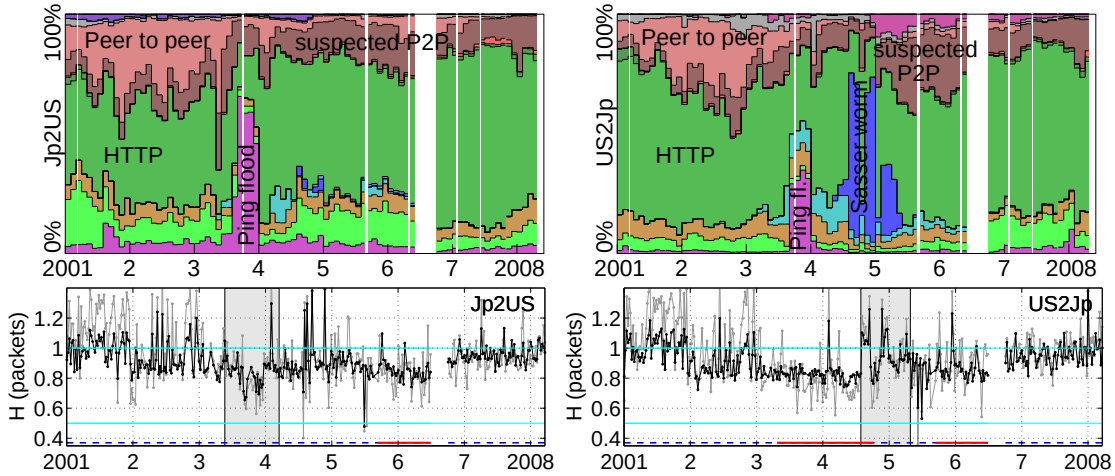


FIGURE 3.11 – Caractérisation du télétrafic transpacifique WIDE, (Japan vers US) de 2001 à 2008 : classement par protocole à gauche ; stabilité du paramètre H de longue mémoire à droite (grâce à la méthode d'estimation robuste à sketch) à droite. En grisé : à gauche, zone où l'anomalie de Ping flooding domine ; à droite, zone grisée où le trafic du ver Sasser domine.

forte même. Un deuxième résultat obtenu par estimation par sketch médian, en figure 3.12, est que le paramètre de Hurst estimé avec sketch pour la série agrégée des paquets X_Δ et celle en octets W_Δ est bien plus souvent le même (sauf lors d'anomalies) que si on faisant des estimations sur le trafic global.

Il nous semble donc que l'on peut répondre aux travaux prédisant la disparition de cette propriété de longue mémoire et la gaussianisation du trafic [CCLS02, KMFB04] en affirmant qu'elle n'a pas lieu. Depuis 2008, nos estimations sur les mêmes traces MAWI n'ont pas changé. Nous continuons l'étude en se penchant cette fois sur les comportements aux temps courts dans un article en préparation.

3.5 Détecter des anomalies

3.5.1 Détection statistique d'anomalie de trafic Internet

Les travaux de détection d'anomalies de trafic sur des bases statiques ont débutées avec [Bru00, BKPR02] et ont été un domaine très actif dans les années suivantes (par exemple [HHP03, LCD04, LBC+06], Workshop LSAD à Sigcomm en 2006 et 2007,...). La combinaison des outils développés ci-dessus, un modèle Gamma-fARIMA adapté de façon pragmatique à du trafic normal et une méthode à sketch pour de l'estimation robuste de moyenne ou de référence, nous permettait de facilement formuler une méthode statistique de détection d'anomalie, ce qui fut fait dans [P15, P17, P21]. Pour évaluer les performances statistiques de ces procédures, il fallait disposer d'une bibliothèque documentée d'anomalies. Celle-ci n'existant pas, nous avons donc décidé de produire nous-mêmes un ensemble documenté, commenté et reproductible d'anomalies en produisant des flash crowds et divers types d'attaques de déni de service – une base de données de traces a été constituée à cette fin dans METROSEC, en procédant à des expériences calibrées et reproductibles sur Renater, voir [P14, P16, J6, J7].

Le premier point vient des études préliminaires dans [J6, J7, P10, P11, P12, P13, P14]

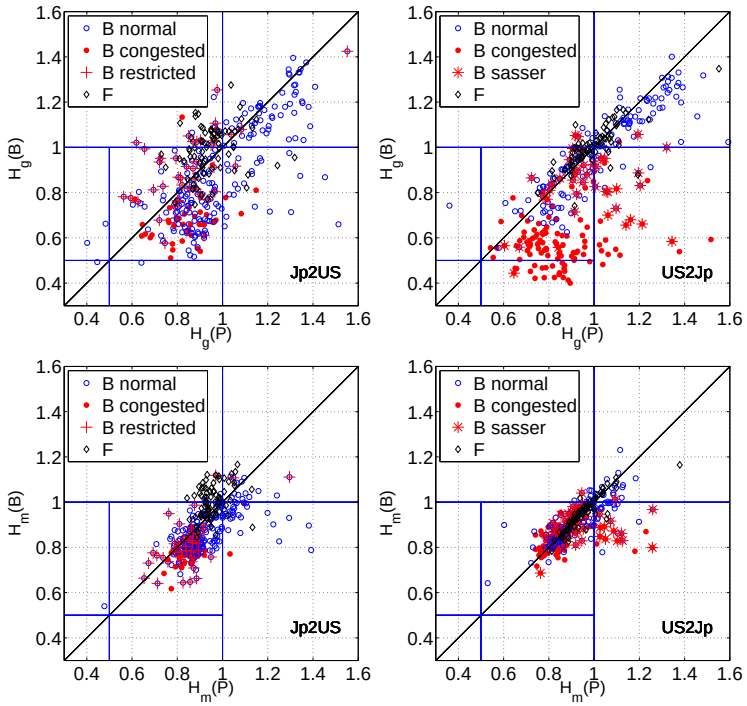


FIGURE 3.12 – Graphes de $H(B)$ (séries en octets) vs. $H(P)$ (séries en paquets). En haut, estimées sur le trafic global. En bas, estimées sur la médiane des sketches. Les symboles sont : **o** : point de mesure **B** sans congestion ; • : point de mesure **B** avec congestion ; + : point de mesure **B** avec anomalie (US2Jp) ou trafic limité (Jp2US) ; ◇ : point de mesure **F**. Gauche : Japon vers US ; droite : US vers Japon.

sur le modèle Gamma-fARIMA, en particulier pour les traces de trafic Metrosec. Nous avons montré que les évolutions des α_Δ et β_Δ avec des Δ assez courts (typiquement pris de 1 ms à 1s) rendent possibles la distinction entre trafic normal, anomalies légitimes (par exemple des flash crowds) et illégitimes (par exemple des attaques par dénis de service). La figure 3.13 montre par exemple l'estimation des paramètres α_Δ de 3 sous-traces après hachage dans une table de taille $M = 32$, chaque estimation étant faite sur une minute. Dans la trace mesurée, on avait généré une anomalie DDoS par UDP flooding contre un ordinateur et on hache sur l'adresse IP destination ; l'adresse cible se trouve dans la sortie $m = 20$ du sketch. Les courbes en rouge sont celles pendant la durée de l'attaque. Visuellement, on voit que ces courbes α_Δ pendant l'attaque sont différentes en $m = 20$ de celles aux autres moments ou pour d'autres sorties du sketch.

À partir des modélisation mises en œuvre sur des fenêtres temporelles successives de trafic, on cherche à détecter des ruptures dans les statistiques du trafic afin de détecter l'occurrence d'anomalies. Le principe de détection retenu exploite la modélisation multirésolution : les paramètres du modèle en loi $\Gamma_{\alpha_\Delta,\beta_\Delta}$ sont estimés successivement pour plusieurs niveaux d'agrégation Δ (de 1ms à 1s). C'est le changement dans la façon dont ces paramètres estimés varient avec le niveau d'agrégation qui est recherché. La détection est réalisée à partir du calcul de distances entre les statistiques estimées dans la fenêtre courante et celles obtenues dans une fenêtre de référence. Les détails sont dans [P17].

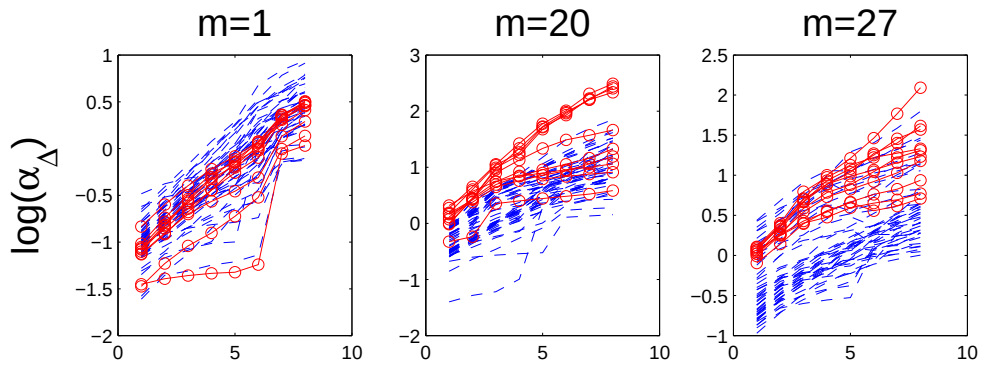


FIGURE 3.13 – Paramètres multi-échelle des lois Gamma. Évolution des $\alpha_\Delta^{n,m}$ en fonction de Δ pour 3 sorties de sketch. Les fenêtres temporelles contenant les attaques sont en rouge (et symboles 'o'). L'attaque (localisée dans la sortie $m = 20$ du sketch) est visible par un changement significatif de la forme de $\alpha_\Delta^{n,m}$.

3.5.2 Détection et identification des anomalies par LD-sketch

Nous avons finalisé un détecteur d'anomalies statistiques dans le trafic à partir des ruptures dans les statistiques α_Δ et β_Δ du trafic mesurées à différents échelles Δ à travers la technique des sketchs, appelé détection par *sketch multi-résolution*. La figure 3.14 esquisse le principe de l'algorithme de détection et ses étapes pour chaque fenêtre temporelle d'analyse, typiquement 1 minute : 1) les sketches forment $M \times N$ sous traces ; 2) agrégation multi-résolution à plusieurs Δ ; 3) estimation du modèle Gamma ; 4) estimation de la situation de référence par la moyenne et la variance des α_Δ et β_Δ sur les M sous-traces pour un n ; 5) calcul par une distance d'un contraste entre les statistiques de chaque sortie de sketch et sa référence 6) alerte d'anomalie pour un sketch (n, m) si la distance dépasse un seuil qui contrôle la proportion de fausses alarmes tolérées. On obtient ainsi, si I anomalies ont lieu dans la fenêtre de temps t, les classes $m_i^n(t)$, avec $i = 1, ...I < M$ pour $n \in \{1, ...N\}$, où une anomalie est présente.

Ensuite, puisque l'on utilise N fonctions de hachage indépendantes h_n, on peut remonter aux ordinateurs impliqués dans les alarmes en cherchant dans les attributs employés (par exemple IP dst) pour clef de hachages, quelles occurrences conduisent systématiquement à des sorties de sketch en alerte quel que soit n. Ceci est réalisé par recherche exhaustive parmi les attributs observés, ce qui est réalisable puisque le hachage direct est très peu coûteux. On peut montrer que le nombre de collisions [1] attendues est proportionnel à M^{-2N} [TZ04] [P17]. Il devient inférieur à 1 dès que $N \geq 6$ pour $M = 32$ pour un espace de départ de 2^{32} adresses IP (cas de l'IPv4). Avec $N = 8$, on peut ainsi identifier les attributs responsables de jusqu'à I anomalies qui arrivent dans la même fenêtre de temps t tant que I est plus petit que M, taille des sketches.

La méthode de détection a été présentée dans [P17] (et déclinée dans [P21] sur des traces de flots plutôt que paquets) et validée sur la base de traces MAWI avec les collègues japonais du NII et de IIJ qui ont collecté ces traces. Les traces quotidiennes fournies, sur plusieurs points de mesure, sont d'une durée de 15 minutes. Comme la méthode de détection repose sur des fenêtres d'une minute environ, la longueur de la trace est suffisante pour qu'elle

1. Une collision se produit quand 2 attributs différents sont envoyés dans les mêmes sorties pour N fonctions de hachage différentes [TZ04].

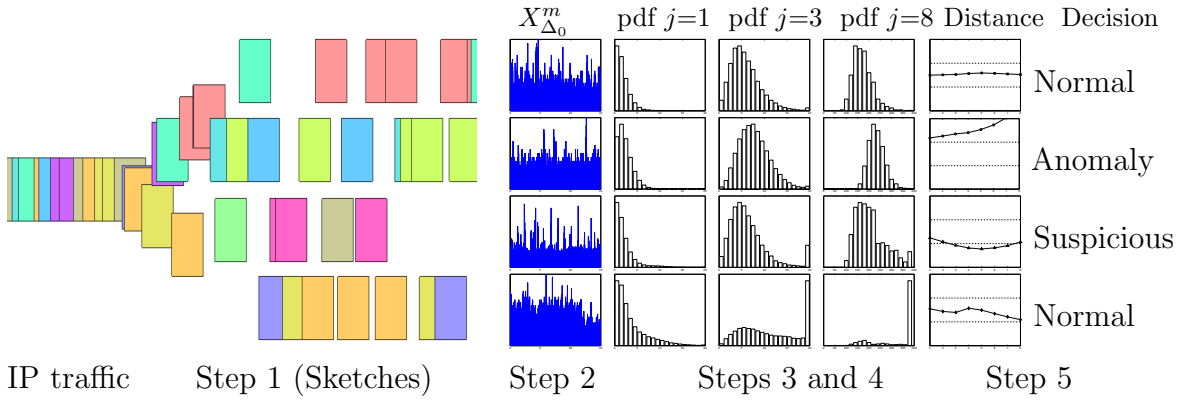


FIGURE 3.14 – Esquisse de la méthode de détection d'anomalie par *sketch multi-résolution*. Elle illustre les étapes de la procédure de détection, partant d'une trace IP de trafic sur laquelle s'applique un sketch pour produire des sous-traces, 4 ici (étape 1). En étape 2, on forme les séries agrégées puis on estime les lois marginales pour différentes échelles (1, 3, et 8 ici) (étapes 3 et 4). La référence est prise comme la moyenne sur les sorties du sketch. Étape 5 : calcul de la distance à la référence pour chaque sortie de sketch, ici représentée en fonction de l'échelle pour décider si il y a une anomalie ou non. Une dernière étape, non représentée, combine les N sketches pour identifier les anomalies.

récupère à la fois des anomalies courtes et des anomalies longues (toute la durée de la trace). On dispose donc dans le trafic MAWI d'une grande variété de trafic et d'anomalies que nous pouvons exploiter par nos méthodes. La figure 3.15 illustre la détection d'anomalie de type *spoofed flooding* de basse intensité, ne correspondant qu'à environ 1% du trafic total au même moment, et étant très loin des flots "éléphants" [PTB+02] les plus massifs dans cette trace. Dans certaines sorties de sketch, d'autres éléments du trafic lèvent des alertes (ici une anomalie DNS et un transfert SSH) mais pour ce dernier qui n'a rien d'anomal, on voit ici que pour une autre sortie de sketch il n'y a pas d'alerte. L'article [P17] détaille d'autres anomalies trouvées et, dans le cadre du projet MAWIlab décrit ci-dessous, toutes les anomalies repérées dans la base MAWI ont donné lieu à une annotation spécifique de la trace. Cela permet à d'autres chercheurs en métrologie des réseaux d'ordinateurs de comparer leurs études à nos résultats.

3.5.3 Analyse longitudinale des anomalies dans le trafic

À l'aide de la procédure de détection des anomalies, on a pu compléter l'analyse longitudinale du trafic MAWI dans [P29] par une étude des anomalies présentes sur les 7 années de la base. La figure 3.11 incorpore des aspects des résultats obtenus : on voit par exemple que vers fin 2003, des attaques de Ping flooding ont régulièrement occupé jusqu'à la moitié de la bande passante. Certaines anomalies durent pendant des semaines voire des mois et sont donc visibles chaque jour pendant ce temps ; d'autres ont une durée très courte (de quelques secondes à quelques minutes) – cela justifie pleinement le besoin d'une méthode multirésolution telle que [P17]. Une anomalie importante est par exemple l'activité du ver informatique Sasser dans le trafic de mai 2005 à mai 2005 principalement visible dans le trafic de US vers Japon. On remarque plusieurs bouffées d'activité (2004/08, 2004/12 et 2005/03) où Sasser fut sur le point de disparaître deux fois mais est revenu en tant que

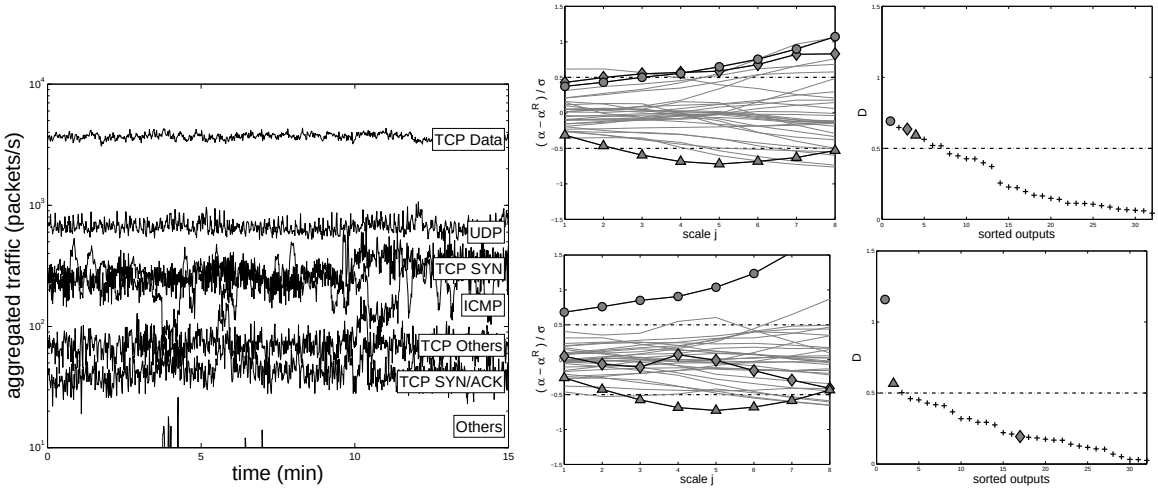


FIGURE 3.15 – Exemple d'anomalie (trafic MAWI, Japon vers USA, 11/10/2005). Le trafic agrégé à 1s est tracé, décomposé par protocole. La deuxième colonne montre les courbes de distance à la référence $(\alpha_{\Delta_j}^{n,m} - \alpha_{\Delta_j}^{m,R})/\sigma_{m,\alpha,\Delta_j}$ en fonction de l'échelle j pour les sorties de deux sketchs différents. La plupart des courbes sont dans la bandes de comportement normal $\pm\lambda$ pour tous les j. À droite : les $D_{\alpha^{n,m},m\in 1,\ldots,M}$ trié en ordre décroissant. On remarque quelques sorties en dehors de $\pm\lambda$. En combinant les sketches, une alarme est retenue seulement si des IPdst conduisent à des sorties dont la distance est toujours en dehors de cette bande pour toutes les fonctions de hachage employées. Ici, les cercles marquent une attaque de *mixed flooding*, les triangles une anomalie DNS et les diamants un transfert par SSH. Seuls les 2 premiers sont des anomalies réelles.

variante de ce ver. Le trafic Sasser représente la moitié du trafic à ce moments. Il est très peu vu du Japon vers les US car les ordinateurs japonais utilisant ce lien appartiennent à des organismes de recherche ou universitaires, vraisemblablement mieux protégés que les ordinateurs du grand public.

L'évolution du nombre d'anomalies par trace de 15 minutes, repérées par la méthode et classées de gravité forte à faible, est donnée en figure 3.16. On se rend compte qu'il n'y a jamais de trace de 15 minutes sans au moins une anomalie réelle de type scan, flooding ou spoofed IP. Cela nous renseigne sur la difficulté de trouver une situation de trafic normal pour étudier les modèles de trafic IP. Cette analyse de trafic réel nous conforte d'ailleurs dans l'intérêt des méthodes développées pour l'analyse robuste du trafic ou la détection d'anomalies. Sans de telles méthodes (ou des méthodes visant à la même chose), les propriétés mesurées dans le trafic que l'on trouve dans beaucoup d'études scientifiques des réseaux d'ordinateur risquent fortement de se révéler être des artefacts ou des caractéristiques spécifiques d'anomalies.

3.5.4 Évaluation des détecteurs d'anomalie de télétrafic Internet

Un dernier problème à propos de la détection d'anomalie dans le trafic sur Internet est la quasi inexistence de vérité terrain pour décider si la détection est bonne ou non. On ne disposait pas de bonne méthode de *benchmarking*, ni même d'approche unifiée pour comparer des détecteurs d'anomalie, eux proposés en nombre croissant dans la littérature au fil des années.

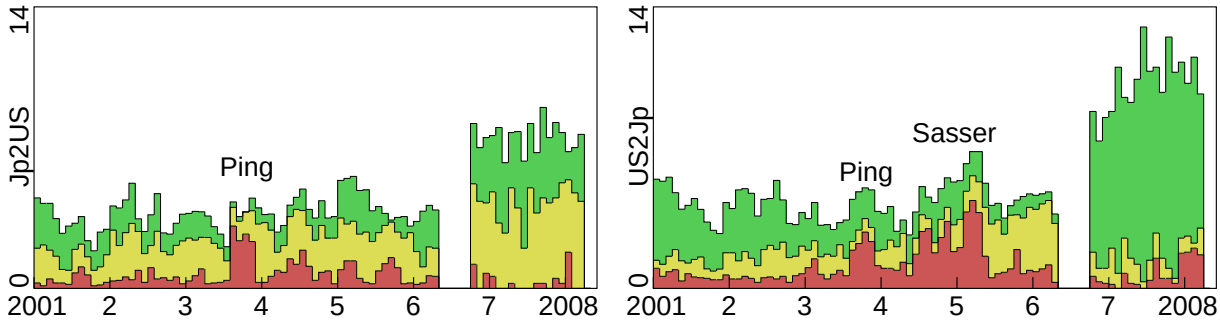


FIGURE 3.16 – Analyse longitudinale des anomalies. Anomalies "suspectées" (en vert) : WWW, P2P, GRE, DNS. Anomalies vraisemblablement des attaques (en jaune) : mécanismes divers. Attaques certaines (en rouge) : Ping/SYN floods, spoofed,...

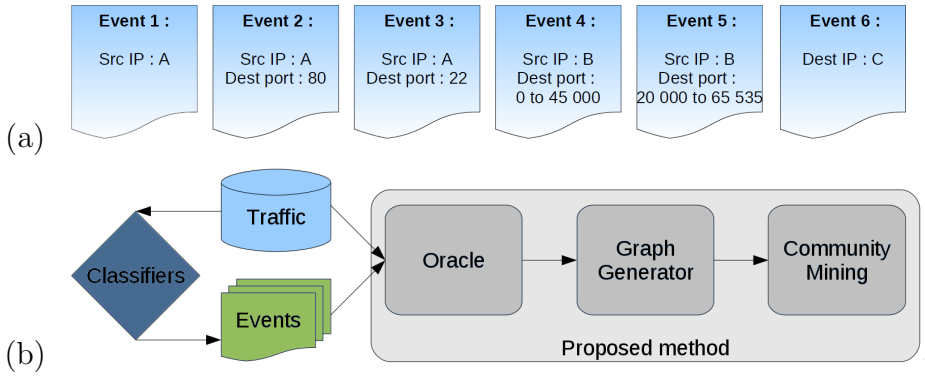


FIGURE 3.17 – Schéma d'évaluation des détecteurs d'anomalie.

Nous avons considéré cette question en s'aidant de ce qu'un détecteur est toujours une procédure qui lit une trace de trafic et renvoie en sortie une liste d'événements qu'elle considère être des anomalies ; on se retrouve alors avec des logs d'anomalies possibles comme en figure 3.17. Ces événements spécifient des caractéristiques des anomalies (IP src, IP dst, ports, etc.) mais pas toujours les mêmes. L'idée mise en avant dans le cadre de la thèse de Romain Fontugne au NII (Tokyo) de 2008 à 2011 et publiée dans [P38, P43, P44] fut d'employer une représentation sous forme de graphe des résultats des différents détecteurs et de leurs similarités. La procédure est esquissée en figure 3.17. Un oracle relit la trace et les résultats des détecteurs d'anomalies analysés en associant à chaque paquet IP les alarmes correspondantes de chaque détecteur ; puis on génère un graphe où chaque somme est une alarme d'un détecteur et chaque arête représente le fait qu'un paquet est associé aux 2 alarmes connectées. On pondère ensuite ce graphe par un indice de Simpson

$$S(E_1, E_2) = |E_1 \cap E_2| / \min(|E_1|, |E_2|) \tag{3.13}$$

où E_i est le trafic associé à l'alarme i. Cette métrique est dans $[0, 1]$ et 0 indique que les alarmes n'ont pas d'intersection de trafic (pas d'arête donc) et sont dissemblables tandis que 1 indique que les trafics sous-jacents sont identiques (noter qu'on pourrait prendre une autre métrique de comparaison d'ensemble, telle que l'indice de Jaccard ; dans [P44], on a

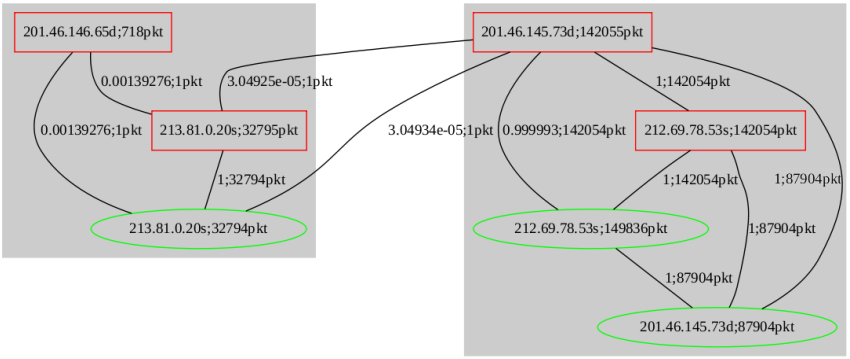


FIGURE 3.18 – Exemple d'anomalie scindée en 2 groupes. Les alertes du détecteur [P17] sont en rouge, celles de [FF11] en vert. A gauche, communauté du scan de réseau (sur le port 3128 d'un serveur proxy) ; à droite, communauté d'un trafic nntp.

préféré prendre celui de Simpson qui a donné de bons résultats).

La dernière étape est de rechercher des groupes d'anomalies qui seraient suffisamment semblables, c'est-à-dire connectées entre elles, pour être considérées comme les mêmes. Ce problème est celui de la recherche de communauté dans des graphes [For10] (que nous retrouverons et discuterons plus en détail dans le chapitre 4). Mentionnons ici seulement que l'on s'est tourné vers les techniques de modularité qui offraient le meilleur compromis entre efficacité et temps de calcul dans notre situation.

Un résultat est présenté en figure 3.18 où l'on voit qu'un ensemble d'alertes pour 2 détecteurs (de [P17] et de [FF11]) qu'on aurait naïvement associés à la même cause sans l'étape de détection de communauté (car les alertes sont liées), sont en fait groupées dans 2 communautés distinctes, l'une étant du scan de réseau (sur le port 3128 d'un serveur proxy) et l'autre du trafic nntp. Ils sont reliés car un paquet du serveur victime du scan va vers le port 3128 des ordinateurs impliqués dans le trafic nntp, sans relation réelle entre les anomalies. L'intérêt de la méthode est de comparer les résultats des détecteurs, d'abord pour les valider en l'absence de vérité terrain ou pour comprendre par exemple quels détecteurs manquent systématiquement quels types d'anomalies, ou pour pouvoir fusionner les résultats des détecteurs pour améliorer les performances générales de détection. Nous avons exploré dans [P43] plusieurs cas comme celui-ci tandis que [P44] finalise la méthode et l'applique à large échelle pour l'annotation de la base de données de trafic internet MAWI. Les annotations mises à disposition de la communauté sur le site web du MAWIlab [2]. Cette base de donnée est à ce jour la plus complète pour trouver du trafic IP annoté quant aux anomalies existantes, en accès libre.

3.6 Classifier les ordinateurs par leur trafic

3.6.1 Classification non supervisée de trafic

Arrivant bientôt au moment du bilan sur ces travaux consacrés à l'étude du télétrafic des réseaux d'ordinateurs, il nous reste à décrire un travail portant sur la caractérisation

2. `www.fukuda-lab.org/mawilab`

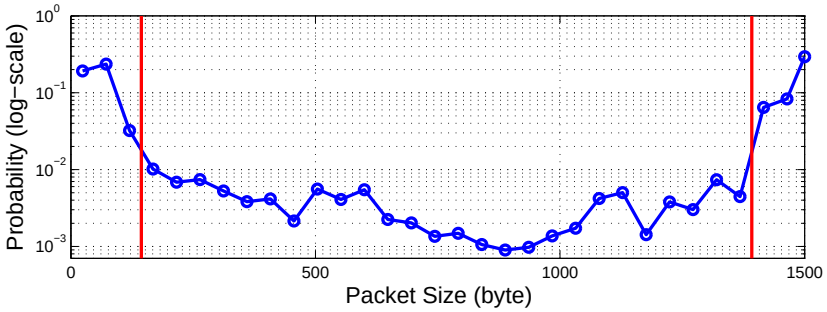


FIGURE 3.19 – Histogramme des tailles des paquets dans le trafic, en échelle log avec des bins de 0 à 1500 octets par pas de 48 octets. Les lignes verticales (rouges) sont les limites employées pour les catégories des petits et grands paquets. Ils correspondent respectivement à 46% et 44% des paquets. Entre les deux, la population des paquets plus grands que 144 octets et plus petits que 1392 octets est beaucoup plus faible et en générale sans spécificités (sauf protocoles particuliers).

des ordinateurs connectés à l'Internet par classification non supervisée de leur trafic [J17, J23]. De telles études sont importantes en administration des réseaux car elles permettent d'identifier des ordinateurs qui diffusent ou sont victimes de virus ou autres attaques, de trouver les traces d'applications particulières (P2P, voie sur IP,...) et plus généralement de suivre les évolutions et nouvelles tendances dans l'emploi des réseaux. Les approches classiques reposent sur l'inspection du contenu du trafic [Roe99, l7-, ope] ou s'appuient sur des ensembles de règles fondées sur les ports employés [Cor]. Des travaux récents ont ajouté à cela des règles heuristiques [KPF05], des approches statistiques de la classification [TPGK03, LCD05, MMN08], l'interrogation des bases de données Google [TRKN08], ou des codages sous forme de graphe macroscopique du trafic [XZB05, JSZ09, IFM09, IGER$^+$10]. La littérature sur la classification de trafic par analyse de profils statistiques est assez large. On y trouve des méthode supervisées, telles que de l'apprentissage par plus proches voisins [ERS$^+$10, LKJ$^+$10], des techniques bayésiennes [MZ05, PCUKEN09, LKJ$^+$10], des arbres de décision [PCUKEN09, LKJ$^+$10], des SVM [KcF$^+$08, LKJ$^+$10],... Des méthodes non supervisées ont été proposées par exemple à base de K-means [LCD05, BTS06, EAM06] ou de classification hiérarchique [LCD05].

Nos travaux entrent dans le catégorie des approches statistiques non supervisées qui présentent deux intérêts : étant non supervisées, elles détectent sans limite des catégories inconnues de trafic et des nouveautés ; étant basées sur des statistiques de trafic, elles ne sont que faiblement affectées par le cryptage des données dans les paquets et fonctionnent plus facilement car on ne regarde pas tout le flot de paquets en détail (elles fonctionnent par exemple sur des traces enregistrées comme celles de MAWI qui, en général, ne gardent pas le contenu en données des paquets). L'originalité de nos travaux est de s'appuyer sur des codages innovants du trafic, soit par un ensemble bien choisi de caractéristiques de trafic [J17], soit par un ensemble plus grand de grandeurs [J23] qui servent à décrire des graphes de trafic [KPF05, KcF$^+$08])

Dans ce mémoire, j'ai décidé de ne donner qu'un aperçu général du principe de ce travail et assez peu des résultats car ils nécessitent d'aller assez loin dans les connaissances en ingénierie de trafic ou sur les problèmes rencontrés en pratique pour la classification des ordinateurs sur Internet. En particulier, avec peu de vérité terrain, il est compliqué

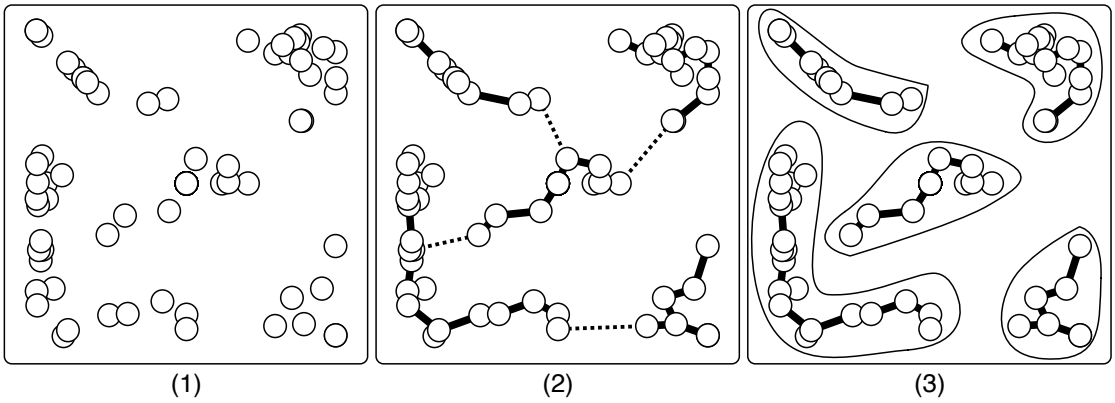


FIGURE 3.20 – Procédure de classification à l'aide d'un arbre minimal recouvrant. (1) On figure un ensemble d'ordinateurs dans l'espace de leurs caractéristiques, ici réduit à 2D. (2) L'arbre minimal recouvrant est tracé avec les arêtes les plus longues en pointillé. (3) Procédure où l'on coupe les arêtes les plus longues qui conduit à des clusters.

de présenter les résultats – faits sur des traces réelles, ils s'appuient sur la comparaison des classifications obtenues à des méthodes qui font partie de l'état de l'art dans ce domaine (classification du trafic selon les numéros de port [KcF+08] ou classification selon les caractéristiques au niveau transport [KPF05, KcF+08]). La discussion se limitera aux idées concernant les méthodes. Le lecteur désirant en apprendre plus, en particulier sur les résultats sur du trafic réel, est renvoyé vers [J17, J23], le 2e proposant des améliorations sensibles par rapport au premier.

3.6.2 Classification avec peu de caractéristiques

Dans un premier article [J17], les caractéristiques utilisées s'inspirent de métriques connues en réseau d'ordinateurs, mais certaines sont nouvelles, en particulier certaines liées aux répartitions dans l'espace des adresses IP, mesurées par une entropie. Pour chaque IP source, on calcule les métriques qui suivent, groupées en trois ensemble, à partir d'une trace IP :

- ***I. Connectivité dans le réseau Internet***
  - **i)** le nombre de pairs avec qui l'IP est en communication (i.e., IP destination)
  - **ii)** le nombre de ports source divisé par le nombre de pairs (IP dst)
  - **iii)** le nombre de porte de destination divisé par le nombre de pairs (IP dst)
- ***II. Dispersion des connexions dans le réseau***
  - **iv)** le ratio de l'entropie du second octet des IP dst divisé par l'entropie du quatrième octet de l'IPdst, où l'on rappelle que l'entropie est $S = -\sum_i p_i \log p_i$
  - **v)** le ratio de l'entropie du troisième octet des IP dst divisé par l'entropie du quatrième octet de l'IPdst
- ***III. Contenu du trafic de l'ordinateur***
  - **vi)** le nombre moyen de paquets par flot
  - **vii)** le pourcentage de paquets de petite taille ($\leq$ 144 octets)
  - **viii)** le pourcentage de paquets de grande taille ($\geq$ 1392 octets)
  - **ix)** l'entropie de la distribution des paquets de taille moyenne.

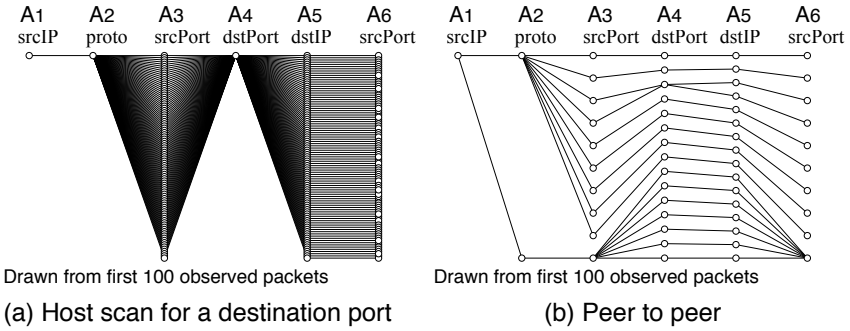


FIGURE 3.21 – Exemples de graphlets. Le trafic d'une seule IP source est tracé sous forme de graphe connectant les attributs (protocole, ports source ou destination, IP destination) de tous les paquets émis par cette source.

L'ensemble de ces caractéristiques a été choisi en fonction de métriques habituelles en réseau (par exemple **i** est souvent employé) mais suit aussi un compromis entre leurs pertinence et la parcimonie de la description, souhaitable pour éviter d'augmenter trop la taille de l'espace de description. Le choix de ces caractéristiques est justifié plus en détail dans [J17]. Sur la figure 3.19 on verra par exemple la distribution empirique des tailles de paquets pour du trafic MAWI qui justifie les choix **vii**, **viii**, **ix** : il y a peu de paquets de taille intermédiaire, tandis qu'un excès de petits paquets va facilement suggérer des scans tandis qu'un excès de grands paquets indique souvent du transfert de données. Notons qu'on peut accélérer le traitement des traces IP pour calculer ces caractéristiques en employant les techniques à base de sketch, en particulier pour estimer le nombre de pairs d'une IP donnée.

Ensuite, nous avons proposé une méthode de classification à partir d'un arbre minimal recouvrant (*Minimum Spanning Tree*, MST), proche d'autres méthodes comme [GCHM06, GMC+09]. Les étapes de la procédures de classification, une fois les 9 caractéristiques estimées pour chaque adresse IP, sont :

— Calcul du MST connectant tous les nœuds que sont les adresses IP dans l'espace à 9 dimensions des caractéristiques retenues, après normalisation des caractéristiques F_n en $f_n = (2/\pi)\arctan(F_n/R_n)$ pour un R_n bien choisi.
— Première classification en coupant les plus grands liens du MST pour ne garder que des cœurs de classes.
— Identification de groupes denses (sous-arbres du MST avec au moins 10 nœuds avec des liens courts) pour refaire croitre les classes à partir de ces groupes.

Les deux premières étapes sont illustrées en figure 3.6.1. Il est très utile de partir des groupes denses car la classication à MST est connue pour être sensible au bruit ou valeurs aberrantes si seules les deux premiers points sont employés.

Les résultats de la classification par cette méthode sont détailles dans [J17] et ils sont cohérents avec des approches supervisées, utilisant des connaissances sur les numéros de port en particulier. Il reste que les classes identifiées doivent être examinées à la main, en se plongeant dans les traces, pour trouver quelle classe décrit quel type de trafic et leur attribuer des labels.

A_i	i-th column (or attribute) of graphlets (from left to right)
$i:j$	Direction from A_i to A_j ($j = i \pm 1$)
n_i	# Nodes in A_i
$o_{i:j}$	# Nodes having degree 1 in direction $i:j$.
$mu_{i:j}$	Average degree in direction $i:j$.
$\alpha_{i:j}$	Maximum degree in direction $i:j$.
$\beta_{i:i+1}$	$= d_{k,i:i-1}$, where $k = \arg\max_l \{d_{l,i:i+1}\}$
$d_{k,i:j}$	In/out-degree of node $v_{k,i}$: in-degree for $i:i-1$ (left half of $v_{k,i}$) and out-degree for $i:i+1$ (right half of $v_{k,i}$)

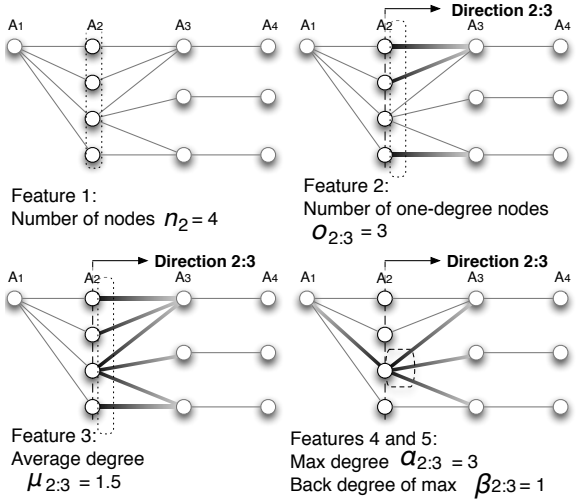


FIGURE 3.22 – Descripteurs statistiques du trafic basés sur les graphlets. À gauche : tableau des notations et des 44 descripteurs retenus. Notez qu'un nœud a ici deux degrés, un dans chaque direction (par exemple, pour la colonne A_2, on sépare le degré 2 : 1 et 2 : 3). À droite : explication graphique des descripteurs.

3.6.3 Interpréter les classes dans la classification non supervisée de trafic

L'amélioration principale que l'on trouve dans [J23] a été de modifier les descripteurs statistiques, en augmentant leur nombre, de manière à ce que ces descripteurs correspondent aux caractéristiques des *graphlets* de chaque ordinateur. Les *graphlets* sont des représentation graphiques du trafic d'un ordinateur proposées pour la méthode BLINC dans [KPF05, KcF$^+$08] dont on trouvera un exemple en figure 3.21. On met en colonne les attributs IP en liste (IP srdc, Protocal, Port src, Prt dst, IP dst, Port src) ; chaque nœud dans une colonne est une valeur possible de l'attribut et on trace une arête entre nœuds de colonnes adjacentes si et seulement si un paquet a ces 2 attributs. Dans [KcF$^+$08], il est montré que les *graphlets* sont une très bonne représentation du trafic et permettent de trouver de bonnes approches de classification par règles du trafic.

Le problème est que les approches utilisant des *graphlets* étaient par construction supervisées et ne s'adaptent pas à l'occurrence de nouveaux types de trafic qui demanderaient de nouvelles règles, ou de reconsidérer les anciennes. Le résultat principal de [J23] est de combiner le meilleur de deux mondes : la classification non supervisée appliquée à des descripteurs statistiques bien choisis du trafic qui nous permet à la fois une bonne classification et la détection de nouveaux comportements ; la classification supervisée qui donne par construction un nom, une identité, aux classes recherchées.

Dans ce travail, la classification est faite sans supervision, en utilisant un algorithme simple de classification hiérarchique avec la méthode de Ward, comme dans [LCD05], qui a l'avantage d'être bien établie et donne de bons résultats. On re-génère ensuite un *graphlet* synoptique pour chaque classe non supervisée, en partant de son centroide. Cela est possible car nous partons de suffisamment de descripteurs statistiques bien choisis. La figure 3.22 indique quels sont ces descripteurs : ce sont les 44 valeurs qui donnent, pour chaque colonne (quand c'est applicable) le nombre de nœuds, les degrés moyens, maximum, et le nombre de

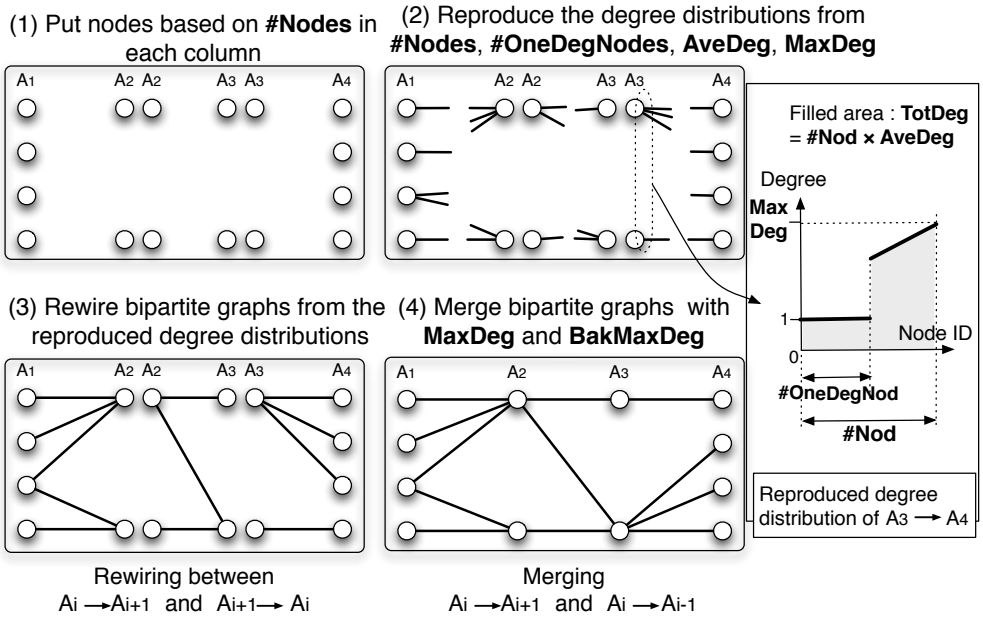


FIGURE 3.23 – Procédure pour re-générer un graphlet synoptique pour chaque centroïde de cluster, de manière à pouvoir visualiser simplement le type de trafic dans le cluster.

nœuds de degrés 1 dans chaque direction et le degré dans la direction inverse du nœud de degré maximum dans une direction. Ces descripteurs sont réellement choisis afin de pouvoir revisualiser facilement un *graphlet* proche à partir de leurs valeurs. La figure 3.23 résume la procédure pour créer un *graphlet* à partir des 44 caractéristiques. Un point important est que l'on fait une approximation pour la distribution des degrés : à l'exception des nœuds de degré 1, les autres ont une distribution que l'on approxime par une loi uniforme entre le degré 2 et le degré maximum pour re-générer le graphlet.

Faisant tout cela, nous disposons d'une méthode de classification de trafic non supervisée, pour laquelle il est possible de visualiser un *graphlet* synoptique qui résume le comportement de la classe de telle sorte qu'il devient simple à un ingénieur réseau de donner un label à chacune. En fait, la classe peut même être labelisée par exemple par les règles heuristiques déduites de BLINC [KPF05] ou Reverse-BLINC [KcF+08]. Néanmoins, notre procédure détecte de nouvelles classes, qui nécessiteront elles l'intervention d'un administrateur réseau, mais n'a pas d'autres besoins de supervision. Des résultats sur des traces IP réelles (MAWI, Keio), comparés à des méthodes état de l'art, sont détaillés et discutés dans l'article paru en 2013 à *IEEE/ACM Trans. on Networking* [J23] ainsi qu'un suivi des graphlets estimés pour chaque classe en fonction du nombre de paquets déjà mesurés, à des fins de détection précoce des classes.

Notons que nous continuons ce type de travaux de classification dans [Js30], cette fois au niveau des applications, en prenant compte l'ordre dans lequel les flots ont lieu et l'existence de causalités entre différents flots. Cela permet une très bonne identification du trafic P2P par exemple.

3.7 Bilan et perspectives

Les travaux que nous avons menés pour étudier le trafic Internet dans une approche de modélisation statistique basée sur la métrologie des réseaux, ont permis de proposer plusieurs méthodes répondant au départ à des motivations très pratiques : méthode de détection d'anomalies (attaques sur le réseau ou comportements inattendus des ordinateurs ou des humains derrière, comme lors des flash crowds) par profil statistique ; méthode d'analyse robuste de trafic ; annotation systématique de la base de données MAWI de trafic IP ; classification des ordinateurs en fonction de leur activité.

Plus généralement, nos travaux ont montré l'intérêt des approches fondées sur le traitement statistique des signaux pour des questions concrètes de métrologie des réseaux informatiques et l'Internet. Au-delà des applications pratiques, ce cheminement a été l'occasion de retravailler les modèles statistiques proposés pour le télétrafic, comme le modèle théorème de Taqqu, et de reprendre des questions telles que la dépendance à longue portée dans ces signaux.

Au passage, nous aurons esquissé une première utilisation des techniques moderne d'analyse de graphes complexes, celles concernant les communautés, dans un cadre original puisqu'il s'agissait de comparer les résultats de détecteurs d'anomalies de trafic. Nous allons dans le suite parler plus longuement de travaux s'inscrivant pleinement dans l'analyse de graphes complexes, puisqu'il s'agit du dernier (et du plus récent) thème sur lequel je travaille. L'étude du trafic IP continue cependant un peu, avec par exemple une nouvelle approche pour la classification de trafic dans [Js30] qui tente cette fois d'intégrer l'aspect temporel et causal de l'ordre dans lequel les flots arrivent pour un même ordinateur (ou plutôt une même IP).

Perspectives. Ces travaux sur les signaux associés aux réseaux d'ordinateurs dégagent néanmoins deux perspectives vers lesquelles nous souhaitons continuer à avancer.

La première a déjà été évoquée, il s'agit de faire pour les temps courts la même étude longitudinale permise par la combinaison des techniques de sketches et des outils à ondelettes. Pour les temps longs, nous pensons avoir montré que le trafic présente de la longue mémoire pilotée par le mécanisme de Taqqu. Pour les temps caractéristiques plus courts, la dossier reste ouvert et la question est en particulier celle l'existence d'une multifractalité [TTW97, FGW98] dans le trafic, ou son absence [HVA05]... Les avancées faites pour les estimations liées au formalisme multifractal [WAJ07] vont nous permettre cette étude.

Une deuxième suite des travaux sera de remettre un peu plus de poids aux étiquettes IP des paquets. On l'a dit, les données brutes ne sont pas des signaux mais une suite de paquets échangés de A à B, à un certain instant. Nous sommes ici passés à des séries agrégées (éventuellement en sous-séries avec les sketches) là où, dans d'autres contextes (par exemple au chapitre 4), on coderait cela comme la trace d'un lien de A vers B dans un réseau. Il y a sûrement mieux à faire dans les deux cas, en s'intéressant directement à une structure de flots de liens et aux traitements qui leurs sont adaptés. Cette idée étant portée par des collègues du LIP6 (à Paris), nous envisagons une collaboration étroite avec eux et avec l'équipe DANTE du LIP (à Lyon) sur ce sujet.

Travaux liés au chapitre 3

Journaux à comité de lecture

[Js30] H. Asai, K. Fukuda, P. Abry, P. Borgnat, H. Esaki, "Network Application Profiling with Traffic Causality Graphs", submitted 11/2013.

[**J23**] Y. Himura, K. Fukuda, K. Cho, P. Abry, P. Borgnat, H. Esaki, "Synoptic Graphlet : Bridging the Gap between Supervised and Unsupervised Profiling of Host-level Network Traffic", *IEEE/ACM Transaction on Networking*, Volume 21, Issue 4, pp. 1284-1297, August 2013.

[**J17**] G. Dewaele, Y. Himura, P. Borgnat, K. Fukuda, P. Abry, O. Michel, R. Fontugne, K. Cho, H. Esaki, "Unsupervised host behavior classification from connection patterns", *International Journal of Network Management*, Vol. 20, No 5, pages 317-337, 30 august 2010.

[**J13**] P. Loiseau, P. Gonçalves, G. Dewaele, P. Borgnat, P. Abry, P. Vicat-Blanc Primet "Investigating self-similarity and heavy-tailed distributions on a large scale experimental facility," *IEEE Trans. on Networking*, Vol. 18 :4, p. 1261-1274, August 2010.

[**J12**] P. Abry, P. Borgnat, F. Ricciato, A. Scherrer, D. Veitch, "Revisiting an old friend : On the observability of the relation between Long Range Dependence and Heavy Tail," *Telecommunication Systems*, Volume 43, Issue 3-4, pp 147-165, April 2010.

[**J7**] P. Borgnat, P. Abry, G. Dewaele, A. Scherrer, N. Larrieu, P. Owezarski, Y. Labit, L. Gallon, J. Aussibal, « Caractérisation non gaussienne et à longue mémoire du trafic Internet et de ses anomalies, » *Annales des télécommunications*, vol. 62, n. 11-12, pp. 1401-1428, november-december 2007.

[**J6**] A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat, P. Abry, "Non-Gaussian and Long Memory Statistical Characterisations for Internet Traffic with Anomalies," *IEEE Transaction on Dependable and Secure Computing*, vol. 4, n. 1, pp.56-70, january-march 2007.

Actes publiés dans des colloques avec actes à comité de lecture

[**P44**] R. Fontugne, P. Borgnat, P. Abry, K. Fukuda, "MAWILab : Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking", *ACM CoNEXT 2010*, Philadelphia (PA), Nov. 30-Dec. 3 2010.

[**P43**] R. Fontugne, P. Borgnat, P. Abry, K. Fukuda, "Uncovering Relations Between Traffic Classifiers and Anomaly Detectors via Graph Theory", *COST-TMA (Traffic Measurement & Analysis) Workshop 2010*, Zurich (CH), April 2010.

[**P40**] F. Ricciato, A. Coluccia, A. D'Alconzo, D. Veitch, P. Borgnat, P. Abry, "On the role of flows and sessions in Internet traffic modeling : an explorative toy-model", *IEEE GLOBECOM'09*, December 2009

[**P38**] R. Fontugne, P. Borgnat, P. Abry, K. Fukuda, "Towards Systematic Traffic Annotation", *CoNEXT'09 Student Workshop*, Rome (Italy), December 1, 2009.

[**P29**] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, K. Cho, "Seven Years and One Day : Sketching the Evolution of Internet Traffic," *Proceedings of the 28th IEEE INFOCOM 2009*, pp. 711–719 Rio de Janeiro (Brazil), May 2009.

[**P22**] F. Chatelain, P. Borgnat, J.-Y. Tourneret, P. Abry, "Parameter estimation for sums of correlated gamma random variables. Application to anomaly detection in Internet Traffic," *IEEE Int. Conf. on Acoust., Speech and Signal Proc. ICASSP-08*, Las Vegas (NV), April 2008.

[**P21**] P. Borgnat, G. Dewaele, P. Abry, « Identification d'anomalies statistiques dans le trafic internet par projections aléatoires multirésolution », *21e Colloque sur le Traitement du Signal et des Images. GRETSI-2007*, Troyes (France), 11-14 septembre 2007.

[**P17**] Guillaume Dewaele, Kensuke Fukuda, Pierre Borgnat, Patrice Abry, Kenjiro Cho, "Extracting Hidden Anomalies using Sketch and Non Gaussian Multiresolution Statistical Detection Procedures," *SIGCOMM 2007, Workshop on Large-Scale Attack Defense (LSAD)*, Kyoto, Japon, 27-31 août 2007.

[**P16**] J. Aussibal, P. Borgnat, Y. Labit, G. Dewaele, N. Larrieu, L. Gallon, P. Owezarski, P. Abry, K. Boudaoud, « Base de traces d'anomalies légitimes et illégitimes, » *6th Conference on Security in Network Architectures and Informations Systems (SAR-SSI 2007)*, Annecy (France), p. 176-185, 12-15 juin 2007.

[**P15**] Patrice Abry, Pierre Borgnat, Guillaume Dewaele, "Sketch based Anomaly Detection, Identification and Performance Evaluation," *Workshop Measurement, IEEE-CS/IPSJ SAINT 2007*, 15-19 janvier 2007, Hiroshima, Japon.

[**P14**] P. Borgnat, N. Larrieu, P. Owezarski, P. Abry, J. Aussibal, L. Gallon, G. Dewaele, K. Boudaoud, L. Bernaille, A. Scherrer, Y. Zhang, Y. Labit, « Détection d'attaques de déni de service par un modèle non gaussien multirésolution », *Colloque Francophone d'Ingénierie des Protocoles (CFIP'2006)*, Tozeur (Tunisie), p. 303–314, 30 octobre - 3 novembre 2006.

[**P13**] A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat et P. Abry, « Une caractérisation non gaussienne et à longue mémoire du trafic Internet et ses anomalies », *5th Conference on Security and Network Architectures (SAR 2006)*, Seignosse (France), p. 176–185, 6-9 juin 2006.

[**P11**] A. Scherrer, N. Larrieu, P. Borgnat, P. Owezarski, et P. Abry, « Non Gaussian and Long Memory Statistical Modeling of Internet Traffic », *4th Workshop IPS-MoMe*, Salzburg (Autriche), p. 176-185, 27-28 février 2006.

[**P10**] P. Borgnat, N. Larrieu, P. Abry et P. Owezarski, « Détection d'attaques de "Déni de Service" : ruptures dans les statistiques du trafic », *Colloque GRETSI-05*, Louvain-la-Neuve (Belgique), 6-9 septembre 2005.

Logiciel

[**L1**] G. Dewaele, P. Borgnat, P. Abry, « IPTools : Analyse de trafic par sketch multirésolution, » logiciel déposé à l'APP (Association de Protection des Programmes) par le CNRS & ENS de Lyon, juillet 2007 (IDDN.FR.001.330007.000.S.P.2007.000.20700).

Chapitre 4

Graphes complexes et traitement du signal

La dernière partie de ce mémoire décrit des activités de ma recherche autour de l'étude de données que l'on peut représenter comme, ou sur, des graphes complexes. Un fil conducteur a été de proposer des approches qui viennent du traitement statistique des signaux, là où les méthodes venues des sciences physiques ou de l'informatique sont les plus répandues dans l'étude des réseaux complexes [BBV08, KKR+99, DM03, New03, PSV04, New10, Kol09, EK10].

Mes études dans ce domaine sont parties de discussion en marge des travaux consacrés à l'analyse et la métrologie du télétrafic Internet (voir le chapitre 3) ; j'ai ainsi abordé ce domaine en collaboration avec le projet INRIA D-NET (E. Fleury), l'équipe "Combining" du LIRIS de l'INSA de Lyon (C. Robardet) et l'équipe "Complex Networks" du LIP6 (C. Magnien et J.L. Guillaume), par l'étude de mesures des contacts entre personnes mesurés par des capteurs Bluetooth, mesures dont l'objectif premier était de suivre leur mobilité (pour l'étude des communications sans fil) et qui se représentent bien par des réseaux dynamiques (publications [J8, P25, P26, P39, C4] en 2008 et 2009).

Par ailleurs, de plus en plus de données peuvent être regardées en tant que réseaux : les réseaux de mobilité ou de déplacement (nous étudierons en 4.3 les déplacements en vélos libre service à Lyon, le système Vélo'v, que nous analysons depuis 2008 [P35, P36, P37, P46, J19, C6]), les réseaux de contacts entre humains (voir 4.2) [P53, J24] et [J8] déjà cité, les réseaux sociaux informatiques, le web, les réseaux de transport d'énergie, les données de réseaux de capteurs, les réseaux de télécommunication, etc. Les traces de trafic IP se plient aussi à des représentations en tant que graphe, on en a vu en 3.6.1. Dans tous ces cas, il est parlant d'ajouter aux données une représentation sous forme de graphe qui résume les relations entre composantes de signaux multi-variés par des liens dans ce graphe.

Cependant, il existait alors peu de travaux proposant des approches venues du traitement du signal sur ces types de données. Or, pour s'intéresser par exemple à la dynamique de ces réseaux, le traitement du signal peut fournir bien des approches puisqu'une des questions y est de savoir comment analyser une information dans le temps. Après des travaux utilisant les approches « signal » sur des données spécifiques ayant des aspects de réseaux (réseaux de contacts sociaux dans 4.2, données de déplacements en Vélo'v 4.3, signaux de capteurs environnementaux ou d'énergie tels que discutés en 2.4.3), nos résultats récents discutés en 4.4 s'inscrivent dans la recherche en **traitement de signaux sur graphes** qui peut proposer des nouvelles méthodes d'analyse pour étudier les réseaux complexes. Dans cette

lignée, la détection de communautés dans des réseaux est reprise dans [P59, P60, P63, P65, Js27] à l'aide de transformées en ondelettes sur graphes (voir 4.4.1). Nous cherchons à étudier des graphes qui sont les signaux d'intérêt, en plus d'analyser des signaux sur une topologie de graphe. Notre étude des réseaux de contacts sociaux [J24, P53] de 4.2.3 a d'ailleurs été principalement méthodologique en proposant une approche de bootstrap sous contrainte pour les réseaux complexes pour définir si un sous-groupe a un comportement attendu ou non. Dans 4.4.2 enfin nous amorçons un programme d'étude des réseaux dynamiques à travers une correspondance menant à l'analyse de signaux et à tracer des équivalents de représentations temps-fréquence pour les réseaux temporels [P57, P62, P64, Ps67].

Avant cela, nous commençons ce chapitre par quelques éclairages bibliographiques sur l'état de l'art de l'analyse des réseaux complexes, en particulier la détection de communautés dans des réseaux [For10], et le domaine émergent du traitement du signal sur graphes [SNF+13].

4.1 État de l'art

Rappelons brièvement qu'un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ est une structure discrète constituée d'un ensemble de nœuds $v \in \mathcal{V}$ et d'un ensemble des liens (ou arêtes) $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ qui dit quels sont les nœuds reliés. On peut aussi représenter $\mathcal{G}$ algébriquement, à travers sa matrice d'adjacence A telle que $A_{ij} = 1$ si et seulement les nœuds i et j sont reliés et $A_{ij} = 0$ sinon. La théorie des graphes (voir par exemple [Bol98]) permet d'étudier une très grande partie des questions liées aux graphes avec une approche de mathématiques discrètes – ou parfois dans un langage d'algèbre linéaire (en particulier via sa matrice d'adjacence), voir par exemple [KG11]. Cependant, un domaine scientifique s'est peu à peu constitué au-delà de cette théorie quand il s'agit d'étudier des graphes qui décrivent des réseaux complexes, en général issus de données mesurées, présentant des propriétés telles que des caractères de petit monde [WS98] ou des lois d'échelles [BA99] ; nous présentons cela en premier avant de nous tourner vers le traitement du signal sur graphes.

4.1.1 Analyse des réseaux complexes

Les réseaux complexes, en particulier statiques, ont été maintenant bien étudiés et des méthodes issues des sciences physiques ou informatiques, voir par exemple [BBV08, DM03, New03, PSV04, EK10, Kol09, New10], offrent des outils pour caractériser des réseaux ou les modéliser, par exemple en vue de l'emergence d'un effet de petit-monde [WS98], ou pour des effets d'hétérogénéités des distributions des dégrés des nœuds [BA99, KKR+99, KRR+00, MR98, DMS01, HK02, SB05]). Dans le cas de réseaux évoluant dans le temps, les travaux sont plus récents, voir par exemple [HS12, KKK+11, CFQS12, HS13] et il reste à trouver comment aborder certains aspects. La question de la co-évolution de processus dynamiques sur des réseaux dynamiques a par exemple été beaucoup étudiée dans des modèles simplifiés [KB08, GS08, HN06, NKB08, VEM08, ZES04], ou dans des modèles plus élaborés de réseaux d'interaction [EMVR06, HB10, TSM+10, GBB09, ZSBB11]. Des travaux s'appuyant sur l'analyse de processus stochastiques (en particulier avec une approche des sciences physiques) on permis aussi d'étudier l'impact de l'évolution temporelle d'un réseau sur les processus de diffusion d'information qui s'y déroulent [CMM+08, CMPS09, GH10, ISB+11, KKP+11, PDC+10]. Enfin, il ne faut pas oublier qu'il y a pour la plupart de

ces réseaux une étape de mesure (donc de métrologie...) associée qui peut faire préférer les appeler du terme graphes de terrain pour indiquer que l'observation de ces graphes, à travers le terrain et les données, ne donne pas forcément un réseau bien déterminé, bien estimé ni même bien fixe dans le temps [Lat07, LM08, Mag10].

Les travaux discutés dans ce mémoire seront parfois porteurs d'un nouveau regard sur un outil maintenant classique en analyse des réseaux complexes (ce sera le cas sur la recherche des communautés dans les réseaux complexes dans 4.4.1), parfois l'approche s'éloignera des approches habituelles pour les réseaux complexes, par exemple les méthodes stastitiques de 4.2 ou nos approches originales pour étudier des réseaux dynamiques dans 4.4.2.

4.1.2 Recherche de communautés dans des réseaux

Une propriété importante rencontrée dans l'analyse des réseaux complexes est la présence très fréquente d'une structure en communautés, aussi appelés modules, et le lecteur consultera avec profit le rapport de S. Fortunato [For10] qui dresse un état de l'art quasiment exhaustif des études liées à cette propriété jusqu'à fin 2009. Cette idée décrit l'existence d'une structuration de réseaux sous forme de groupes de nœuds qui sont plutôt fortement connectés entre eux dans le groupe tandis qu'ils ne sont que moins connectés avec les autres nœuds du réseau. Rechercher des communautés c'est rechercher ces groupes, en général sous la forme de partitions, parfois sous la forme de groupes qui peuvent se recouvrir.

Il existe bien des méthodes pour détecter les communautés dans des réseaux, les premières méthodes découlant d'abord de travaux en informatiques ou en théorie de graphes (clustering spectral, approches informationnelles, méthodes par coupure dans le graphe,...) [VL07, RB10, TCR12]. Une approche venue des sciences physiques a une place clef dans le développement de l'étude des communautés dans des réseaux : la notion de modularité de Newman [New06]. La modularité d'un réseau est au départ une métrique qui permet d'évaluer quantitativement si une partition proposée des nœuds en groupes est bien une structure de communauté du réseau. Elle s'écrit dans un cas simple (réseau non pondéré, non dirigé) :

$$Q = \frac{1}{2m} \sum_{i,j \in \mathcal{V}} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j), \tag{4.1}$$

où m le nombre de liens dans le graphe, k_i est le degré du nœud i et C_i est le numéro du groupe dans lequel est mis le nœud i. L'idée est que la modularité quantifie à quel point on a partitionné les nœuds en groupes ayant un densité plus forte de liens entre eux qu'avec l'extérieur. La modularité devient alors une fonction objectif pour chercher les communautés et bien des travaux ont construit des méthodes efficaces pour les trouver en maximisant Q – jusqu'à la méthode gloutonne de l'algorithme à la Louvain [BGLL08] qui permet d'aborder la question dans des réseaux de très grande taille.

Dans notre travail en 4.4.1, la structure en communauté recherchée pourra être à différentes échelles, appelant le besoin de méthodes multi-résolution (voir le chapitre 12 de [For10]). Plusieurs méthodes existent déjà pour cela, en particulier celles qui modifient l'équation (4.1) pour proposer une modularité ayant un facteur de résolution [Pon06, RB06, AFG08], ou en ré-exprimant la modularité sous la forme d'une marche aléatoire où le temps contrôle la résolution [Lam10, SDYB12]. D'autres méthodes conduisent plutôt à des structures hiérarchiques en communautés [SPGMA07, LF09, CMN08]. Nous reviendrons en 4.4.1 sur

les raisons qui nous conduisent à proposer une méthode multi-résolution de plus détecter des communautés dans un réseau.

4.1.3 Traitement du signal sur ou pour des graphes

Le traitement du signal (ou des images) sur graphe regroupe des méthodes qui visent principalement à étudier des signaux qui sont indexés par des graphes (plutôt que par le temps ou l'espace), et [SPM13] propose un panorama récent d'articles sur ce domaine émergent pour ce qui concerne le traitement du signal sur graphes [SNF+13], ou qui proposent des méthodes de traitement d'image en les codant à l'aide de graphes – on consultera [LG12] en tant qu'ouvrage récent sur ce sujet [1]. Ce domaine connaît actuellement un intérêt grandissant dans la communauté, car il permet d'aborder de nouvelles modalités de données comme celles suggérées plus haut.

Parmi les questions qui ont connu des avancées récentes, celles visant à trouver des représentations ou des transformations adaptées à des signaux sur graphes ont particulièrement attiré les efforts. Par exemple, on dispose maintenant de plusieurs travaux définissant des transformées en ondelettes sur graphes [CK03, CM06, NO09, HVG11], des ondelettes discrètes aussi [LV13, SWHV14], des notions de transformée de Fourier et de filtre sur graphes [SNF+13, SM13] des analyses temps-fréquence [SRV13], etc. Bien entendu, les questions de clustering ont été regardées par les méthodes de signal sur graphes, par exemple pour les réseaux multi-couches [DFVN12, DFVN13, DFVN14] ou pour des réseaux évoluant dans le temps [XKH13, XKH14], ou même pour de la détection de communautés (au sens de 4.1.2) dans le temps [XKH11].

Notre approche est actuellement orientée vers des contributions méthodologiques qui s'appuient sur notre expertise en analyse multi-échelle et en traitements non stationnaires. L'enjeu est de pouvoir aller à terme vers l'étude de propriétés multi-échelles évoluant dynamiquement dans les réseaux complexes (des communautés par exemple) ce qui reste un sujet d'étude. L'idée est que les outils de traitement de signal pour graphes sont pertinents pour étudier les réseaux complexes eux-mêmes, sans se limiter à analyser les signaux indexés sur des réseaux.

4.2 Les réseaux de contacts entre humains

4.2.1 Mesurer des interactions sociales ou des contacts entre humains

Discuter de réseaux pour étudier les interactions sociales n'a rien de neuf et remonte au moins aux années 1950 et 1960, déjà avec des concepts de graphes, sans oublier des précurseurs plus anciens [Mer04]. Ce qui nous a amené sur ce terrain d'étude est le développement dans ces dix dernières années d'approche expérimentales pour mesurer les interactions sociales entre personnes à l'aide de capteurs distribués qui autorisent une collecte

1. Il ne faut cependant pas oublier que le traitement des images a une très longue histoire de modélisation à l'aide de graphes, que ce soit par des champs de Markov (liés aux modèles graphiques) dans la suite des travaux [GG84] ou toutes les approches venues de l'informatique et de la vision par ordinateur, par exemple par *graph-cut* [GPS89].

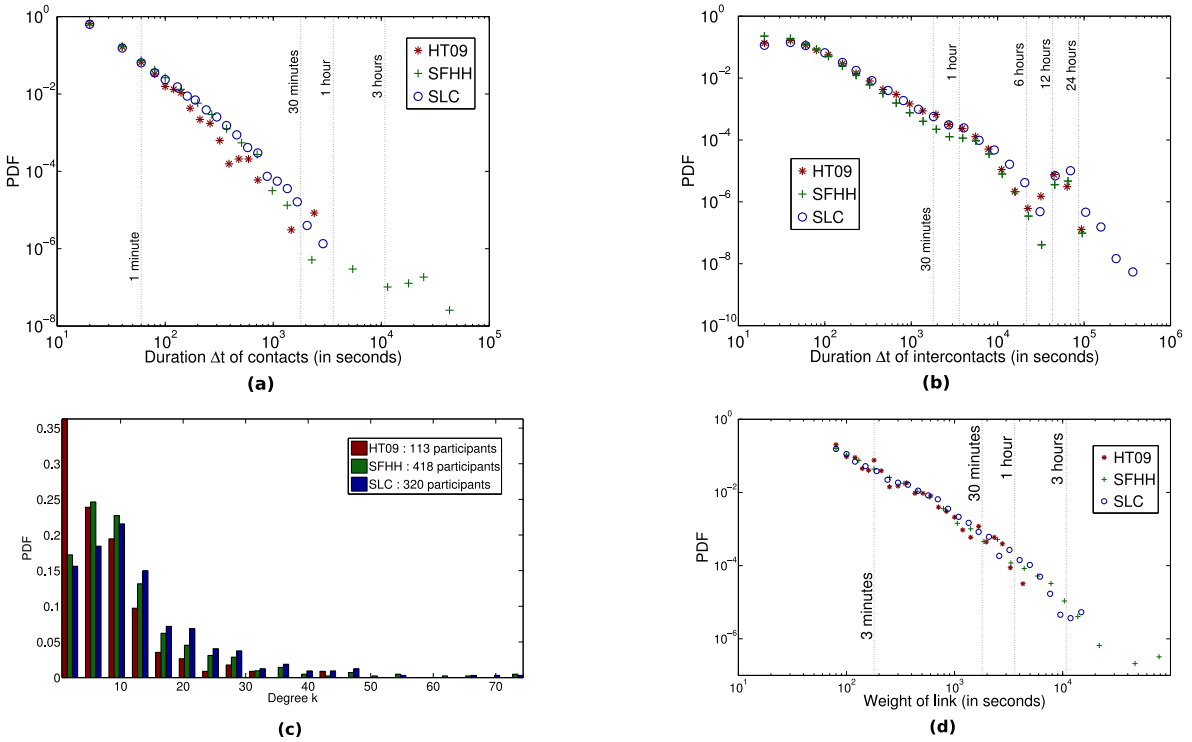


FIGURE 4.1 – Données de contacts entre humains (expériences Sociopatterns). (a) Comparaison des distributions des durées de contacts pour trois expériences différentes. (b) Distribution des durées d'inter-contact. Un intervalle d'inter-contact est défini, pour chaque nœud, comme le temps entre deux contacts successifs avec n'importe qui. (c) Distribution des degrés dans le réseau agrégé. Le degré d'un participant correspond au nombre total de participants avec qui il a été en contact pendant le temps de mesure. (d) Distribution des poids sur les liens. Le poids d'un lien est le temps de contact total entre les 2 participants reliés par le lien.

de données résolues dans le temps et potentiellement l'espace, là où les réseaux en sociologie s'appuyaient sur des observations et des enquêtes qui ne permettent que rarement l'exhaustivité ou la résolution amenée par les expériences avec des capteurs.

Différentes expériences. Au rang des expériences de ce type, le projet *MIT Reality Mining* est précurseur et a permis d'enregistrer sur des longues durées (quelques mois) la proximité entre une centaine de personnes munies des capteurs employés (des téléphones instrumentés pour, communiquant par Bluetooth), principalement dans un contexte d'analyse des organisations [EP06, Pen08]. En parallèle, les capteurs Bluetooth *Imote* ont permis la collecte des contacts entre personnes à des conférences [HCS+05] ou pendant des sorties rollers [TLB+09], cette fois sur des données plus courtes mais des densités de capteurs plus grandes, donc avec beaucoup plus de contacts enregistrés. Le Bluetooth ayant comme défaut d'avoir une portée un peu trop grande (quelques mètres) pour une bonne résolution des contacts (on parle plutôt d'une proximité entre les personnes qui les portent), la technologie des RFID (Radio Frequency Identification Device) a permis un gain de résolution considérable avec des portées contrôlables de 1 à 3m seulement et une directivité des interactions (deux personnes se tournant le dos ne seront pas en contact) ; c'est par exemple ce qu'on retrouve avec les données issues de la plate-forme Sociopatterns.org [Soc, CBB+10, CVB+10,

SKL+10, SVB+11, ISB+11, IRB+11] ou par les capteurs des projets iBird et MOSAR pour des mesures dans un environnement hospitalier [FCF+11, LCL+11, LLC+12, GGF13]. Les contacts entre les participants sont mesurés grâce à des petits badges RFID portés par chacun : quand deux participants sont face-à-face et à moins d'un mètre cinquante à deux mètres l'un de l'autre, un contact est mesuré et enregistré si il y a une antenne collectrice à portée (de 20m à 30m).

Nos travaux [J8, P25, P26, P39, C4] s'appuient sur les données Imote et MIT Reality Mining collectées par d'autres tandis que les travaux [J24, P53] utilisent des données Sociopatterns collectées par nos soins ou par les partenaires du projet Sociopatterns.

Caractéristiques usuelles des réseaux de contacts humains. Les mesures sont initialement sous une forme de logs des contacts entre deux capteurs (RFID, Bluetooth, etc.) avec une information de temps (instant, éventuellement durée – la durée pouvant être reconstruite a posteriori si on a une suite non interrompue de contacts en temps) et parfois en espace (souvent : proximité avec telle base de collecte de données dont on connaît la position). Pour les données de [J24, P53], partant donc du lignes du log de forme (τ, r, i, j) où τ est le temps auquel un collecteur r reçoit l'information que les capteurs RFID des individus i et j ont été en contact à courte portée face à face, nous construisons un réseau dynamique dont les badges i sont les nœuds et les contacts les liens. Avec les paramètres de fonctionnement du matériel Sociopatterns [CBB+10], la probabilité de détecter au niveau du collecteur que 2 individus portant des badges RFID sont en contact est de plus de 99% sur des fenêtres de 20s ou plus. Nous agrégeons donc les données sur des fenêtres de 20s. À chacune de ces périodes centrées au temps t, nous calculons une matrice d'adjacence A^t telle que $A^t_{ij} = 1$ si et seulement les nœuds i et j ont échangé au moins un paquet radio durant la fenêtre t, sinon $A^t_{ij} = 0$.

Partant des graphes dynamiques $\mathcal{G}^t = (\mathcal{V}, \mathcal{E}^t)$ associés à ces matrices d'adjacence dynamiques, où l'ensemble des nœuds possible $\mathcal{V}$ est cependant fixe, plusieurs propriétés usuelles pour des réseaux complexes ont été regardées. Une première caractéristique associée à la dynamique est que les temps entre les contacts ou les durées d'inter-contacts sont distribués avec des lois larges, qui s'apparentent à des lois de puissance. Pour mesurer cela, un contact entre i et j est défini comme un événement en temps par une suite ininterrompue de 1 dans la suite des $\{A^t_{ij}\}$. Sa durée est la longueur d'une telle séquence (ramenée en seconde). La figure 4.1 (a) et (b) compare ces caractéristiques pour les données de contact collectées en trois occasions : la conférence ACM HyperText de 2009 (HT09) [ISB+11], le congrès de la *Société Française d'Hygiène Hospitalière* (SFHH) [CBB+10] et les conférences de l'APS à Salt Lake City (SLC) co-localisées : la GEC (*Gaseous Electronic Conference*) et la rencontre de la DPP (*Division of Plasma Physics*) de novembre 2011 [J24, P53]. Les comportements sont similaires et conformes avec les premières mesures des données du MIT [EP06], ou Imote [HCS+05], ré-analysées dans [J8]. Bien entendu, à temps long devenant de l'ordre de grandeur d'une partie de la journée (12h par exemple), on voit les lois de puissance laisser apparaître des temps caractéristiques de la journée qui rythme les activités humaines. Ensuite, on peut se tourner vers des propriétés statiques telles que celles en figure 4.1 (c) et (d), où l'on a agrégé la suite de réseaux dynamique (par une somme de A^t_{ij} sur tous le temps). On retrouve des lois de distributions larges pour les degrés des nœuds et pour les poids des liens (voire même avec une queue en loi de puissance), les poids étant définis comme la somme du temps total de contact entre 2 nœuds. D'autres propriétés sont intéressantes

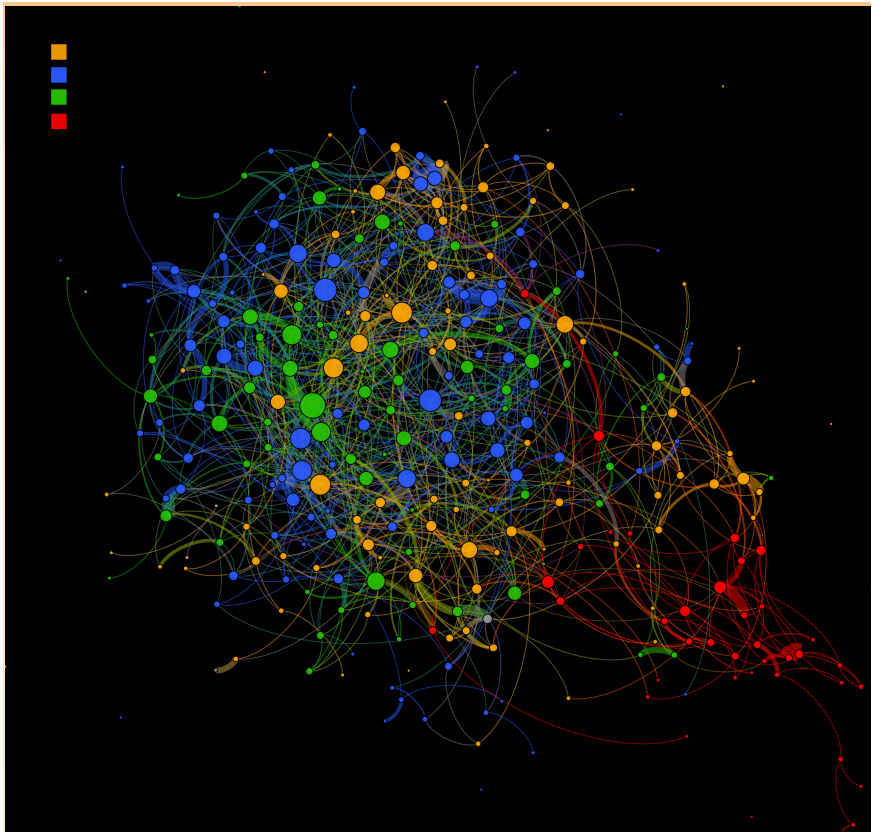
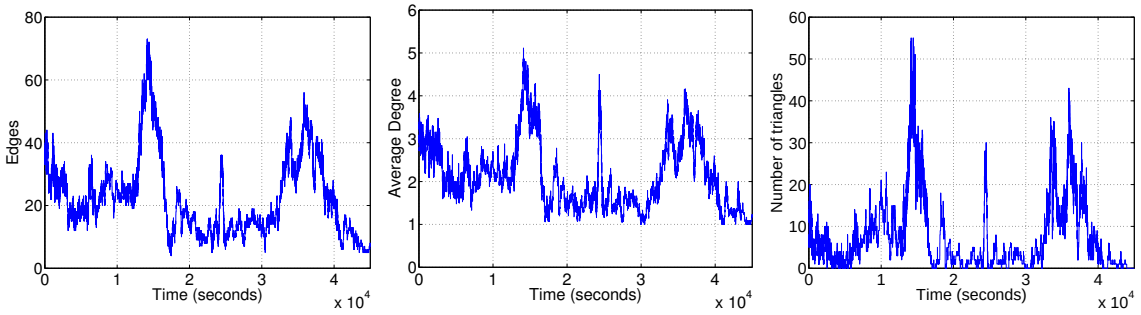


FIGURE 4.2 – Réseaux de contacts entre humains. Les contact mesurés par tags RFID lors de la conférence DPP-GEC à Salt Lake City en novembre 2011 sont représentés de manière agrégée comme un graphe où chaque nœud est une personne, et chaque arête est d'épaisseur proportionnelle au temps de contact entre les deux personnes. Les couleurs indiquent si la personnes est un doctorant de DPP (STP), chercheur junior (JUP) ou senior (SEP) de DPP ou un participant à GEC.

(et amplement trouvées dans la littérature) pour ces réseaux : ils sont peu denses (peu de liens en proportion de ce qui serait possible), il ont un fort taux de clustering local (donc beaucoup de triangles dans ces données), il y a souvent beaucoup de composantes connexes à un instant donné mais un effet de petit monde (le diamètre du graphe est petit) dans les graphes agrégés (voir les figures dans [J8, J24]).

Visualisation du réseau. Un dernier élément pour décrire des données de contacts humains, telles que celles de SLC, est de visualiser des graphes induits. Ici, nous donnons le graphe le plus simple que l'on peut extraire des données, qui est le graphe de contact agrégé sur toute la durée de la conférence – c'est donc un graphe statique, représenté en figure 4.2. On y voit déjà que, sur ce type de représentation simple spatialisée (et un peu arbitraire) à l'aide d'un outil de visualisation de graphes qui met d'autant plus proches des nœuds qu'ils sont liés et repoussent ceux qui ne le sont pas [JHVB12], un groupe semble se détacher : ce sont les participants à GEC tandis que les autres participent à la rencontre DPP. Nous étudierons en 4.2.3 par la méthode de [J24, P53] à quel point ce sous-groupe du réseau interagit normalement ou pas avec le restant des individus.



(a) (b) (c)

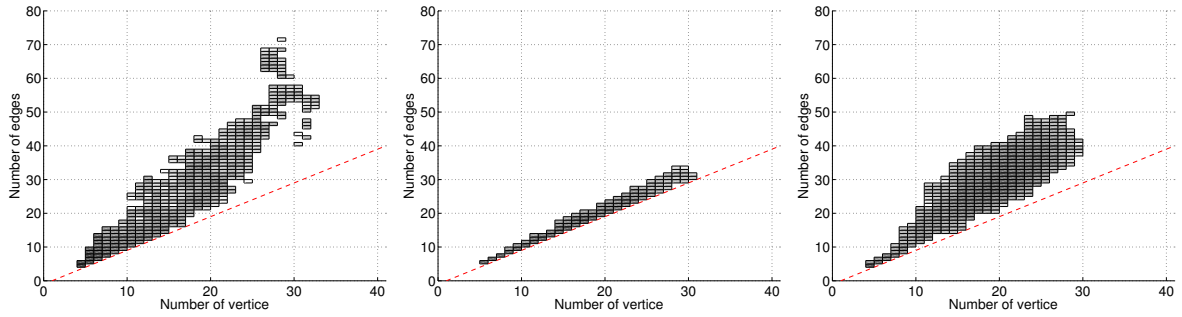
FIGURE 4.3 – Signaux associés aux données Imote. On montre les évolutions non stationnaire du nombre d'arêtes actives, du degré moyen et du nombre de triangles en fonction du temps.

4.2.2 Un modèle dynamique pour les réseaux de contacts humains

Une première contribution, en étudiant les données MIT et Imote [J8, P25, P26, P39, C4], a été de proposer un modèle dynamique des contacts entre les capteurs. Ce modèle part du constant suivant [J8] : les caractéristiques des $\mathcal{G}^t$ telles que le nombre de liens, de composantes, de triangles, le degré moyen, ont des temps de corrélation souvent assez longs, de 1h à 2h par exemple pour les données Imote, et des séries plutôt corrélées entre elles, comme on peut le voir en figure 4.3. Si l'on s'intéresse à la série du nombre d'activations et d'inactivations de liens individuels à chaque instant, elle est à corrélation plutôt courte par rapport aux autres quantités (environ 10 minutes). De plus, le nombre d'activations ou d'inactivations de liens reste toujours faible à un instant donnée. Il est alors possible d'approcher la dynamique du graphe par un processus markovien d'activations et d'inactivation des liens. C'est le modèle de réseau dynamique que nous avons développé dans [J8] ; en parallèle ou ensuite, des travaux plus théoriques [CMM+08, CMPS09, GH10, GP11] ont étudié des propriétés telles que des temps de diffusion dans le réseau pour des modèles de ce type. Plus récemment, le modèle de *Time-Varying Graphs* de [CFQS12] offre un cadre générale pour des réseaux dynamiques qui peut englober le modèle que nous proposions – il resterait à déterminer si ce nouveau cadre général est opérationnel.

Modèle markovien pour les réseaux dynamiques. Le modèle proposé dans [J8] permet de retrouver toutes les propriétés de temps de contact (approximativement en loi de puissance), le fort clustering (c'est-à-dire le ratio entre le nombre de triplets fermés dans le réseau (groupe de 3 nœuds complétement connecté) divisé par le nombre de triplets connectés) et l'existence de communautés qui se propagent dynamiquement dans le réseau. Nous avons étudié aussi la stabilité dans le temps des composantes connexes et des groupes de nœuds obtenus par une approche de data mining dans le réseau dynamique (plus de détails sont dans [J8] à ce propos).

Ici, seul le principe du modèle sera donné avec quelques illustrations. Pour obtenir une séquence de graphes $\mathcal{G}^t = (\mathcal{V}, \mathcal{E}^t)$ où l'ensemble $\mathcal{E}^t$ varie au cours du temps t discrétisé (ici par pas unité pour simplicité), notre idée principale est de construire un processus de Markov pour chaque arête possible qui aura deux états : actif ou inactif, et un paramètre de mémoire τ qui retient le temps écoulé depuis le dernier changement d'état du lien. Comme les lois de distributions du temps durant lequel un lien est actif (2 nœuds sont en contacts)



(a) (b) (c)

FIGURE 4.4 – Modèles pour les données Imote de contacts entre humains. On indique les distributions jointes du nombre d'arêtes actives avec le nombre de nœuds connectés (non isolés), pour les données réelles en (a), le modèle **1** en (b) et le modèle **2** en (c).

sont distribués quasiment en loi de puissance dans les données réelles (voir 4.2.1), on propose de déduire des distributions empiriques les probabilités $P_+(\tau)$ qu'un lien s'active si il était inactif depuis τ pas de temps et $P_-(\tau)$ qu'il s'inactive si il était actif depuis une durée τ. Si on connaît les distributions de la durée des contacts $P_{ON}(\tau)$ ou des inter-contacts $P_{OFF}(\tau)$ (définis cette fois pour les liens) (ON codant pour actif et OFF pour inactif), ou si l'on se donne un modèle, on doit avoir :

$$P_-(\tau) = \frac{P_{ON}(\tau)}{\prod_{i=1}^{\tau-1}(1-P_-(i))}, \quad \tau \geq 2, \quad P_-(1) = P_{ON}(1) \tag{4.2}$$

$$P_+(\tau) = \frac{P_{OFF}(t)}{\prod_{i=1}^{\tau-1}(1-P_+(i))}, \quad \tau \geq 2, \quad P_+(1) = P_{OFF}(1) \tag{4.3}$$

Avec ces règles pour les probabilités de changement d'état des liens, un simple algorithme de simulation de Monte-Carlo qui, à chaque pas de temps, considère tous les liens et teste si ils changent ou non, conduit à des réseaux dynamiques qui ont les distributions attendues pour les temps de contact ou inter-contacts. En figure 4.4 est indiqué, en la supposant stationnaire pour l'estimation, la distribution jointe du nombre de liens actifs en fonction du nombre de nœuds connectés (c'est-à-dire ayant au moins une connexion, non isolés), dans le cas des données Imote en (a) et d'une simulation de ce modèle **1** en (b). On voit ici que ce principe de modélisation n'est pas suffisant : il ne capture pas correctement la densité du réseau qui, pour un nombre donné de nœuds connectés, a plus de liens actifs que ce que le seul respect des distributions de temps de contact et d'inter-contact impose.

Il est cependant possible de corriger cela en ajoutant un deuxième aspect au modèle : nous imposons une coordination globale au réseau qui force son évolution vers un plus fort clustering (et du coup une plus forte densité) en favorisant la création de triangles. Ce principe vient de la mesure des probabilités d'activation ou d'inactivation de liens selon que ce lien ferme un triangle ou non. Notons $P_{+/tri+}$ (respectivement $P_{+/tri=}$) les proportions d'activation de liens qui augmentent (resp. ne changent pas) le nombre de triangles dans le réseau. Soit $f_{+/tri+}$ (resp. $f_{+/tri=}$) la proportion moyenne de liens inactifs qui fermeraient un triangle si activé resp. ne changeraient pas le nombre de triangles). Ces proportions sont données dans le tableau 4.1 (en pourcentage), mesurées sur les données Imote, du MIT et sur le modèle **1** proposé. Le constat est que les triangles se ferment bien plus souvent dans

les données que dans le modèle **1** où les liens évoluent en toute indépendance.

	$P_{+/tri+}$	$P_{+/tri=}$	$f_{+/tri+}$	$f_{+/tri=}$
Imote	44 %	56 %	6 %	94 %
MIT	40 %	60 %	7 %	93 %
Modèle **1**	10 %	90 %	5 %	95 %

TABLE 4.1 – Proportion d'arêtes qui forment de nouveaux triangles ou non lors de leur activation (P) ; proportion d'arêtes inactives qui, si elles deviennent activent, forment ou non de nouveaux triangles (f).

Le modèle **2** que nous proposons favorise donc la création des triangles en pondérant la loi de probabilité de transition entre les états actifs et inactifs comme il suit :

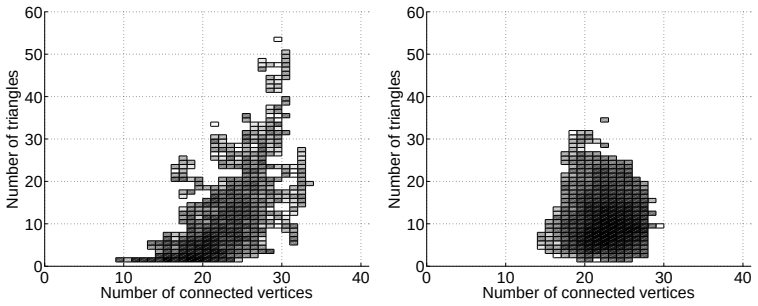
$$P_{tr}(e, G_t) = \begin{cases} P_+(\tau(e))\frac{P_{+/tri=}}{f_{+/tri=}} & \text{pour l'activation du lien } e \text{ sans fermer de triangle,} \\ P_+(\tau(e))\frac{P_{+/tri+}}{f_{+/tri+}} & \text{pour l'activation du lien } e \text{ fermant un nouveau triangle.} \end{cases} \tag{4.4}$$

Cette règle, qui s'appuie sur une contrainte globale favorisant les triangles, permet de rapprocher significativement le modèle **2** du comportement des réseaux expérimentaux. La figure 4.4 compare en (c) la distribution jointe du nombre de liens actifs en fonction du nombre de nœuds connectés au précédents modèle et aux données Imote. La figure 4.5 compare celle du nombre de triangles en fonction du nombre de nœuds connectés pour ce modèle **2** et les données Imote. Sans surprise (surtout dans le second cas), on a un bon accord. Notez que [J8] propose d'imposer d'autres types de coordination globales dans le réseau (imposer les distributions jointes de certaines caractéristiques) mais elles se révèlent en fait peu utiles ou de peu d'effet en comparaison de favoriser les triangles.

Discussion. Les parties des distributions jointes qui ne semblent pas atteintes dans la simulation sont en fait liées aux instants où le comportement non stationnaire du réseau dynamique de contacts humains finit par intervenir : dans les donnée Imote, il y a des moments lors de la pause café dans la conférence ou le nombre de liens devient grand par rapport à la dynamique habituelle (on voit deux pics large et un fin de ce type sur les séries de la figure 4.3). Il faudrait en fait varier les paramètres du modèle pour tenir compte de cette non stationnarité pour capturer les comportements lors de ces pics.

Le modèle que nous proposons avait pour objectif initial de proposer un outil de simulation de réseaux dynamiques représentatifs de la dynamique des contacts entre humains, qui peut servir de modèle étalon pour des travaux sur les réseaux dynamiques. Cependant, il nous enseigne qu'un élément est important pour capturer la dynamique des contacts humains : les triangles sont localement plus probables que le hasard. Cela n'est pas vraiment surprenant : si un individu est en contact avec 2 autres, il y a de grandes chances que ces 2 autres personnes soient elles aussi en contact !

Ce modèle n'a en réalité pas été tellement employé jusqu'ici. Nous avons récemment repris des travaux dessus en le comparant au modèle de [SBB10, ZSBB11], formulé sur des données Sociopatterns. Ce modèle propose lui un mécanisme d'évolution à l'échelle du groupe d'individu connecté. Il postule donc de fermer les triangles dès l'analyse des données expérimentales. Nous pensons qu'il sera possible de mélanger les deux approches. Notre modèle permet d'avoir une évolution dynamique plus fine (puisqu'on est à l'échelle de l'individu et des changements de liens) mais avec une approche d'ingénierie qui utilise



(a) (b)

FIGURE 4.5 – Modèles pour les données Imote de contacts entre humains. On indique les distributions jointes du nombre de triangles dans le graphe avec le nombre de nœuds connectés (non isolés), pour les données réelles en (a) et le modèle **2** en (b).

des paramètres et principes (tel que favoriser le nombre de triangles) aisément assimilables à un *deus ex machina*, là où le modèle de [SBB10, ZSBB11] (formulé au niveau des groupes et des changements de groupes) propose une analyse théorique plus élaborée.

4.2.3 Test par bootstrap contraint sur des sous-groupes du réseau

Le travail sur les réseaux de contacts entre personnes a justement repris cette fois en collaboration avec certains collègues physiciens, A. Barrat (du CPT, Marseille) en particulier, qui ont proposé le modèle [SBB10, ZSBB11], un collègue physicien au sein du laboratoire, J.-F. Pinton, et le projet Sociopatterns à Turin (C. Cattuto). Nous avons en particulier effectué pour les travaux [J24, P53] des mesures lors de la conférence scientifique jointe DPP et GEC à SLC, ainsi qu'indiqué en 4.2.1.

La question initiale qui motivait notre étude était de savoir si l'on peut inférer de manière statistiquement fiable si oui ou non les deux sous groupes correspondant aux deux communautés GEC et DPP se sont mélangés ? Le problème est que le réseau dynamique mesuré est une réalisation unique d'un événement qui ne peut être reproduit. Comment savoir alors si les différentes mesures de mixité auxquelles on peut penser sont statistiquement valables ? Plus généralement, comment donner un sens statistique à des propriétés au sein d'un réseau mesuré une seule fois ?

Le travail dans [J24, P53] a été de développer une une méthode basée sur des ré-échantillonnages du graphe (par du bootstrap de graphes contraints) qui permet de définir des intervalles de confiance à certaines propriétés d'un sous-groupe de nœuds donné dans le réseau, en fonction d'une hypothèse nulle définissant le comportement à tester. Le résultat est de pouvoir déterminer si l'interaction entre les deux conférences a été correct ou non. Je donne les éléments principaux de la méthode dans ce mémoire (les détails étant bien sûr dans [J24]) ainsi que le résultat obtenu sur les données SLC.

Adapter la méthode de bootstrap à des graphes contraints. Le bootstrap [Efr82, ZI04] est une technique en statistique non paramétrique pour calculer des intervalles de confiance ou faire des tests pilotés par les données. Face à une seule réalisation des données, nous nous tournons naturellement vers une telle technique pour tester si le sous-groupe des participants à GEC se comporte comme les autres sous-groupes du réseau de contact

général DPP + GEC ou non – ce sera notre hypothèse à tester. Néanmoins, un obstacle pour déployer les techniques de bootstrap pour tester des propriétés d'un sous-groupe dans le réseau est que les nœuds ne sont pas indépendants les uns des autres ; cette difficulté est proche de celle pour utiliser le bootstrap pour des séries corrélées qui appellent des techniques telles que le bootstrap par blocs [Lah99, Pol03]. Dans notre étude, nous proposons de construire l'ensemble de bootstrap en imposant des contraintes aux sous-graphes choisis, les contraintes étant choisies pour explorer différentes hypothèses à tester pour le sous-groupe d'intérêt.

Plus concrètement, soit un sous-groupe de nœuds du réseau, $X^0 \subset \mathcal{V}$. Une hypothèse nulle est formulée à propos du comportement normal attendu des sous-groupes du réseau. Dans notre contexte d'analyse des données SLC, les observables employées caractérisent à quel point les individus de X^0 (de GEC) se sont mélangés (pour leurs contacts) avec les autres (ceux de la rencontre DPP) : nombre de liens (ou somme des poids) dans un groupe, l'autre ou entre les deux, modularité de la partition entre X^0 et le reste. Notant Z les observables, le principe du test est que certaines vont être contraintes pour créer un ensemble de bootstraps représentatif de l'hypothèse nulle à avoir des valeurs Z vérifiant

$$Z^0(1-\delta) \leq Z \leq Z^0(1+\delta) \tag{4.5}$$

où Z^0 est la valeur de l'observable pour X^0 et δ un paramètre à fixer qui dit à quel point la contrainte est relaxée. Par exemple, on comparera X^0 à des sous-groupes de même taille et ayant la même modularité à δ près que celle de X^0 (pour avoir un comportement similaire en tant que communauté dans le réseau). Cet ensemble de bootstrap est construit en générant, par recuit simulé [BM95] et de manière indépendante pour chaque nouveau tirage, un grand nombre N_B de graphes satisfaisant la (ou les) contraintes de type (4.5). Ensuite, pour un ensemble d'observables que l'on pense pertinentes, on élabore les intervalles où l'on ne rejette pas l'hypothèse nulle pour Z^0 (observable Z pour X^0) sous une probabilité de fausse alarme α' en utilisant les distributions empiriques des Z sur les N_B sous-groupes tirés par bootstrap. Comme on peut tester en même temps plusieurs observables (non contraintes) en nombre F', la probabilité de fausse alarme du test joint est alors majorée par $\alpha = \alpha' F'$ selon la correction de Bonferroni.

Compromis entre contraintes et puissance du test. En fait, la méthode reposant sur une hypothèse nulle définie via une contrainte relaxée par δ, on est face à un compromis entre la force de la contrainte et la force du test. Plus δ est petit, plus les sous-groupes bootstraps seront représentatifs de l'hypothèse nulle mais plus la taille de l'ensemble des sous-groupes admissibles se réduit ; à la limite de $\delta = 0$, il pourrait ne contenir que le sous-groupe X^0 à tester et donc tout test bootstrap acceptera X^0 comme étant cohérent avec cette hypothèse nulle mal décrite par les bootstraps – le test perd toute puissance. À l'inverse, si δ augmente, on s'éloigne de plus en plus de l'hypothèse nulle supposée être testée. Il faut donc trouver comment fixer δ, le plus petit possible, pour que le test garde quand même la puissance attendue.

Nous avons analysé cela par simulation dans [J24] sur des graphes de Chung-Lu [CL02, MH11], et nous en avons déduit un seuil δ^* pour δ, qui se transpose en des seuils σ_u^* et χ^{2*} sur critères plus opérationnels pour caractériser la taille de l'ensemble bootstrap :
- on mesure l'écart type σ_u de la distribution du nombre de fois que chaque nœud de $\mathcal{V}$ est choisi dans une sous-groupe de l'ensemble de bootstrap. Cela mesure à quel point on choisi

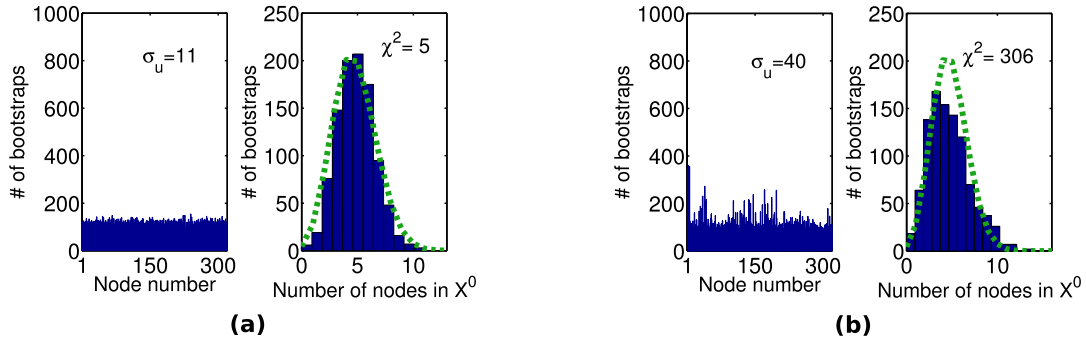


FIGURE 4.6 – Valeurs de σ_u et χ^2 obtenues pour la méthode de bootstrap pour X^0 = GEC, en vue d'un test (a) comparant aux groupes de même taille ou (b) aux groupes de même taille et même modularité (avec $\delta = 15\%$). Gauche : histogramme du nombre d'utilisation de chaque nœud dans l'ensemble bootstrap et son écart-type σ_u. Droite : histogramme du nombre de fois qu'un nœuds de X^0 est pris dans l'ensemble de bootstrap et distance χ^2 entre l'histogramme et la loi hypergéométrique théorique (ligne pointillée).

les nœuds uniformément ;
- pour chaque sous-groupe dans l'ensemble de bootstrap, on regarde combien de nœuds viennent de X^0. Sans contraintes (et sans dépendance donc), la loi théorique est une distribution hypergéométrique. On calcule alors un χ^2 entre cette loi et la distribution empirique pour quantifier à quel point les sous-groupes bootstraps ressemblent à X^0.

L'étude sur δ se traduit en des seuils σ_u^* et χ^{2*} au-dessus desquels on ne peut plus conclure sur le test formulé car on a des indications que l'ensemble de bootstrap construit n'est pas assez large pour être discriminant.

Etude de cas : données SLC. L'hypothèse nulle retenue pour comparer si les participants à GEC se sont bien mélangés à ceux de la DPP est de tester leur comportement en comparaison de tous les sous-groupes qui auraient la même taille et la même modularité – donc le même comportement à rester plus ou moins entre soi. Suite à l'analyse du compromis précédent, on fixe $\delta = \delta^* = 15\%$ et on a $\sigma_u^* = 60$ and $\chi^{2*} = 950$. La figure 4.6 montre les distributions sous-jacentes aux calculs de σ_u et χ^2, pour illustrer en quoi elle mesure l'uniformité des sous-groupes bootstraps. Ici, les seuils ne sont pas dépassés.

En figure 4.7, on compare le résultat du test sans contrainte (comparaison donc aux sous-groupes de même cardinalité) et avec la contrainte sur la modularité, pour le sous-groupe GEC et pour le sous-groupes constitué par tous les étudiants de la conférence. Le choix de prendre ce groupe pour comparaison vient de ce que la modularité mesurée est alors de 0.145 alors qu'elle n'est que 0.100 pour la partition en GEC et le reste. Les étudiants auraient-ils donc une tendance à la grégarité plus grande que GEC ? Le résultat du test sans contrainte ne nous apporte pas vraiment de réponse : pour les deux sous-groupes, l'hypothèse nulle comme quoi ils se comporteraient comme des sous-groupes pris au hasard (de même taille) dans les données est rejetée. En revanche, cela change pour le test sous contrainte de modularité. Cette fois, le groupe des étudiants revient à un comportement quasi normal sous l'hypothèse nulle tandis que GEC continue à être singulier, plus particulièrement vis-à-vis du temps de discussion entre soi (plus important que sous hypothèse nulle) ou avec les autres (plus faible) .

La réponse finale apportée à notre question : les participants aux deux conférences se

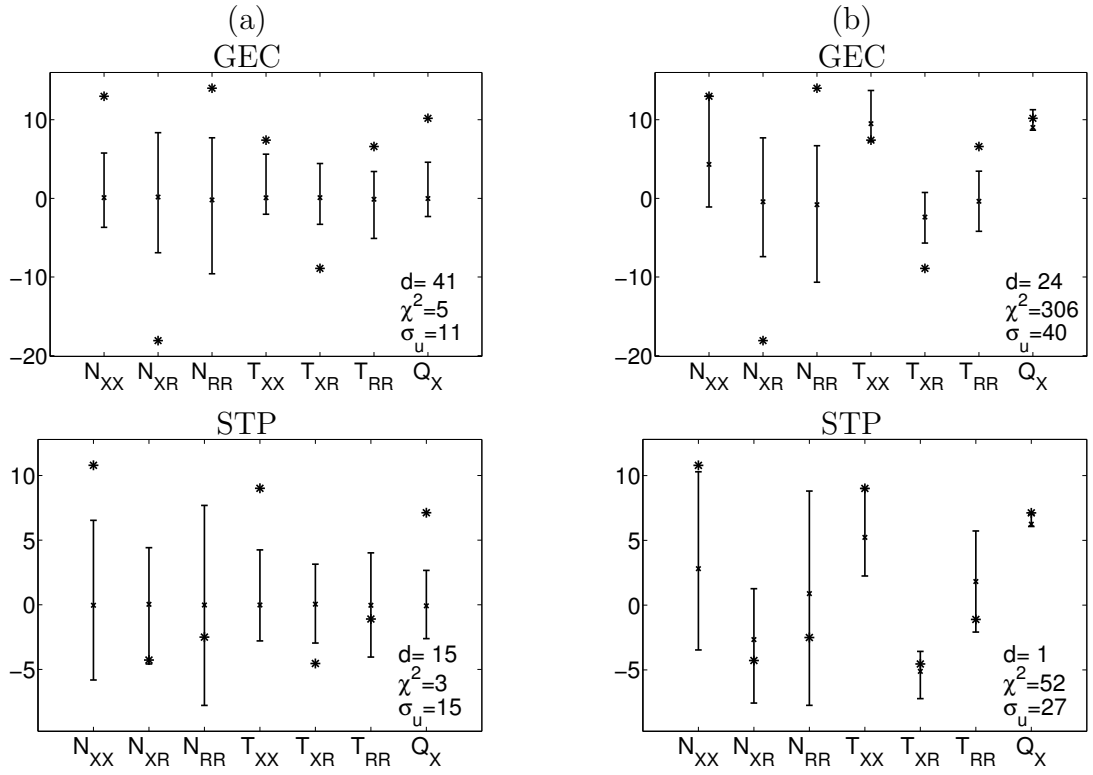


FIGURE 4.7 – (a) Résultat du test bootstrap en comparant avec des groupes de même taille pour GEC et STP. (b) Résultat du test bootstrap en comparant avec des groupes de même taille et de même modularité (avec $\delta = 15\%$) pour GEC et STP. Pour chaque observable Z, on trace sa version réduite de Fisher z (normalisée et centrée par l'aléatoire) et les intervalles d'acceptations de l'hypothèse nulle avec une confiance $1 - \alpha'$. L'étoile noire indique la valeur z^0 du groupe testé. Ici, $\alpha = 5\%$, i.e. $\alpha' = \frac{\alpha}{F'} = \frac{0.05}{6} = 0.8\%$. Dans chaque cas, le scalaire d est une distance entre les intervalles d'acceptation de l'hypothèse nulle et les valeurs pour ce groupe. χ^2 et σ_u sont les paramètres de contrôle de l'ensemble de bootstrap et du test.

sont-ils effectivement mélangés ? peut sembler plutôt décevante. Elle est en effet plutôt négative puisque le sous-groupe GEC se distingue encore par des discussion faibles avec les autres personnes de la rencontre DPP par rapport aux autres sous-groupes comparables. On pouvait s'en douter car en réalité les lieux principaux des deux conférences étaient assez éloignés dans le bâtiment (voir la carte dans [J24]) ! Cependant, la réponse a été statistiquement validée à l'aide du protocole bootstrap décrit ici et, sutout, ce protocole et l'étude qui l'établit sont ré-utilisables pour s'intéresser à bien d'autres tests d'hypothèse dans des réseaux pour lesquels on ne mesurerait qu'une instance, ce qui est bien souvent la norme.

4.3 Le réseau de déplacement en Vélo'V

Les systèmes de location instantanée de vélos en libre-service (VLS) urbains ont connu récemment un développement très rapide à l'échelle internationale. Ils préfigurent des mutations sans précédent des mobilités et urbanités contemporaines. Cette innovation s'apparente à un transport public individuel mais s'en écarte car l'usager est bien plus libre dans son

déplacement. Le succès public des VLS amène à ce poser des questions, selon des visées diverses : planification, gestion opérationnelle, évaluation de l'action publique, analyse des mutations sociales. Nous présentons ici les études menées sur le système de location de vélos en libre service à Lyon, appelé Vélo'v, en fonctionnement depuis mai 2005. Notre travail d'étude des données Vélo'v a démarré en 2009 et se poursuit jusqu'à maintenant [J19, C6, P35, P36, P37, P46, P62, P64, Js25, Js28].

4.3.1 L'étude des systèmes de vélos libre service

Étudier les VLS conduit à deux questionnements complémentaires : un VLS est un système socio-technique dont on souhaite analyser, et éventuellement modéliser, le fonctionnement et les usages effectifs ; un VLS est aussi une innovation dont on cherche à étudier les modalités de son appropriation sociale. Le premier angle d'approche s'inscrit dans les recherches en transport où l'on tente de mieux comprendre la mobilité des gens. Deux caractéristiques des VLS, en plus de leur popularité, justifient qu'on s'y intéresse dans ce cadre : le déplacement en vélos en ville est une petite partie des mouvements qui est mal étudiée via les « Enquêtes Ménage Déplacement » qui servent usuellement à cela et l'accès aux mouvements en VLS permettent de sonder plus généralement les pratiques de déplacements en vélos ; le deuxième point est que tous les déplacements sont nécessairement enregistrés pour assurer le bon fonctionnement du système – cela permet des analyses exhaustives mais nécessite alors des méthodes d'analyse dédiées. Noter que sur ce dernier aspect on retrouve la même chose par exemple avec les systèmes de métro [LM02, RKBB11], ou avec le réseau de transport aérien qui a beaucoup été étudié comme un graphe complexe [GMTA05, CBBV06, CBBV07].

Nos travaux sur Vélo'v visent à la description et la caractérisation de la dynamique spatio-temporelle de ce système, avec un objectif plus lointain de modélisation. Le deuxième angle d'approche s'inscrit dans une démarche d'étude sociologique, où l'on ne cherche pas tant à saisir le fonctionnement du système VLS qu'à questionner les usagers, leurs motivations pour l'employer et les usages qu'ils révèlent – quitte à discuter aussi les non usagers.

Nos travaux actuels portent plutôt sur l'étude du système socio-technique VLS et de son fonctionnement mais, comme ce travail se poursuit maintenant dans le cadre du projet ANR Vel'innov (Programme « Sociétés Innovantes » 2012, de 02/2013 à 02/2016) dans lequel je coordonne un groupe à l'ENS de Lyon qui associe des chercheurs en signal, physique, sociologie et géographie, aux côtés d'informaticiens (LIRIS) et de spécialistes en transport (LET), le deuxième aspect sera abordé en perspective à la fin de la section.

L'analyse des données VLS dans le monde. Le système Cyclocity, dont Vélo'v est celui déployé a Lyon, est installé dans 67 villes en France – incluant Paris avec Velib' –, en Europe, en Australie et au Japon ; dans le monde, tous opérateurs inclus (publics ou industriels), on est passé de 5 programmes avec une flotte de 4 000 bicyclettes au début des années 2000, à 375 programmes avec une flotte totale de 236 000 bicyclettes en 2011 [P.11]. Il n'est donc pas surprenant de trouver dans la littérature de plus en plus d'études portant sur l'analyse de données de tels systèmes. En Europe, on trouvera par exemple des études de données de Vélib' [Gir08, NMHHB13, CO13, FMG12, FG13], Bicing à Barcelone [FNO08, FNO09], OYBike à Londres [LAC12], Bicikelj à Ljubljana [DM12], sans oublier Vélo'v à travers nos travaux et ceux de collègues [MR10, JROR10, NCB09, NCPB11, NCPB13]. Un

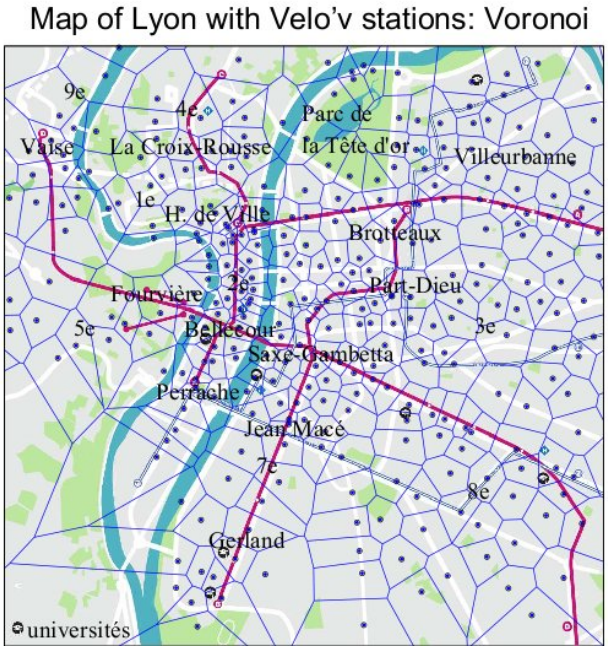


FIGURE 4.8 – Carte de Lyon où sont figurées les stations Vélo'v (points), leur diagramme de Voronoi (lignes bleues), les lignes de métro (lignes épaisses rouges), les cours d'eau (en bleu), et les parcs (en vert).

objectif partagé par une grande partie de ces études est de tenter de mesurer dans le temps l'activité de la ville (ou le « pouls de la cité » selon l'expression de [FNO08]) à travers les mouvements en VLS, ou les disponibilités en vélos ou attaches au niveau des stations.

Le cas lyonnais du Vélo'v. Le système Vélo'v, déployé à Lyon depuis mai 2005, est le premier système de location de VLS (vélos en libre service) automatisé de grande taille a avoir été mis en opération. Il permet aux usagers de louer des vélos (3000 fin 2007, 4000 disponibles aujourd'hui), à retirer à l'une quelconque des 334 stations distribuées dans la ville (voir la carte en figure 4.8) et à reposer à n'importe quelle autre station [Vél], soit avec à un abonnement à l'année ou avec des cartes de courte durée obtenues par une carte bancaire. À travers un partenariat impliquant le Grand Lyon et la filiale Cyclocity de la société JCDecaux, les empreintes numériques générées par chaque location de Vélo'V nous ont été rendues accessibles pour être analysées. Ces empreintes forment un enregistrement sur plusieurs années (2005-2008, 2011) de la dynamique du système Vélo'v. Les travaux discutés ici sont en quelque sorte des étapes préliminaires au travail de recherche que nous allons mener dans le projet ANR démarré il y a peu. Nous avons procédé en trois temps : (1) étudier un modèle statistique qui est non stationnaire pour les temps au-delà de la semaine, cyclique à l'échelle de la semaine et avec des corrections corrélées à l'échelle de quelques heures, et qui décrit bien la dynamique globale du réseau [J19, P35, P36] ; (2) utiliser avec profit des techniques d'analyse des réseaux complexes pour mettre en évidence les motifs (au sens de "patterns") fréquents du trafic et faire une typologie spatiale et temporelle du trafic [J19, C6, P37] ; (3) s'intéresser aux usagers des Vélo'v à travers une classification basée sur leurs déplacements. Les paragraphes qui suivent couvriront les deux premiers points.

4.3.2 Un système non stationnaire

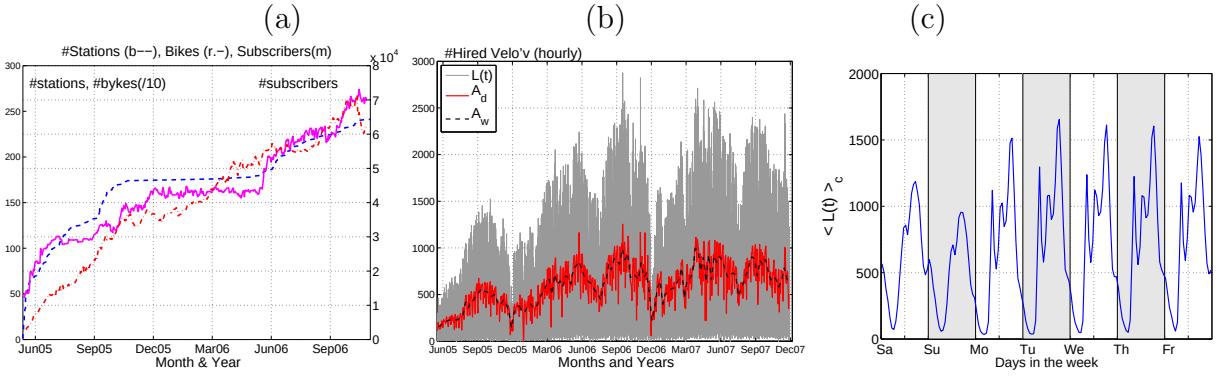


FIGURE 4.9 – Séries globales pour Vélo'v. (a) Évolutions du nombre de stations (ligne en tirets bleus), de vélos disponibles (ligne en traits mixtes rouge) et d'abonnés à l'année N_s (ligne pleine en magenta). (b) Location horaires $L(t)$ tracées heure par heure, $A_{\rm d}$ moyenne sur la journée et $A_{\rm w}$ sur la semaine du 17.XII.2005 au 14.XII.2007. (c) Moyenne cyclique sur la semaine $< L(t) >_c$ des locations horaires.

En Fig. 4.9 (b), on représente sur 2 ans (de décembre 2005 à décembre 2007) le nombre de locations horaires superposé à ce même nombre moyenné au jour et à la semaine. On constate un comportement non stationnaire de la moyenne dont l'interprétation est aisée : le système n'est pas dans un régime stationnaire au début. Au fil des mois le nombre d'utilisateurs, en particulier les abonnés, augmente et le nombre de stations existantes aussi, ainsi que montré en Fig. 4.9 (a). L'utilisation de Vélo'v dépend de plus du temps qu'il fait, donc de la saison. Une deuxième caractéristique est une modulation forte sur une journée, de période égale à une semaine qui ne fait que révéler les habituels 3 pics journaliers (matin, midi, soir) de semaine, bien connus en transport, avec un comportement différent le week-end et une intensité plus faible sans pic le matin.

Modèle statistique pour le nombre horaire de locations. Pour obtenir un modèle de ce nombre de locations, nous estimons un cycle moyen hebdomadaire, puis la tendance (multiplicative) non stationnaire plus lente et la prédisons en fonction des facteurs extérieurs ; enfin, nous étudierons les fluctuations plus rapide (à l'heure) par l'analyse des corrélations résiduelles. Le modèle pour le nombre de locations par heure $L_{\rm mod}(t)$ se formule ainsi :

$$L(t) = L_{\rm mod}(t) + F(t) = A_{\rm d}(d)\,\frac{< L(t) >_c}{A_{\rm mod}(d_7)} + F(t), \tag{4.6}$$

où $A_{\rm d}(d)$ est le nombre de locations au jour d :

$$A_{\rm d}(d) = \sum_{t\in(d)} L(t), \tag{4.7}$$

$A_{\rm mod}(d_7) = \sum_{t\in(d_7)} \langle L(t)\rangle_c$ est le nombre moyen de location au jour de la semaine d_7, où d_7 indique le jour de la semaine, de lundi à dimanche (d_7 est égal à d (la variable de jour)

Facteur	$\delta N_s(d)$	$\delta T(d)$	$R(d)$	$J_h(d)$
Unité	abonnés	°C	mm	
réf.	62 250	13.0	0.11	
std.	8 030	7.7	0.37	
coeff.	α_1	α_2	α_3	α_4
est.	1 860	2270	-1280	-2900
IC_-	1 210	1980	-1520	-3700
IC_+	2 560	2560	-1030	-2100

TABLE 4.2 – Modèle linéaire pour $A_{\rm d}(d)$, Éq. (4.9). On donne les facteurs en jeu (et leur unité), leur valeur de référence ($N_s(d)$ fin décembre 2007, $< T(d) >$; pour la pluie on donne $< R(d) >$ même si la référence choisie est 0) et leur écart quadratique moyen (std.). En-dessous, on trouve les coefficients obtenus par régression linéaire multi-variée : la valeur estimée (est.) et les intervalles de confiances [IC_-, IC_+] à 95 % (sous hypothèse gaussienne).

modulo 7). Enfin, la moyenne cyclique est estimée (en s'inspirant des techniques pour les signaux cyclostationnaires [GNP06]) par

$$\langle L(t)\rangle_c = \frac{1}{N_w}\sum_{k=0}^{N_w-1} L(t+k\,w_\Delta). \tag{4.8}$$

$F(t)$ est la déviation qui reste entre le modèle et le nombre de locations. La moyenne cyclique estimée est représentée en Fig. 4.9 (c) et le signal est classique d'un mode de transport employé en premier pour aller et revenir du travail : pic fin le matin, secondaire à midi et un peu plus large en fin d'après-midi, avec une baisse d'utilisation le soir et le week-end. Il est très comparable à ce qui est mesuré pour d'autres villes.

Prédiction de $A_{\rm d}(d)$. Nous écrivons les amplitudes journalières comme pouvant venir d'un modèle de régression linéaire prenant en compte les variables pertinentes du problème : les conditions météorologiques (déviation à la température moyenne $\delta T(d) = T(d) - < T(d) >$, en °C, et pluie $R(d)$ en mm), le nombre d'abonnés $N_s(d)$ (qui évolue très significativement sur les 2 premières années, voir Fig. 4.9 (b)) ainsi qu'un indicateur de vacances scolaires J_h car l'on constante une réduction des mouvements en Vélo'v à ces périodes. D'autres facteurs ont été envisagés qui sont peu significatifs. Le modèle est alors :

$$\widehat{A_{\rm d}}(d) \;=\; \alpha_0(d_7) + \alpha_1\delta N_a(d) + \alpha_2\delta T(d) + \alpha_3 R(d) + \alpha_4 J_h(d) \tag{4.9}$$

Le terme constant $\alpha_0(d_7)$est ajusté avec le jour d_7 qui ne rend compte que de la position du jour dans la semaine (de lundi à dimanche) :

$$\alpha_0(d_7) = A_0 + c_1\left(A_{\rm mod}(d_7) - < A_{\rm mod}(d_7) >_{d_7}\right). \tag{4.10}$$

Cette dépendance est nécessaire puisqu'on a vu en Fig. 4.9 (b) que le nombre de locations attendu un jour donné dépend de sa position dans la semaine ; il est plus faible par exemple le week-end. Par minimisation de l'écart quadratique aux données, on obtient pour les deux constantes $A_0 = 17370 \pm 320$ (en nombre de Vélo'v) et $c_1 = 1.05 \pm 0.14$ (coefficient

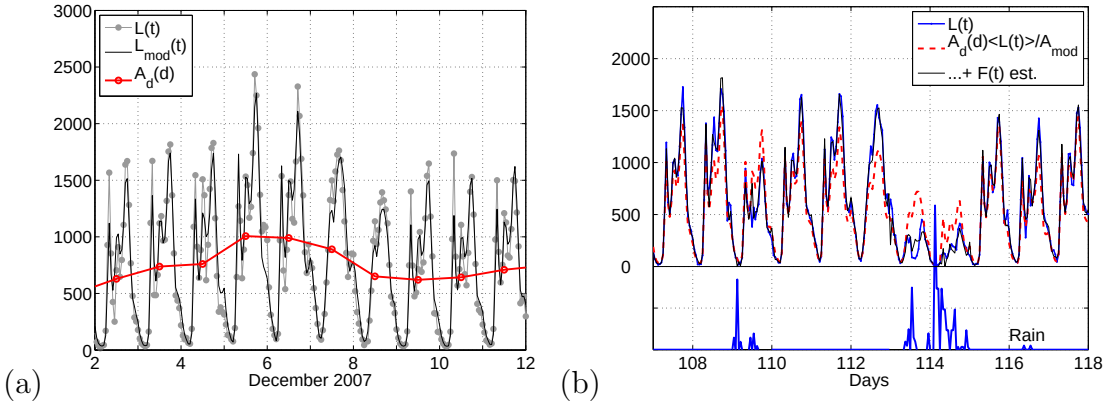


FIGURE 4.10 – Fluctuations du nombre de locations à l'échelle horaire : comparaison entre mesures et modèle. (a) Détail autour du 356e jour, le 8.XII.2006 (jour de la Fête des Lumières à Lyon) montrant que le modèle cyclique (voir éq. (4.6)) rend compte des évolutions usuelles, sauf quand une anomalie telle que le 8.XII apparaît dans la répétition cyclique. (b) Effet de la pluie (courbe en bleue du bas, échelle arbitraire) : on voit que les données réelles (ligne pleine) sont mieux approchées en cas de pluie par le modèle ARX(1) (trait fin) que par le modèle AR(1) sans la pluie (ligne pointillée).

sans dimension). Le terme A_0 est caractéristique du nombre moyen de locations dans une journée (à l'état de référence), soit en moyenne 725 vélos loués par heure. Le terme correctif selon le jour est linéaire (c_1 est estimée proche de 1) qui prend en compte les différences d'un jour sur l'autre dans la semaine type. Ensuite, les autres facteurs sont obtenus par régression linéaire multi-variée (en minimisant l'écart quadratique aux données) [DS98] et on indique dans la table 4.2 les valeurs estimées. Les résultats du modèle $\widehat{A_{\mathrm{d}}}(d)$ ont été comparé aux amplitudes journalières réelles et permettent déjà une prédiction avec seulement 12% d'erreur quadratique moyenne relative [J19].

Ces résultats nous éclairent donc sur la pertinence des facteurs explicatifs. Comme attendu, la température moyenne du jour joue en positif et la pluie en négatif sur le nombre de locations, tandis que le nombre d'abonnés agit comme facteur positif pour rendre compte de la croissance générale du système au début et que les vacances pèsent en négatif.

Analyse des déviations (ou fluctuations) $F(t)$. Une fois l'amplitude journalière et le cycle hebdomadaire estimés, il nous reste à étudier les corrélations résiduelles dans $F(t)$. Leur déviation standard est de 210 (en nombre de vélos loués par heure), la moyenne étant nulle. Ces fluctuations sont de plus corrélées d'une heure à la suivante. Dans un premier temps, un simple algorithme de Levinson (voir par exemple [Pri81]) permet d'estimer les paramètres d'un modèle AR pour $F(t)$: on trouve principalement un processus AR(1) avec un paramètre de l'ordre de 0.60 (et pas d'autres dépendances significatives à plus d'un pas de temps, d'où le modèle d'ordre 1). Cependant, l'analyse des données suggère que les pluies, étant des phénomènes parfois de courte durée, peuvent jouer beaucoup sur les fluctuations horaires. Une modèle de prédiction de F est donc pris comme un modèle ARX(1) [Pri81, Lju99], auto-régressif d'ordre 1 avec un facteur exogène, la pluie :

$$F(t) = a_1 F(t-1) + \beta_1 R(t) + I(t), \tag{4.11}$$

où a_1 est un coefficient d'AR(1), β_1 le coefficient de régression linéaire associé à la pluie dans l'heure $R(t)$ (en mm) et $I(t)$ une innovation. À nouveau le critère d'adéquation est l'erreur quadratique. L'estimation (aux moindres carrés) donne : $a_1 = 0.59 \pm 0.02$ et $\beta_1 = -40 \pm 4$ (en vélos/h/mm de pluie).

En combinant ce modèle pour F avec le modèle journalier, on arrive à réduire l'erreur quadratique moyenne pour la prédiction du nombre de locations à une heure donnée. La prédiction sans information horaire est : $A_{\mathrm{d}}(d) < \ L(t) \ >_c / A_{\mathrm{mod}}(d_7)$ et a une erreur de 210 locations de vélos. Avec l'éq. (4.11), on préconise l'estimée $\widehat{F(t)} = a_1 F(t-1) + \beta_1 R(t)$ qui réduit cette erreur quadratique moyenne de prédiction à 104 locations horaires.

On montre en Fig. 4.10 le résultat sur quelques jours de la prédiction du modèle, et en (b) celui de la prédiction corrigée par heure avec les observations passées et la pluie dans une situation où la pluie est importante. On voit l'amélioration apportée par la prise en compte des termes à l'heure, en particulier les jours de grande pluie.

Discussion. Si l'on cherche à déployer maintenant une telle analyse station par station, on se confronte à un écueil : individuellement, une station a en général assez peu d'activités. Beaucoup ont moins de 2 déplacement par heure en moyenne (tandis que la moyenne du nombre horaire de déplacement par station était entre 7 et 8). Le décompte faible fait que la stratégie précédente échoue à l'échelle de la station là où elle fonctionnait à l'échelle globale. De plus, le nombre de vélos disponible à une station donnée est un facteur important. Nous avons débuté dans [P46] un travail de régression parcimonieuse (de type lasso [Tib96, EHJT04]) sous contrainte pour tenir compte des capacités des stations mais, bien que la méthode fonctionne sur des données simulées, on se rend compte qu'il est encore difficile de modéliser l'activité à l'échelle de la station unique.

La suite de notre démarche est alors plutôt de trouver quelle échelle spatiale est pertinente pour étudier les mouvements en Vélo'v, ou, plus exactement : quels groupes de stations sont pertinents et seraient intéressants à faire pour une future modélisation.

4.3.3 Vélo'v comme réseau complexe

Le réseau complexe dynamique de Vélo'v. Pour cela, nous avons regardé les déplacements en Vélo'v comme une instance de réseau dynamique pondéré (et dirigé) entre les stations du système [P37, J19, C6]. Soit $\mathcal{V}$ l'ensemble des stations et $n \in \mathcal{V}$ une station qui sera un nœud du réseau. On pose $\mathcal{D} = \{(n, m, \tau)\}$ l'ensemble des trajets de la station $n \in \mathcal{V}$ vers $m \in \mathcal{V}$ au temps τ. Le réseau dynamique Vélo'v est défini comme $\mathcal{G} = (\mathcal{V}, \mathcal{E}, T)$ où l'ensemble des liens est $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ et T est une fonction définissant une matrice d'adjacence dynamique pondérée. Soit $\mathcal{T}$ contenant les instants t auxquels on s'intéresse, et $\mathcal{S}$ une collection de durées d'agregation Δ. La fonction $T : \mathcal{E} \times \mathcal{T} \times \mathcal{S} \to \mathbb{N}$ s'écrit formellement

$$T[n, m](t, \Delta) = \#\left\{(n, m, \tau) \in \mathcal{D} \mid t \leq \tau < t + \Delta\right\} \tag{4.12}$$

où $\#$ est le cardinal de chaque ensemble. Le résultat $T[n, m](t, \Delta)$ est la matrice d'adjacence pondérée du réseau Vélo'v. Le poids sur chaque lien est donc le nombre de vélos allant de la station n à station m entre t et $t + \Delta$.

Nous avons vu l'aspect cyclique sur la semaine des locations et il se transmet au réseau. T peut donc être estimé en faisant des moyennes périodiques sur une semaine. Pour le temps d'agrégation Δ, on sait qu'il est très important de bien le définir ou de choisir plusieurs

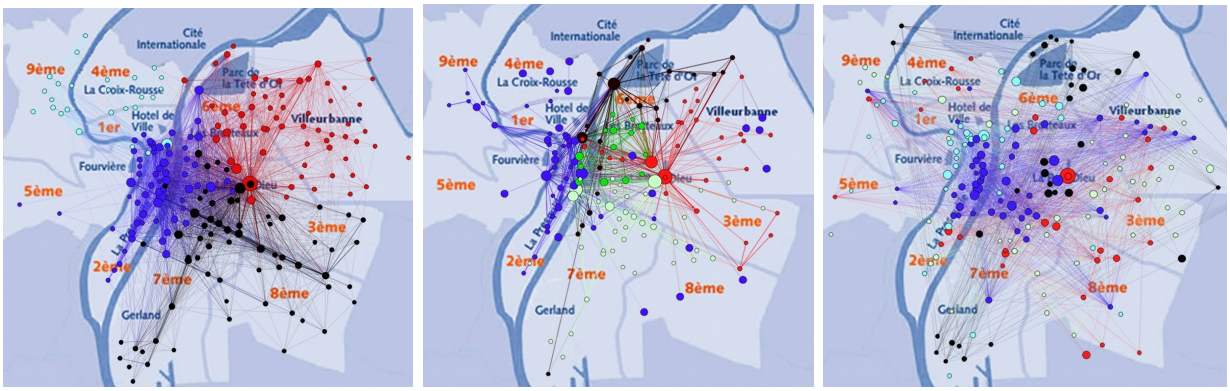


FIGURE 4.11 – Graphe des stations Vélo'v et des communautés dans le réseau Vélo'v. Chaque communauté est indiquée par une couleur et reflète un usage spécifique. À gauche, réseau agrégé global ; au milieu, réseau pour le samedi après-midi ; à droite, réseau pour typologie par similarités entre profils temporels. Les nœuds ont une taille proportionnelle au nombre de trajets effectués depuis ou à destination de cette station pendant la période couverte par un graphe donné.

fenêtres. Cela a souvent été remarqué, par exemple pour étudier le télétrafic Internet (voir le chapitre 3 ou [ABF+02]), les réseaux de communication [KKB+12], la diffusion d'information dans un réseau [MML11], ou encore l'étude des mouvements de troupeau [BBN+11], etc. Pour Vélo'v, la durée de location moyenne est inférieur à 30 minutes (du fait du mécanisme de prix) et prendre un Δ plus grand permet de lisser les comportements de trajets individuels. Nous prenons donc usuellement $\Delta = 1$h (comme dans les figures 4.9 (c) et 4.10), ou 2h ou 3h pour couvrir plus largement chaque pic de la journée.

Dans la suite, nous réduisons de plus la dimension de $T[n,m](t,\Delta)$ en ne gardant que les 19 pics principaux d'activité temporelle, ceux des jours ouvrés à 8h, 12h et 17h et ceux à 12h et 16h les week-ends. Ces 19 pics sont ici justifiés comme étant les moments principaux d'activité du système et ceux que d'autres études de transport regarderaient en premier. Néanmoins, une analyse en composante principale comme celle faite dans [J19] révèle que ces 19 pics sont justement ceux qui permettent de réduire la dimension de $T[n,m](t,\Delta)$ au sens de l'ACP.

Agrégation dans l'espace du réseau Vélo'v. Un premier résultat dans [J19, C6] fut de regarder quelles sont les communautés dans le réseau Vélo'v défini ci-dessus, au sens des communautés décrites dans 4.1.2. La méthode employée est celle de la recherche de communautés par optimisation de la modularité à l'aide de l'algorithme de Louvain [BGLL08]. La modularité dans ce cas est :

$$Q = \frac{1}{2W} \sum_{\{n,m\}\in\mathcal{V}\times\mathcal{V}} \left[T[n,m] - \frac{\sum_{j\neq n} T[j,n] \cdot \sum_{k\neq m} T[m,k]}{2W} \right] \delta_{c_n,c_m}, \tag{4.13}$$

où $W = \sum_{n,m} T[n,m]$ est le poids total du réseau et c_n la fonction de partition en groupe. Le niveau le plus haut de communauté est indiqué en figure 4.11 (a) pour le réseau agrégé dans le temps sur les 19 pics et on trouve avec un peu de surprise que les communautés ne se recouvrent quasiment pas en espace (la même chose est obtenue en [J19] pour des découpes plus fines en nombre de communautés si l'on utilise l'approche hiérarchique permise

par [BGLL08]). Ceci nous montre que les trajets ont une certaine distance préférée dans l'espace et qu'une première découpe en groupes de stations qui échangent le plus de vélos entre elles est déjà à trouver sur une base géographique, alors même qu'aucune information géographique n'a été mise à l'entrée de l'étude.

Si on applique la même recherche de communautés au réseau pris à des instants qui ne couvrent pas toute la semaine, le résultat change un peu, voir ici la carte donnée pour le samedi 4.11 (b). Pendant le week-end, les stations les plus actives changent : les zones de shopping (Presqu'île) et autour du grand Parc de la Tête d'Or au nord sont plus actives que dans la semaine, même si sont conservés les hubs principaux des transports publics (Part-Dieu, Vieux Lyon au bas de la colline de Fourvière, Hôtel de Ville,...). Les communautés se réorganisent donc et on trouve par exemple une communauté de stations qui échangent beaucoup de vélos entre elles du nord au sud, le long des pistes cyclables du Rhône. La partition est donc moins fondée sur la géographie et une distance typique des trajets. En semaine, les partitions du matin et de la fin d'après-midi sont en revanche assez semblables et proches de celle de la figure 4.11 (a) : ici s'illustre le fait que Vélo'v sert de manière régulière de transport pour aller à son travail et en revenir.

Communautés de stations selon leurs profils d'activités. Un deuxième choix est de transformer le réseau dynamique $T[n, m](t, \Delta)$ basé sur un décompte de trajet en un réseau de similarités des profils d'activité des stations. Le détail est donné dans [C6] et revient essentiellement à corréler les activités des stations au fil de la semaine. Faisant cela, on peut alors trouver cette fois les groupes de stations qui ont un comportement en communauté pour cette métrique de corrélation de manière à obtenir une classification des stations en fonction de leurs profils d'activités. Le résultat est affiché en figure 4.11 (c). L'analyse de cette carte est très différente des précédentes : plutôt qu'une découpe des stations en zones géographiques, c'est une structure qui s'articule autour d'un centre (communauté en bleu), de zones frontières au centre connectées à des points principaux de l'accès aux autres transports en commun (rouge, bleu ciel) et des périphéries (campus autour de Lyon en noir, zones d'habitation uniquement en blanc).

Discussion. La combinaison des différentes cartes nous indique qu'il faudra chercher, dans nos futurs travaux de modélisation du système Vélo'v, à décrire les stations à travers plusieurs caractéristiques : elles échangent plus facilement des vélos en réseau avec les stations qui ne sont pas trop éloignées (la distance rend les trajets moins fréquents, sûrement par concurrence des autres moyens de transport) ; leurs profils d'activités sont hétérogènes et dépendant de leur position en ville (centre vs. périphérie en particulier). Un autre élément sera alors à prendre en compte : les stations fonctionnent plus ou moins en groupe avec leurs voisines selon où elles se trouvent. La figure 4.12 montre quelles stations ont une activité qui se corrèle avec les 5 plus proches voisines plus que la moyenne, moins que la moyenne ou à peu près comme la moyenne des stations. Cette carte montre qu'une zone comme la Presqu'île au centre de Lyon a des stations qui vont sûrement bien fonctionner en groupe, comme celles autour des gares (Part-Dieu, Perrache ou Vaise) là où les stations plus périphériques et éloignées des métros sont peu corrélées avec leurs voisines. Tous ces aspects rendent la lecture des groupes de stations complexe et continue à faire l'objet de nos travaux.

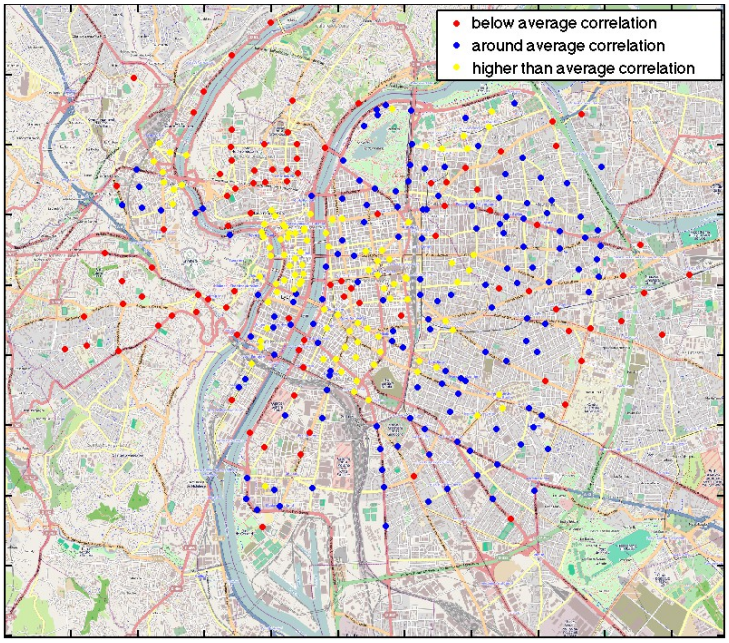


FIGURE 4.12 – Carte de la classification des stations en fonction de la corrélation avec leurs 5 plus proches stations (corrélation des probabilités de trouver une station vide). Les stations en jaune (clair) sont plus corrélées que la moyenne, celles en rouge (gris) moins, et en bleu (foncé) entre les deux.

4.3.4 Perspectives

Affiner l'analyse des VLS. Notre objectif est d'aller au-delà de la description empirique de ces données en les enrichissant, grâce à un travail avec L. Merchez (géographe de l'UMR 5600 « Environnement, Ville, Société » à l'ENS de Lyon), M. Vogel et I. Mallon (sociologues à l'ENS de Lyon, UMR 5283 « Centre Max Weber »), P. Jensen (Laboratoire de Physique et IXXI) et C. Raux (UMR 5593, Laboratoire d'Économie des Transports) par des données géographiques, démographiques, sociales et économiques qui permettent de caractériser l'environnement des stations de Vélo'V en décrivant à la fois les populations résidentes aux alentours et les espaces en termes d'emplois, de loisirs et de services. Formuler un modèle statistique de « demandes » en Vélo'v nous a orienté vers les études de régression parcimonieuses, en particulier les techniques de lasso. Nous avons commencé à voir dans [P46] comment les adapter aux données Vélo'v qui, en plus de leur aspect réseau, présentent des troncatures non-linéaires dans les relations demandes – trajets effectués (et donc mesurés). Mais l'on a vu dans 4.3.3 qu'il faut pouvoir travailler sur des groupes de stations et que définir comment les grouper n'est ni immédiat, ni achevé.

Remettre les utilisateurs dans le discours. Un deuxième aspect complique le travail pour formuler un tel modèle donnant le nombre de trajets effectués en fonction de l'environnement géographique, humain et économique des stations : les utilisateurs des VLS sont eux-mêmes très variés dans leur pratique. Un travail en cours [Js28] prend presque le contrepied des études menées jusqu'ici sur les VLS en s'intéressant non plus au système et aux trajets, mais aux utilisateurs et à leur pratique du Vélo'v. Bien entendu, ce questionnement

est plutôt conduit sous la direction des collègues sociologues et nous apportons les outils d'analyse de données pour dresser en particulier une typologie des utilisateurs de Vélo'v. Sur 2011, l'étude nous permet de grouper les 50000 abonnés longue durée actifs en quatre catégories principales à partir de l'intensité de leur pratique dans l'année et dans la semaine : les vélo'veurs extrêmes (en moyenne dans les 700 trajets par an, tout au long de l'année et la semaine), les utilisateurs réguliers (dans les 25 à 30 trajets en moyenne par mois, certains tout au long de l'année, d'autres pour quelques mois seulement, en général à la belle saison), les utilisateurs multi-modaux (50 trajets par an, espacés dans l'année, employés sûrement en complément d'autres modes de transport) et les utilisateurs rares, voire déçus (1 ou 2 trajets dans une année...). La typologie est construite simplement par classification sur des descripteurs d'intensité de la pratique dans la semaine et l'année. L'enseignement de ce travail préliminaire est qu'aux côtés de l'hétérogénéité des stations, il faudra prendre en compte l'hétérogénéité des comportements des humains qui utilisent le système Vélo'v.

4.4 Les graphes vus comme signaux

Traitement du signal sur ou pour des graphes. Faisant suite à ces travaux initiés dans le cadre d'applications spécifiques (système VLS, mesure de contacts humains) je développe maintenant des approches de traitement du signal sur ou pour des graphes. Nos résultats récents cherchent à promouvoir le traitement de signaux sur graphes en les croisant avec les questions (et aussi les méthodes) des études sur les réseaux complexes. Par exemple, nous avons considéré la détection de communautés dans [P59, P60, P63, P65] (et un article soumis [Js27]), voir 4.4.1, l'étude statistique de propriétés des sous-groupes dans un réseau [J24, P53], voir plus haut en 4.2.3, ou des manières d'étudier des réseaux dynamiques par des méthodes traitement du signal [P57, P62, P64] (et deux articles soumis [Js31, Ps67]). Nous cherchons donc par ces travaux à étudier **des graphes qui sont des signaux** et pas seulement des signaux sur une topologie de graphe fixée – mais la perspective est réellement de coupler les deux.

4.4.1 Détection multi-échelle des communautés dans les graphes avec des ondelettes

Positionnement de la contribution. Nous avons tracé rapidement l'état de l'art de la détection de communautés dans des réseaux en 4.1.2. Un défaut des différentes méthodes existantes est qu'elle ne peuvent conduire à des communautés à différentes échelles que par l'introduction de paramètres ad-hoc, voir par exemple [Pon06, KSKK07, RB06, AFG08, Lam10, SDYB12] pour les approches liées à la modularité. Elles sont d'ailleurs souvent limitées par la résolution minimale de cette technique [FB07]. De plus, il n'est pas aisé de trouver un test statistique pour savoir si une partition donnée est pertinente ou non (a contrario, voir les discussions dans [SDYB12]). Troisième défaut, dans les cas où l'on souhaite trouver des groupes pour ensuite travailler sur des signaux dans des réseaux (ce qui est par exemple intéressant pour des réseaux de capteurs ou des réseaux de communication ad hoc), l'analyse en communautés par les méthodes usuelles rappelées dans [For10] n'a rien en commun avec les traitements ultérieurs que permettraient les méthodes de traitement du signal sur graphes.

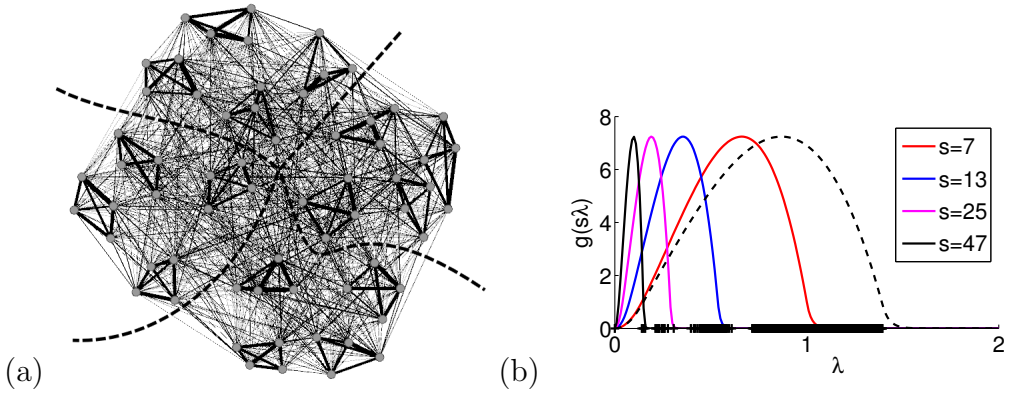


FIGURE 4.13 – (a) Une réalisation d'un graphe hiérarchique à trois niveaux de communautés (modèle de [SPGMA07]) avec $N = 640$ nœuds. Pour la clarté de la représentation, chaque communauté de 10 nœuds est dessinée comme un seul nœud. Les tirets marquent le découpage en quatre communautés à la plus grande échelle. ; la structure intermédiaire en 64 communautés est discernable visuellement. (b) Filtres passe-bande g de l'ondelette dilaté pour 4 paramètres d'échelle $s = 7\,(s_{min})$, 13, 25, and $47\,(s_{max})$. Les paramètres de g sont ceux dans [Js27] : $x_1 = 1$, $x_2 = 7$, $\alpha = 2$ and $\beta = 41$. Un 5e filtre est indiqué en pointillés ; il correspondant à une échelle plus petite que s_{min} et les expériences ont montré qu'une telle échelle n'est pas utile pour détecter les communautés. Sur l"abscisse, on a mis des croix là où sont les valeurs propres du réseau (ces filtres et valeurs propes ont été calculés pour le réseau de gauche).

Notre approche dans [P59, P60, P63, P65, Js27] vise à répondre à ces deux difficultés. Nous utilisons le cadre du traitement du signal sur graphes [SPM13] pour détecter des communautés dans des réseaux, ce qui nous permet d'introduire la notion d'échelle des communautés en utilisant des ondelettes sur graphes comme descripteurs des nœuds aux différentes échelles. Ensuite, notre travail montre comment la résolution d'un problème de classification non supervisée avec ces descripteurs permet de détecter efficacement les communautés à différentes échelles dans un réseau et de tester leur pertinence. Nous avons aussi relié en [P60] notre approche à celles développées à partir de la modularité, telles que [SDYB12, Lam10]

Transformées en ondelettes sur graphes. Depuis 2003, il y a plusieurs travaux définissant des ondelettes sur des graphes (ayant des propriétés compatibles avec celles discutées en 3.2), depuis les approches dans le domaine spatial des nœuds [CK03], ou par lifting [NO09] aux travaux dans un domaine spectral pour les graphes, par exemple par diffusion [CM06] ou à l'aide du spectre du Laplacien du graphe [HVG11]. Voir [SNF+13] pour une bibliographie plus complète à ce sujet. C'est cette dernière construction [HVG11] que nous utilisons, car le spectre du laplacien est déjà impliqué dans les méthodes spectrales de détection des communautés [For10].

Rappelons quelques notions d'analyse spectrale d'un graphe par son laplacien [Chu97, CL06, Mie11]. Soit $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ un graphe non dirigé et pondéré, N le cardinal de $\mathcal{V}$ et sa matrice d'adjacence $\mathbf{A}$ pondérée telle que $\mathbf{A}_{ij} = \mathbf{A}_{ji} > 0$ est le poids de l'arête entre les nœuds i et j, $\mathbf{A}_{ij} = 0$ si ils ne sont pas reliés et $\mathbf{A}_{ii} = 0$. Sa matrice laplacienne est $\mathbf{L} = \mathbf{D} - \mathbf{A}$ où $\mathbf{D}$ est une matrice diagonale avec $\mathbf{D}_{ii} = \mathbf{d}_i = \sum_{j \neq i} \mathbf{A}_{ij}$, la force du nœud i.

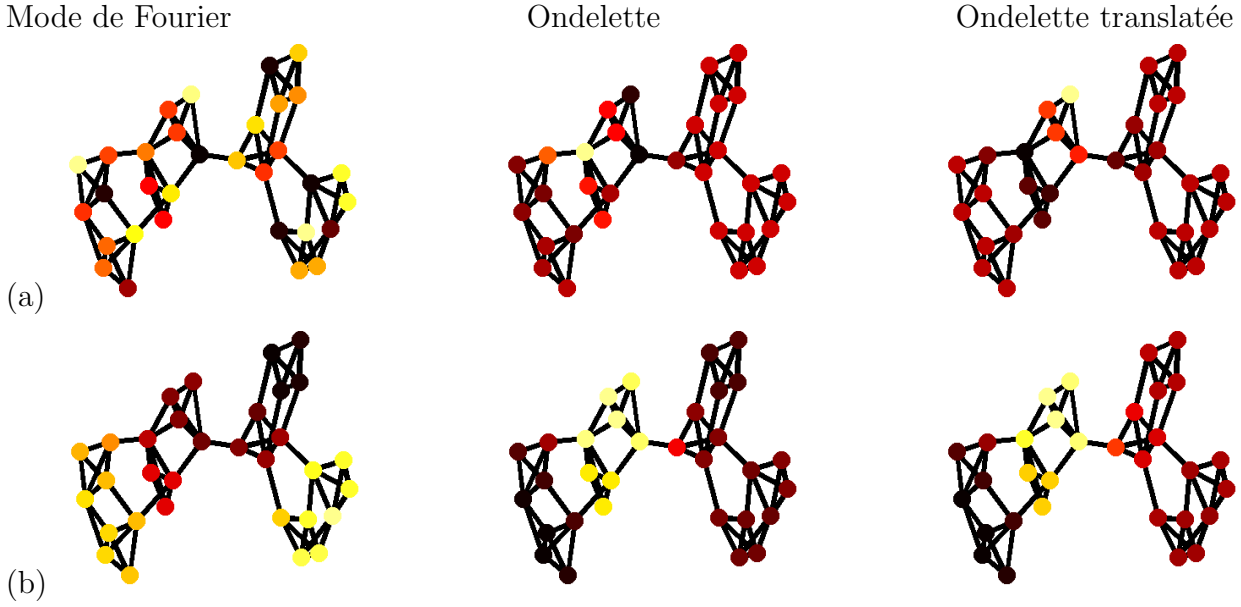


FIGURE 4.14 – Exemples de modes de Fourier et d'ondelettes sur un graphe simple. (a) À haute fréquence ($\sqrt{\lambda_i}$) pour Fourier et petite échelle s pour les ondelettes ; (b) à plus basse fréquence pour Fourier et plus grande échelle pour les ondelettes. Les valeurs des ondelettes sur les nœuds sont codées par les couleur, jaune pour les grandes valeurs, rouges pour les valeurs proches de 0, noires pour les valeurs négatives.

Pour notre problème, on emploie plutôt la matrice du laplacien normalisé :

$$\boldsymbol{\mathcal{L}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I}_N - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}, \tag{4.14}$$

où $\mathbf{I}_N$ est la matrice identité de taille N. $\boldsymbol{\mathcal{L}}$ est réel, symétrique et donc diagonalisable. Son spectre est l'ensemble des valeurs propres $(\lambda_l)_{l=1\ldots N}$, triées en ordre croissant : $0 = \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_N \leq 2$ [Chu97]. On note $\boldsymbol{\chi}$ la matrice des vecteurs propres normalisés : $\boldsymbol{\chi} = (\boldsymbol{\chi}_1|\boldsymbol{\chi}_2|\ldots|\boldsymbol{\chi}_N)$. On considérera ici seulement des graphes connectés, de telle sorte que $\lambda_1 = 0$ a une multiplicité de 1 [Chu97]. Il est possible de définir une transformée de Fourier sur graphes à l'aide de $\boldsymbol{\mathcal{L}}$ en procédant par analogie avec l'opérateur de Laplace 1D ou 2D usuel [SNF+13]. La transformée de Fourier d'un signal f défini sur le graphe est définie comme :

$$\hat{f} = \boldsymbol{\chi}^\top f. \tag{4.15}$$

Comme la transformée de Fourier usuelle, elle est inversible et offre donc un second domaine de représentation d'un signal, pour lequel on retrouve des fonctions de représentations qui peuvent êtres des modes lents, délocalisés dans le graphe, ou plus rapides mais délocalisés (voir figure 4.14), parfois des modes plus localisés proches de nœuds de forts degrés [NSW13]. Noter qu'il a été proposé d'autres notions de transformée de Fourier sur graphe [SM13] mais que nous suivons dans cette présentation [HVG11]. Il est alors simple de définir ce qu'est un filtre sur le graphe. Soit g un noyau de filtre ; il agit sur une fonction f dans le domaine de Fourier en multipliant chaque coefficient de sa décomposition de Fourier sur $\boldsymbol{\chi}_i$ par $g(\lambda_i)$.

Étant donné qu'on peut relire le terme [2] suivant : $\frac{1}{a}\psi\left(\frac{u-t}{a}\right) = \psi_{a,t}(u)$ de l'équation

2. Ici, on a remplacé $1/\sqrt{a}$ par $1/a$ ce qui change seulement l'espace fonctionnel de définition de la

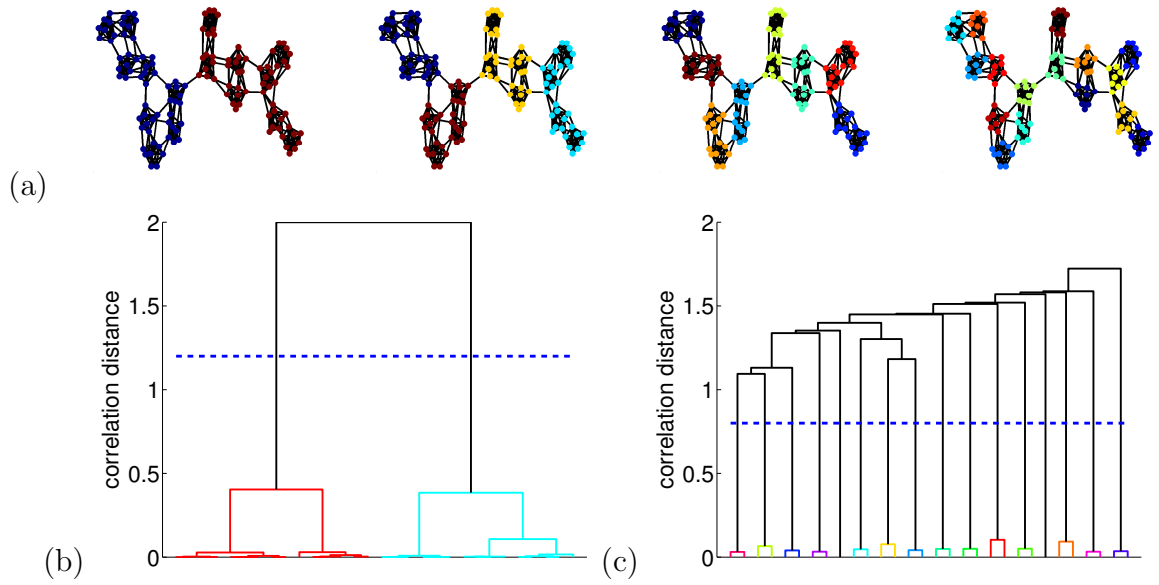


FIGURE 4.15 – (a) Exemple de graphe avec des communautés à plusieurs échelles (2, 4, 8 ou 16 communautés) et dendrogrammes associés : en (b) dendrogramme obtenu à large échelle, en (c) à petite échelle, avec le plus grand saut associé qui coupe (selon la droite en pointillés) en 2 (en b) ou 16 (en c) communautés

(3.1) définissant ce qu'est une transformée en ondelettes continue, en passant dans le domaine de Fourier pour avoir :

$$\psi_{a,t}(u) \;=\; \int_{-\infty}^{\infty} \hat{\delta}_t(\omega)\hat{\psi}(a\omega)\exp^{i\omega u}\, d\omega, \tag{4.16}$$

il est aisé de définir une transformée en ondelettes sur graphe en mettant dans cette équation le filtrage sur graphe dans le domaine de Fourier [HVG11]. Pour ce faire, soit g un noyau de filtre passe-bande qui sera étiré par le paramètre d'échelle $s > 0$. Nous notons sa représentation matricielle à l'échelle s : $\hat{G}_s = \mathrm{diag}(g(s\lambda_0), \ldots, g(s\lambda_{N-1}))$. Sur un graphe un peu complexe [SPGMA07], on montre en figure 4.13 comment se placent les filtres $g(s\cdot)$ étirés par l'échelle. La base des ondelettes à cette échelle s'écrit pour le graphe :

$$\Psi_s = (\psi_{s,0}|\psi_{s,1}|\ldots|\psi_{s,N-1}) = \boldsymbol{\chi}\hat{G}_s\boldsymbol{\chi}^\top, \tag{4.17}$$

où $\psi_{s,v}$ est l'ondelette centrée autour du nœud v. Contrairement aux modes de Fourier du graphe (vecteurs propres du laplacien), les ondelettes sont localisées [HVG11]. Une illustration sur un graphe simple en est donnée en figure 4.14 : on y voit des ondelettes centrées autour de deux nœuds différents, à deux échelles différentes (grande échelle en s qui correspond à des fréquences basses et la petite échelle à des fréquences hautes).

Trouver des communautés par clustering sur les ondelettes. La figure 4.14 précédente nous illustre aussi la base de la méthode développée : quand l'échelle est de la taille

transformée en ondelettes ; à nouveau, nous suivons dans chacun des chapitres les conventions employées par les travaux sur lesquels nous nous appuyons.

d'une communauté (ligne du bas), les ondelettes de 2 nœuds d'une même communauté sont quasiment identiques. En effet, l'ondelette centrée autour d'un nœud est une signature de l'environnement topologique de ce nœud. Si deux nœuds ont un voisinage similaire à l'échelle de résolution de l'ondelette, alors leurs ondelettes associées sont corrélées et ils ont d'autant plus de raisons de se retrouver dans la même communauté à cette échelle. Une ondelette fournit en fait un vue égo-centrée du réseau, autour du nœud où elle est centrée. La méthode que nous proposons part de la matrice de corrélation entre les ondelettes, mise en entrée d'un algorithme de partitionnement hiérarchique [HTF+01] qui réalise une classification des nœuds sur cette base et propose ainsi un regroupement en communautés. Les étapes de la méthode sont les suivantes. Pour un ensemble d'échelles d'intérêt $S = \{s_1 = s_{min}, s_2, \ldots, s_M = s_{max}\}$, on répète pour chaque échelle s :

1. **Calcul des vecteurs caractéristiques.** À chaque nœud v est associé l'ondelette centrée en v, $\boldsymbol{\psi}_{s,v}$, comme vecteur caractéristique. Notons qu'on peut employer aussi une fonction d'échelle (au sens des ondelettes continues) que nous avons définie en [P59, P60] mais nous de détaillerons pas cela ici, les résultats étant assez semblables avec les ondelettes.
2. **Distance entre deux nœuds.** La distance entre deux nœuds est la distance de corrélation entre leurs ondelettes :

$$\boldsymbol{D}_s(u,v) = 1 - (\widetilde{\boldsymbol{\psi}}_{s,u})^\top \widetilde{\boldsymbol{\psi}}_{s,v}, \tag{4.18}$$

   où l'ondelette est normalisée par $\widetilde{\boldsymbol{\psi}}_{s,u} = \frac{\boldsymbol{\psi}_{s,u}}{||\boldsymbol{\psi}_{s,u}||_2}$. Noter que l'ondelette est par constuction de moyenne nulle car la moyenne pertinente d'un signal $\boldsymbol{f}$ sur le graphe analysé par le laplacien normalisé est :

$$\bar{\boldsymbol{f}} = \boldsymbol{\chi}_1^\top \boldsymbol{f} = \frac{1}{\sqrt{\sum_i \boldsymbol{d}_i}} \sum_{i=1}^{N} \sqrt{\boldsymbol{d}_i} \boldsymbol{f}(i). \tag{4.19}$$

   Or $\forall (s,v) \quad \bar{\boldsymbol{\psi}}_{s,v} = \boldsymbol{\chi}_1^\top \boldsymbol{\psi}_{s,v} = \boldsymbol{\chi}_1^\top \boldsymbol{\chi} \boldsymbol{G}_s \boldsymbol{\chi}^\top \boldsymbol{\delta}_v = g_s(1) \boldsymbol{\chi}_1(v)$ puisque $\boldsymbol{\chi}$ est une matrice orthogonale (les vecteurs propres forment une base). La composante constante $g_s(1)$ étant nulle par définition du filtre des ondelettes (condition d'admissibilité comme en 3.2) on obtient que $\bar{\boldsymbol{\psi}}_{s,v} = 0$.
3. **Algorithme de partitionnement.** Nous utilisons un algorithme de partitionnement hiérarchique avec la méthode de chaînage moyenné (*average linkage* [HTF+01]). Sur des exemples simples, utiliser cette méthode plutôt qu'un chaînage complet (*complete linkage*) ou simple (*single linkage*) ne change rien mais nous préférons utiliser de base la méthode de chaînage moyenné qui est réputée plus stable. Cet algorithme produit un dendrogramme, comme il est illustré en figure 4.15, dont la forme dépend fortement de l'échelle d'analyse.
4. **Choisir où couper le dendrogramme.** Chaque coupe de ce dendrogramme définit une partition possible en communautés ; possible ne veut pas dire pertinent et on doit décider quelle coupe est la meilleure. Inspirés par l'approche des statistiques des sauts (cf. *gap statistics* [HTF+01]), le dendrogramme sera coupé au niveau du plus grand saut entre deux de ses nœuds. Cette idée est connue mais nécessite pour être justifiée un test statistique. Ici, la justification est donnée par la notion de stabilité de la partition associée que nous discutons dans la suite.

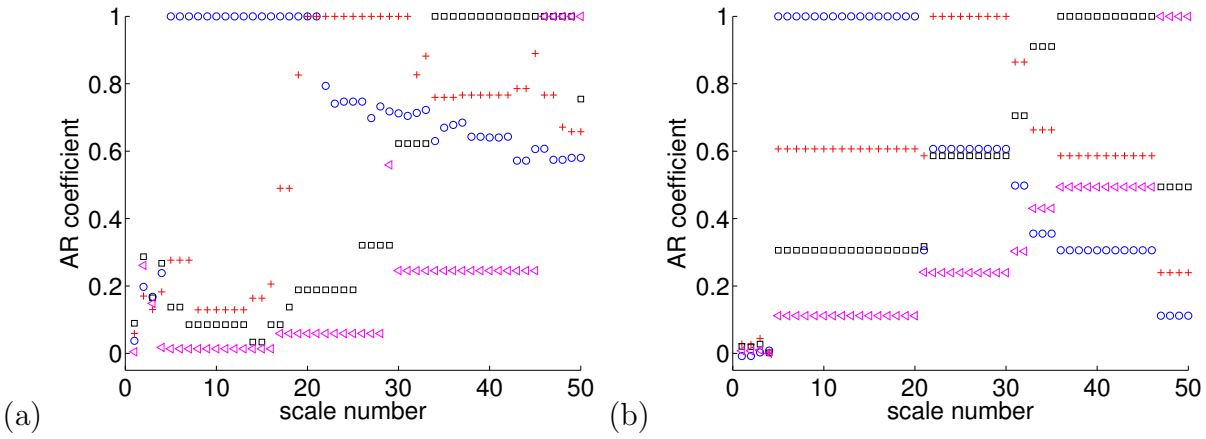


FIGURE 4.16 – Comparaison entre (a) ce qui est trouvé en coupant aux nombres exacts de communautés à différentes échelles (2, 4, 8, 16) et (b) ce qu'on trouve effectivement en coupant au saut le plus grand dans chaque dendrogramme. Chaque couleur compare à une découpe en communautés théorique existante dans le graphe : en 2 communautés en rose, 4 en noir, 8 en rouge et 16 en bleu (le graphe de l'exemple est celui de la figure 4.15). La métrique du résultat est l'indice de Rand ajusté [HA85] (1 indiquant une estimation parfaite des communautés).

Noter que nous avons étudié comment choisir la fonction de filtre passe-bande g définissant les ondelettes (illustrée en 4.13), en gardant à l'esprit que les méthodes de *clustering spectral* qui s'appuient sur $\boldsymbol{\chi}_2$ (vecteur de Fiedler) ne sont pas si mauvaises que cela et qu'on veut donc lui donner un rôle particulier. Voir la discussion dans [P60] pour les paramètres fixés pour g et pour les limites en échelle s_{min} et s_{max} ; entre ces deux limites, les échelles sont espacées logarithmiquement avec typiquement K échelles par voie comme pour les ondelettes continues classiques et on a donc $M = K \log_2(N)$ (K est typiquement inférieur à 10).

La figure 4.15 montre cela pour 2 échelles, qui trouve ici une découpe en 2 communautés ou en 16 communautés, toutes les deux justes là où une optimisation de modularité trouverait uniquement la partition en 8 communautés. En figure 4.16 on quantifie en terme d'indice de Rand ajusté [HA85] [3] ce que la méthode permet de retrouver. On voit que les dendrogrammes contiennent bien les 4 structurations en communauté et qu'on voit préférentiellement l'une ou l'autre en fonction de l'échelle : plus elle augmente, plus on voit les communautés larges (donc découpe en peu de groupes). La méthode de coupe du dendrogramme au plus grand saut semble aussi effective : le résultat (b) est quasiment celui qu'on obtiendrait en utilisant en (a) la connaissance a priori du nombre de communautés. Des validations plus détaillées de la méthode sont présentées dans [Js27] sur le graphe étalon de Sales-Pardo [SPGMA07] (aussi utilisé dans [Lam10]) et sur le modèle de graphe de [LF09], proposant des modèles de réseaux hiérarchiques plus réalistes que le modèle simple des figure 4.15 et 4.16 de ce mémoire.

Passer à de plus grands graphes. Le problème de notre approche est de devoir calculer toutes les fonctions d'ondelettes $\boldsymbol{\psi}_{s,v}$, pour tous les N nœuds v du réseau et pour les M

3. L'indice de Rand ajusté compare deux partitions et ramène cela à un indice entre 0 et 1, en l'ajustant pour que l'indice soit à 0 si la comparaison ne fait pas mieux que le recouvrement de 2 partitions aléatoires de même cardinalité. Sa définition précise se trouve en annexe B de [Js27].

échelles, avant de calculer la matrice des distances. Cela est coûteux algorithmiquement. Dans [P65, Js27], une approximation rapide de D est discutée qui ne nécessite que de calculer la transformée en ondelettes de η signaux aléatoires pris sur le réseau. Comme on peut employer la méthode de [SVF11, HVG11] pour calculer approximativement la transformée en ondelettes sur graphes de n'importe quelle fonction sans diagonaliser explicitement le laplacien $\boldsymbol{\mathcal{L}}$, et qu'on montre qu'on peut prendre η bien plus petit que N, le gain en temps de calcul est important. Les vecteurs caractéristiques qui remplacent les $\boldsymbol{\psi}_{s,v}$ est pris comme :

$$\boldsymbol{f}_{s,v} = \left(\boldsymbol{\psi}_{s,v}^{\top}\boldsymbol{R}\right)^{\top} \qquad (\text{avec } \boldsymbol{f}_{s,v} \in \mathbb{R}^{\eta}), \tag{4.20}$$

où $\boldsymbol{R} = (\boldsymbol{r}_1|\boldsymbol{r}_2|\ldots|\boldsymbol{r}_\eta) \in \mathbb{R}^{N\times\eta}$, matrice qui contient des colonnes $\boldsymbol{r}_i$ qui sont des réalisations i.i.d. du vecteur aléatoire $\boldsymbol{r} \in \mathbb{R}^N$ dont les composantes sur chaque nœud sont i.i.d. et selon un distribution aléatoire gaussienne, de moyenne nulle et de variance σ^2. Chaque colonne de $\boldsymbol{f}_{s,v}$ se calcule en appliquant la transformée en ondelettes rapide de [SVF11, HVG11]. Il est ensuite aisé de montrer qu'on peut approcher la matrice de distance par la corrélation empirique entre ces vecteurs caractéristiques. Cette corrélation empirique est :

$$\hat{C}_{uv,\eta} = \frac{(\boldsymbol{f}_{s,u} - \bar{\boldsymbol{f}}_{s,u})^{\top}(\boldsymbol{f}_{s,v} - \bar{\boldsymbol{f}}_{s,v})}{||\boldsymbol{f}_{s,u} - \bar{\boldsymbol{f}}_{s,u}||_2||\boldsymbol{f}_{s,v} - \bar{\boldsymbol{f}}_{s,v}||_2}, \tag{4.21}$$

et sa limite quand le nombre de vecteurs η augmente est :

$$\lim_{\eta\to+\infty} \hat{C}_{uv,\eta} = \mathrm{Cor}(F_{s,u}, F_{s,v}) = \frac{\boldsymbol{\psi}_{s,u}^{\top}\boldsymbol{\psi}_{s,v}}{||\boldsymbol{\psi}_{s,u}||_2||\boldsymbol{\psi}_{s,v}||_2} = 1 - \boldsymbol{D}_s(u,v). \tag{4.22}$$

On remplace donc dans l'étape **2** de la méthode la distance par $1 - \hat{C}_{uv,\eta}$. En pratique, on a par exemple une bonne approximation dès que $\eta = 30$ pour un graphe de $N = 640$ et le temps de calcul est ainsi réduit. Cela permet d'étudier des réseaux de quelques milliers ou dizaines de milliers de nœuds sans difficulté ; par exemple, pour $N = 6400$ et $M = 50$, la méthode implémentée en Matlab rend un résultat en 8 minutes sur un ordinateur portable Intel i7 Core@2.6GHz avec 8GB de RAM. On est loin des méthodes gloutonnes (mais sous-optimales) employant l'approche à la Louvain de [BGLL08] pour la modularité, mais il est tout de même possible d'aborder des réseaux de taille suffisante pour bien des applications. Le temps de calcul a par exemple été trouvé comparable aux méthodes multi-résolution de [SDYB12].

Stabilité des partitions multi-échelles et test statistique. La méthode propose une partition en communautés qui est la meilleure pour chaque paramètre d'échelle. Elle ne nous dit pas dans un premier temps si cette partition est à retenir ou si elle n'est pas pertinente ; on voit déjà sur 4.16 que quelques échelles ne donnent aucune des partitions exactes (mais en fait un mélange de communautés à un niveau de résolution et à un autre). Pour ne garder que les partitions pertinentes, nous avons avancé plusieurs notions de stabilité des partitions dans [P59, P63], certaines appuyées sur les travaux [Lam10].

Finalement, la méthode pour les grands réseaux étant la plus efficace algorithmiquement, nous avons proposé une nouvelle mesure de stabilité qui est ancrée dans la diversité intrinsèque à cette méthode. On considère J estimation des communautés, à partir de J matrice aléatoires $\boldsymbol{R}$. Les partitions associées $\{P_s^j\}_{j\in J}$ vont être légèrement différentes du fait de l'approximation en prenant $\eta \ll N$ et une mesure de stabilité $\gamma_a(s)$ des partitions à

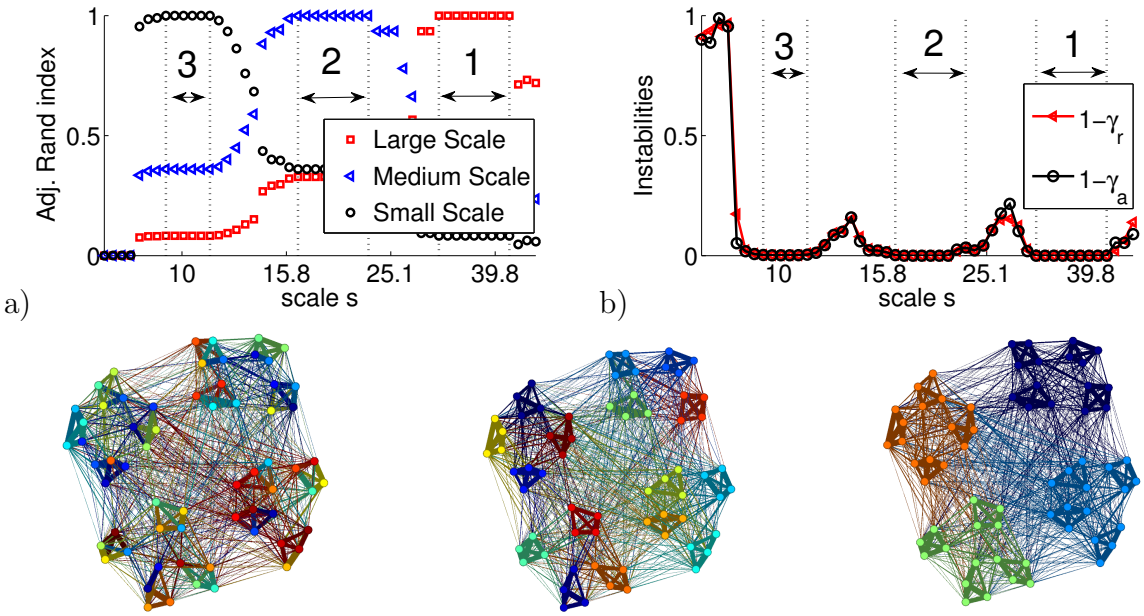


FIGURE 4.17 – a) Résultat de la méthode de détection multi-échelle de communautés sur une réalisation du graphe de [SPGMA07], voir fig. 4.13 (paramètres $\rho = 1$ et $\bar{k} = 16$) [Js27]. On a pris $M = 50$ échelles entre s_{min} et s_{max} et les corrélations entre ondelettes sont estimées avec $\eta = 60$ vecteurs aléatoires. Une partition en communautés est trouvée à chaque échelle et nous traçons sa similarité avec les structurations théoriques en communautés à petite, moyenne et grande échelle. Les intervalles 1, 2 et 3 représentent les zones en échelle où l'on trouve exactement les communautés théoriques. b) Instabilités $1 - \gamma_r$ (et $1 - \gamma_a$ de [Lam10]) en fonction de l'échelle s. Les 3 intervalles sont les mêmes qu'en a) ; les partitions associées à des instabilités minimales sont celles à retenir et correspondent bien aux communautés théoriques. c) Représentation du réseau coloré en fonction des 3 structurations en communautés trouvées (les 3 minima locaux des intervalles figurés).

une échelle s est prise comme la moyenne d'une similarité entre toutes les paires de partitions à cette échelle :

$$\gamma_a(s) = \frac{2}{J(J-1)} \sum_{(i,j)\in J, i\neq j} \texttt{simi}(P_s^i, P_s^j). \tag{4.23}$$

Si la partition est très stable, les $\{P_s^j\}_{j\in J}$ vont être proches et $\gamma_a(s)$ près de 1. Sinon, $\gamma_a(s)$ est proche de 0. La mesure de similarité sera à nouveau l'indice de Rand ajusté. Pour J, on prend typiquement 20. Comme dans [Lam10], on retiendra que les partitions les plus stables (maximum locaux). Dans [Js27], le travail de détection des communautés pertinentes est poursuivi en mettant en place un test statistique par comparaison avec un modèle de graphe ré-échantillonné à partir du réseau initial, de manière à donner un seuil en $\gamma_a(s)$ au-dessous duquel les partitions trouvées sont instables et ne sont donc pas à considérer. Il reste alors des intervalles où l'on trouve une structure en communautés stable et l'on garde par exemple le maximum local de chaque intervalle.

Exemple sur un graphe modèle. Le résultat de la procédure et du calcul de la stabilité (tracée ici en tant qu'instabilitée $1 - \gamma_a(s)$) sur un graphe suivant le modèle de [SPGMA07] est en figure 4.17. La procédure trouve bien les communautés sur les trois plages d'échelles où elles existent et la stabilité pointe bien vers trois découpes en communautés à retenir.

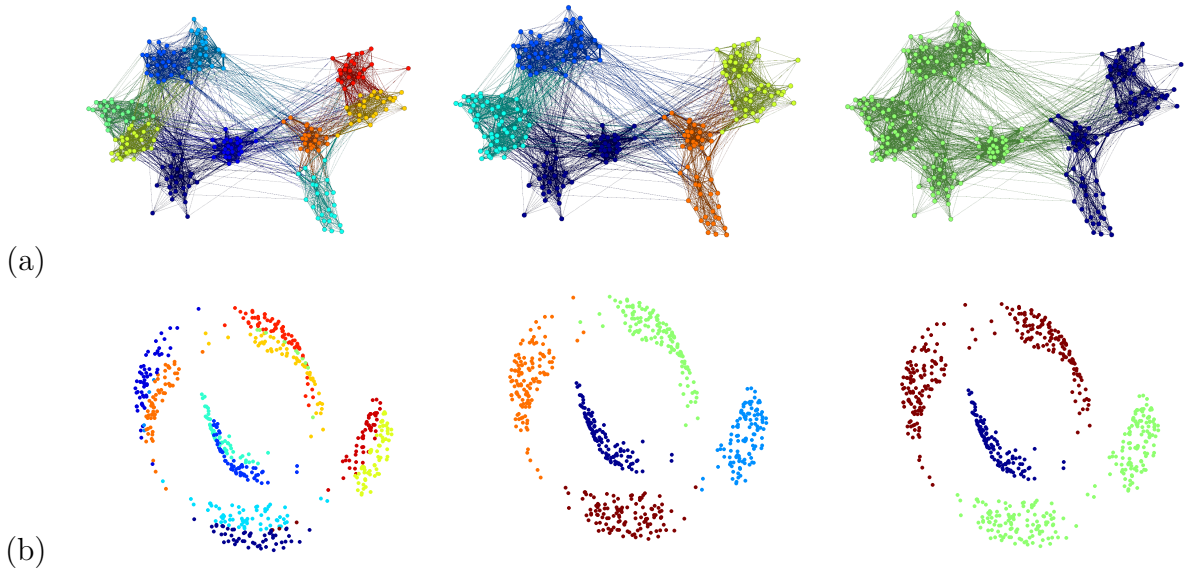


FIGURE 4.18 – Exemples de communautés détectées par l'approche multi-échelle sur des graphes de terrain ; chaque couleur est une communauté. En (a), interactions sociales entre des enfants dans une école primaire (où l'on trouve 10, 5 et 2 communautés reflétant les structures en 10 classes de 5 niveaux avec des grands (4 classes) et des petits (les 6 autres classes)). En (b), trois partitions de la variété non linéaire dite "Swiss roll", échantillonnée de façon non uniforme par des points (le réseau est un graphe de voisinage entre les points), en 13, 5, et 3 communautés.

Exemples sur des graphes de terrain. La méthode est en cours d'applications à d'autres données représentées sous forme de réseaux complexes, ou graphes de terrain. Les dessins des communautés trouvées pour deux graphes étudiés dans [Js27] sont en figure 4.18 et on trouvera dans ce travail plus de détails sur ces résultats. Ils ouvrent déjà la perspective de revenir sur les questions de communautés dans les réseaux de contacts entre humains étudiées de 4.2 puisque le premier réseau est celui des mesures Sociopatterns de [SVB+11], et vers l'étude des réseaux de capteurs (avant de se pencher sur les signaux dessus) puisque le deuxième est une situation modèle de points échantillonnant une variété un peu compliquée dans l'espace [HVG11]. L'intérêt de cette dernière découpe n'est pas vraiment pour l'analyse mais par exemple pour faire des estimations ou détection partielles dans chaque groupe avant de combiner les résultats, ce qui fait un pas vers des protocoles distribués.

Compléments et perspectives. La méthode multi-échelle de détection de communautés dans des graphes à partir des ondelettes vient s'inscrire dans un panorama déjà large de méthodes proposant de trouver des communautés, certaines dans une optique multi-résolution. De nombreuses comparaisons ont donc été nécessaires pour montrer que la méthode se compare très bien à l'état de l'art sur cette dernière question, en particulier [Lam10, SDYB12], ce qui est le cas. Nous avions d'ailleurs d'abord formulé dans [P60] notre méthode en utilisant une version filtrée en échelle de la modularité, proche de ce que propose les approches par marche aléatoires de [Lam10, SDYB12]. Le résultat est qu'on trouvait le même résultat qu'avec la technique de coupure au plus grand saut, mais avec un coût algorithmique plus élevé. Reste que ce passage par des modularités filtrées à une échelle a deux intérêts potentiels : il montre que la méthode proposée n'est pas étrangère

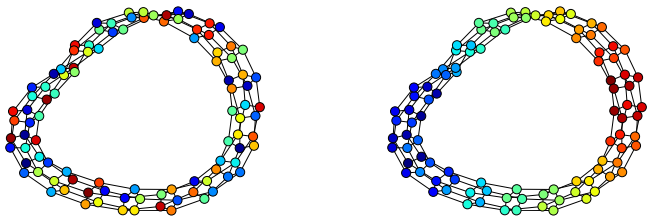


FIGURE 4.19 – Réseau en grille 2D de 5 nœuds par 20, chacun connecté à ses voisins et avec des conditions limites périodiques. (a) couleurs des labels pris aléatoirement sur le graphe ; (b) couleurs des labels après bon étiquetage par le méthode proposée pour minimiser le *cyclic bandwidth*. L'étiquetage est ainsi régulier et les couleurs suivent la structure du graphe.

aux travaux de recherche de communautés par optimisation d'une modularité ; il ouvre aussi la perspective d'employer peut-être un jour des algorithmes gloutons à la méthode de Louvain [BGLL08] pour passer des graphes à quelques dizaines de milliers de nœuds que nous pouvons étudier à des graphes de centaines de milliers de nœuds ou au-delà.

4.4.2 Les réseaux non stationnaires étudiés comme signaux

Le lecteur aura noté que nous avons pris avantage de nos connaissance sur les méthodes multi-résolution déjà rencontrées dans le chapitre 3 dans la section précédente. Il n'est inattendu de voir maintenant le point de vue non stationnaire du chapitre 2 revenir s'adjoindre à l'étude des réseaux complexes.

Dans les situations dynamiques, nous développons une approche originale consistant à transformer des graphes en une collection de séries qui décrivent ce graphe, puis à faire l'analyse spectrale de ces séries. Ces caractéristiques spectrales révèlent des éléments de la structure du graphe (présence de chemins, de communautés, structuration très déterministe ou aléatoire) et permettent ensuite de représenter et étudier dans le temps des réseaux dynamiques dans une optique "temps-fréquence" [P57, P62, P64] (et deux articles soumis [Ps67, Js31]).

Principe de transformation d'un graphe statique en signaux. Shimada et al. [SIS12] ont proposé une méthode pour transformer un graphe à N nœuds en séries à N points indexées par les sommets en utilisant le positionnement multidimensionnel classique (*classical multidimensional scaling* ou *CMDS*) [BG05]. L'intérêt d'une telle transformation réside dans le fait que si les N signaux permettent de reconstruire exactement le graphe d'origine, il est également possible de réduire la représentation en ne conservant qu'un nombre réduit de signaux. On part d'une matrice de distance entre nœuds d'un graphe notée $\mathbf{\Delta} = (\delta_{ij})_{i,j\in 1,..N}$ et définie pour $i, j \in \mathcal{V}$ par

$$\delta_{ij} = \begin{cases} 0 & \text{si } i = j \\ 1 & \text{si } A_{ij} = 1 \text{ et } i \neq j \\ w > 1 & \text{si } A_{ij} = 0 \text{ et } i \neq j \end{cases} \tag{4.24}$$

avec $w > 1$ et A qui est la matrice d'adjacence du graphe. En suivant [SIS12] on prendra $w = 1.1$ par la suite.

Le *CMDS* est une technique permettant d'obtenir les coordonnées de points dans un espace euclidien dont on ne connaît que les distances entre chaque paire de points. La matrice $\mathbf{X}$ des coordonnées peut être calculée analytiquement : on effectue un double centrage de la matrice $\mathbf{\Delta}$ dont les termes sont élevés au carré : $\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{\Delta}^{(2)}\mathbf{J}$ avec $\mathbf{J} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ où $\mathbf{I}_N$ est la matrice identité de taille N par N et $\mathbf{1}_N$ est un vecteur colonne de taille N rempli de 1. La solution du *CMDS* est donnée par $\mathbf{X} = \mathbf{Q}_+\mathbf{\Lambda}_+^{\frac{1}{2}}$ avec $\mathbf{\Lambda}_+$ une matrice diagonale dont les termes sont les valeurs propres positives triées par ordre décroissant de la matrice $\mathbf{B}$ et $\mathbf{Q}_+$ la matrice des vecteurs propres correspondants. Les signaux caractérisant le graphe sont les composantes de la matrice $\mathbf{X}$ avec $\mathbf{X}^{(j)}$ j-ème composante de $\mathbf{X}$. Elles sont indexées par les nœuds. Noter qu'il existe une autre méthode permettant de transformer des signaux en graphes et inversement, développée dans [CSM+11] mais elle concerne principalement le sens signal vers graphe, même si depuis le sens graphe vers signal a été repris et amélioré par nos collègues voisins de l'équipe DANTE du LIP [GGF14].

Du bon étiquetage du graphe. L'étiquetage d'un graphe (*graph labeling*) consiste à attribuer à chaque sommet d'un graphe un entier naturel compris entre 0 et $N-1$ en vertu d'un critère à minimiser [Chu88]. L'étiquetage des nœuds a une importance cruciale pour la méthode de transformation puisque deux nœuds très proches dans la numérotation mais distants dans le graphe vont avoir des valeurs sur chaque composante éloignées. Pour éviter des variations brusques du signal, nous avons proposé une méthode d'étiquetage qui résout empiriquement le *cyclic bandwidth problem*. Il cherche à minimiser la somme des distances dans la numérotation pour chaque paire de sommets voisins, cette distance étant définie pour deux sommets voisins numérotés i et j par $d(i,j) = \min(|i-j|, N-|i-j|)$. Nous avons proposé dans [P64] un nouvel algorithme plus rapide qui ne sera pas détaillé ici (il est l'objet d'étude poussée dans [Js31] et ce travail est disponible en [HBFR13]). Il consiste à effectuer une recherche en profondeur du graphe où le choix de l'ordre de parcours des sommets est déterminé par l'indice de Jaccard entre un sommet et ses voisins. Il permet par exemple de passer de l'étiquetage aléatoire du graphe en figure 4.19 (a) à l'étiquetage plus régulier de cette même figure en (b)

Exemples. La figure 4.20 propose 5 exemples de graphes de 100 noeuds. Pour chaque exemple, la colonne (a) propose un aperçu du graphe. La colonne (b) affiche les 4 premières composantes de la collection de signaux obtenue après transformation. On peut comparer les caractéristiques de ces signaux avec les graphes dont ils sont issus : la présence de régularité dans le graphe se traduit par des sinusoïdes (graphes 1 et 2) alors que la présence de communautés (graphe 3) donne des signaux avec des paliers correspondant aux communautés du graphe. Il est intéressant de noter que la présence de ces deux propriétés (graphe 4) permet de retrouver des signaux avec des paliers et des signaux sinusoïdaux. Le graphe 5 illustre d'autres types de signaux qu'il est possible d'obtenir et qui présentent également des caractéristiques particulières. Enfin, la transformation sur un graphe aléatoire (graphe 6) montre l'absence de structure particulière sur les signaux si ce n'est celle ajoutée du fait du tri décroissant des composantes par énergie.

Analyse spectrale des signaux. Les signaux obtenus sont étudiés par analyse spectrale sur les composantes de la matrice $\mathbf{X}$. Soit une collection de K signaux indexés par N nœuds,

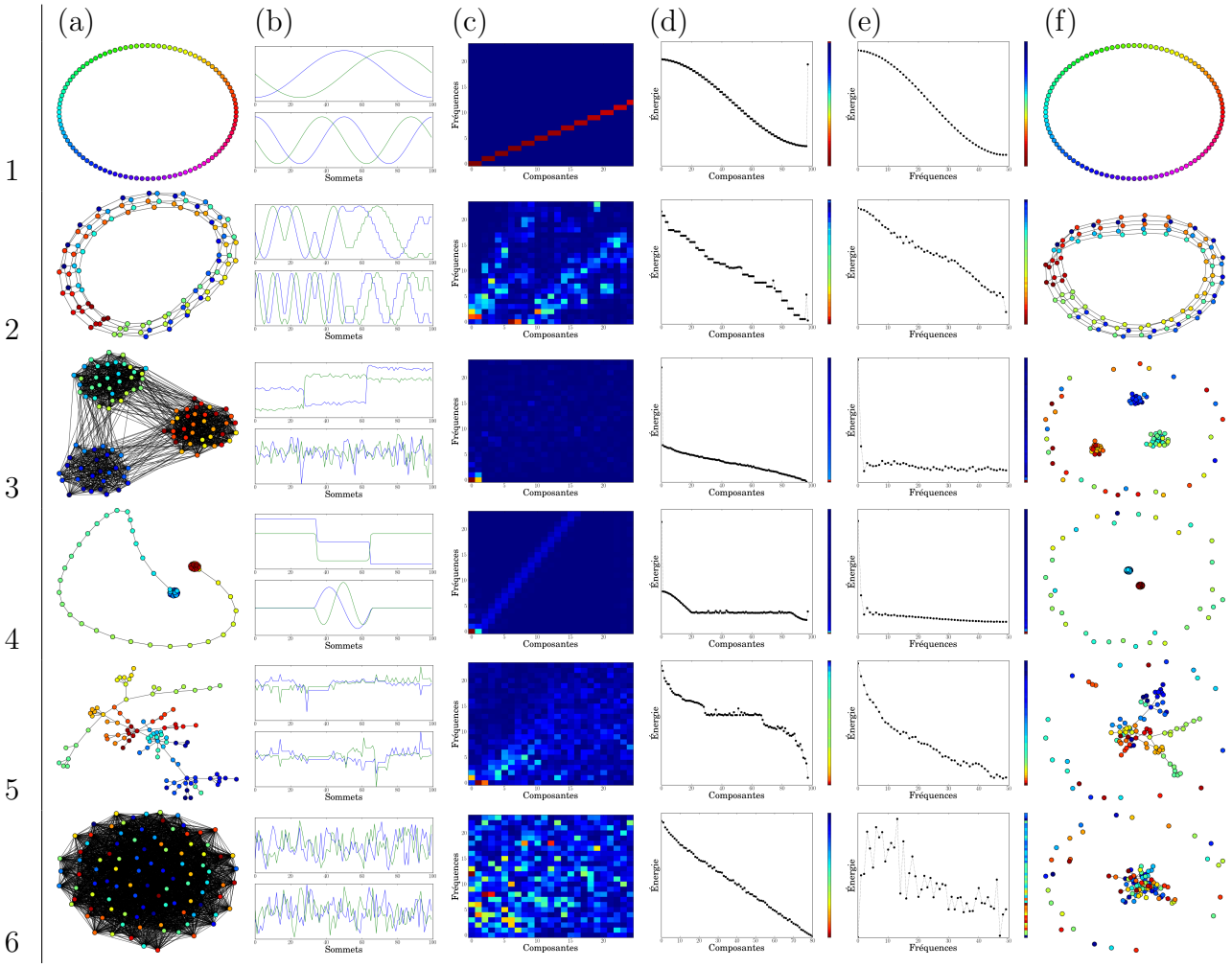


FIGURE 4.20 – Exemples de graphes et signaux associés. Colonnes : (a) Représentation du graphe (la couleur représente le numéro d'étiquette du noeud). (b) Les 4 premières composantes associées au graphe. (c) Énergie en fonction de la composante et de la fréquence pour les 25 premières composantes et 25 premières fréquences en code couleur. (d) Énergie en fonction de la composante moyennée sur les fréquences avec représentation en code couleur. (e) Énergie en fonction de la fréquence moyennée sur les composantes avec représentation en code couleur. (f) Représentation du graphe après reconstruction (r arbitraire). La couleur est codée du bleu foncé (valeur faible) au rouge foncé (valeur élevée). Lignes : 1. Modèle de Watts-Strogatz : 100 nœuds, degré 2 avec probabilité nulle de changer le lien entre 2 nœuds. 2. Grille 2D 20×5 : 100 nœuds. 3. Graphe de 100 nœuds répartis en 3 communautés. 4. Modèle de Barbell : 2 cliques de 35 noeuds reliées par un chemin de 30 nœuds. 5. Modèle de Barabasi-Albert : 100 nœuds. 6. Modèle de Erdös-Rényi : 100 nœuds, $p = 0.6$

le spectre de la composante k est donné par

$$S(k, f) = |\mathbb{F}\mathbf{X}^{(k)}(f)|^2 \tag{4.25}$$

évalué, pour les fréquences positives, sur $\frac{N}{2} + 1$ échantillons, $\mathbb{F}$ étant la transformée de Fourier. Il est alors possible de calculer les caractéristiques suivantes :

— Énergie des composantes : $\forall k \in \{1, \cdots, K\}, E_k = \|\mathbf{X}^{(k)}\|^2 = \sum_{f=1}^{\frac{N}{2}+1} S(k, f)$

— Énergie des fréquences : $\forall f \in \{1, \cdots, \frac{N}{2} + 1\}, \bar{S}(f) = \sum_{k=1}^{K} S(k, f)$

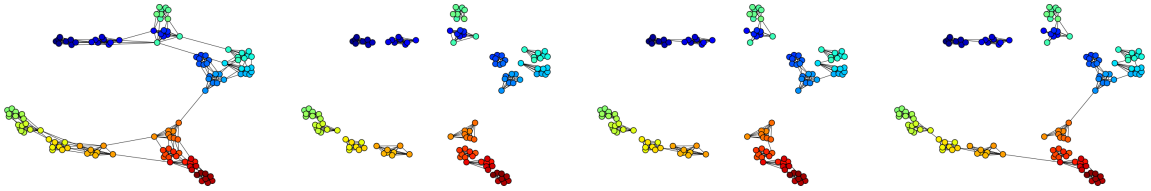


FIGURE 4.21 – Exemple de reconstructions partielles du graphe sur un graphe avec communautés imbriquées à différentes échelles. Pour $r = 0.1$, seuls les nœuds appartenant aux communautés à petite échelle sont sélectionnés. Quand r augmente, des liens entre ces communautés se forment pour mettre en évidence les communautés à plus grande échelle.

Ces caractéristiques peuvent être reliées à des propriétés du graphe comme illustré sur la figure 4.20. Les propriétés des graphes se retrouvent dans l'analyse spectrale : la colonne (c) représente l'énergie de chaque fréquence pour chaque composante, tandis que les colonnes (d) et (e) sont respectivement les marginales des composantes et des fréquences. Ces deux grandeurs permettent de représenter synthétiquement les caractéristiques des signaux et donc les propriétés du graphe.

Reconstruction du graphe à partir de signaux. La reconstruction d'un graphe à partir d'une collection de signaux est triviale lorsque toutes les composantes sont prises en compte : selon le principe du *CMDS*, la matrice de distance $\mathbf{D}$ entre les points est identique à la matrice $\mathbf{\Delta}$ définie sur le graphe. Un simple seuillage suffit alors à retrouver la matrice d'adjacence du graphe et donc le graphe. Si moins de composantes sont retenues, les distances entre points ne sont plus égales à 1 ou à w mais ont une distribution dont la largeur dépend du nombre de composantes retenues. Ces distributions peuvent être utilisées pour sélectionner les liens les plus significatifs : pour une paire de sommets reliés entre eux, plus la distance entre ces sommets dans l'espace euclidien réduit va être faible, plus le lien est considéré comme significatif. Il est alors possible de définir un seuil pour déterminer la significativité d'un lien. Connaissant la matrice d'adjacence du graphe complet, il est possible d'identifier les distributions correspondants aux distances entre les paires de nœuds liées et non-liées, et ainsi de connaître l'intervalle sur lequel les deux distributions se chevauchent. Cet intervalle correspond à une zone où la distance entre deux nœuds n'est pas suffisante pour déterminer leur connectivité avec certitude dans le graphe complet. Le seuil est ainsi choisi de façon à ce que les arêtes sélectionnées soient « fiables » c'est-à-dire ne correspondent pas à des distances appartenant à cet intervalle de chevauchement. On note r la proportion de composantes retenues sur le nombre total de composantes.

Un exemple détaillé de reconstruction d'un graphe à partir de signaux est donné en Figure 4.21 pour un graphe hiérarchique avec communautés : à petite échelle, 16 communautés sont visibles qui fusionnent progressivement pour former 8, puis 4 et enfin 2 grandes communautés lorsque l'échelle augmente. La sélection des liens les plus significatifs permet de mettre en évidence, pour différentes proportions de composantes retenues, ces différentes échelles de communautés. Sur la figure 4.20, la colonne (f) montre plus systématiquement des exemples de reconstruction des graphes pour une valeur r fixée pour chaque graphe de façon à obtenir une version réduite, c'est-à-dire avec moins de liens, du graphe.

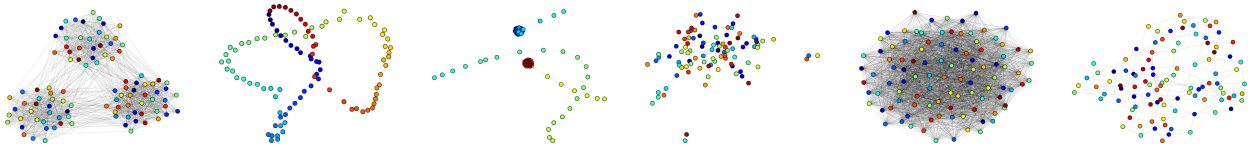


FIGURE 4.22 – Instantanés du graphe dynamique utilisé comme exemple, montrés à différents temps juste avant un changement du graphe de référence.

	$e \in \mathcal{E}_m$	$e \notin \mathcal{E}_m$
$e \in \mathcal{E}^{t-1}$	$1 - 10^{-4}$	0.90
$e \notin \mathcal{E}^{t-1}$	0.1	10^{-4}

TABLE 4.3 – Probabilités fixées qu'une arête e soit dans $\mathcal{E}^t$ ensemble des arêtes du graphe dynamique au temps t en fonction de la présence dans $\mathcal{E}_m$ ensemble des arêtes du graphe modèle et de $\mathcal{E}^{t-1}$ ensemble des arêtes du graphe dynamique au temps $t-1$

Extension de la transformation et analyse « temps-fréquence » des graphes dynamiques. L'analyse fréquentielle des signaux issus d'une transformation d'un graphe permet de mettre en évidence des motifs caractéristiques des propriétés du graphe transformé. Cette approche peut être étendue aux graphes dynamiques : le suivi dans le temps de ces signatures fréquentielles permet de mettre en évidence l'évolution de la structure globale du graphe. On considère le cas où le nombre de sommets dans le graphe est fixé : seuls les liens évoluent au cours du temps. On note $\mathbf{X}_t^{(k)}$ la k-ème série issue d'un graphe au temps t, son spectre, que l'on qualifie (abusivement) de « temps-fréquence » est donné par

$$S(k, f, t) = |\mathbb{F}\mathbf{X}_t^{(k)}(f)|^2. \tag{4.26}$$

À chaque pas de temps, il est ainsi possible de calculer les caractéristiques détaillées dans la section précédente et de suivre au cours du temps l'évolution de la structure du graphe. On note $\mathcal{G}^t = (\mathcal{V}, \mathcal{E}^t)$ un graphe dynamique où $\mathcal{E}^t$ est l'ensemble des arêtes à l'instant t.

Par exemple, un graphe de 100 noeuds est généré à partir d'une série de graphes de référence : à chaque pas de temps, deux nœuds sont liés avec une probabilité dépendant de la présence ou non du lien dans le graphe de référence et dans le graphe dynamique au temps précédent (ce qui fait que le graphe est construit dans le même esprit que le modèle de la section 4.2.2). La table 4.3 donne les probabilités utilisées alors que la figure 4.22 représente des instantanées du graphe à l'instant précédant un changement du graphe de référence tous les 50 pas de temps.

La relation entre l'évolution des énergies par composante et par fréquence au cours du temps présentées en figure 4.23 et les motifs discutés en figure 4.20 permet de mettre en évidence les changements de structure du graphe à chaque instant et son évolution au cours du temps. Les ruptures brutales de structure dues au changement de graphe de référence sont visibles, ainsi que les transitions vers le nouveau graphe de référence.

Développements futurs. La transformation de graphe en signaux permet la réalisation de deux objectifs : elle donne tout d'abord la possibilité de suivre la structure globale du graphe au cours du temps en rattachant des motifs fréquentiels des signaux à des propriétés

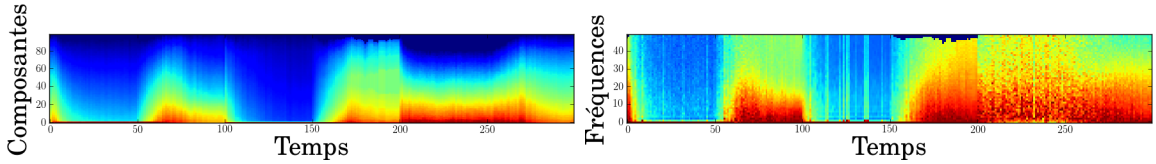


FIGURE 4.23 – Représentations « temps-fréquence » d'un réseau dynamique. À gauche : énergie de chaque composante moyennée en fonction du temps. À droite : énergie de chaque fréquence moyennée sur les composantes en fonction du temps. La couleur est codée du bleu foncé (valeur faible) au rouge foncé (valeur élevée).

du graphe. Chaque changement de la topologie du graphe est repérable et permet d'identifier des changements brusques dans la structure ou au contraire des lentes transitions vers une structure qu'il est possible d'identifier. Un deuxième aspect concerne la réduction du graphe en ne considérant qu'une proportion réduite de composantes afin de ne sélectionner que les liens les plus significatifs dans la structure du graphe.

Il est clair que ce travail n'est qu'un début. Nous avons appliqué la méthode de suivi « temps-fréquence » des graphes dynamiques et des reconstructions partielles pour analyser les réseau des déplacements en Vélo'v en 2011. On trouvera des premiers résultats dans [P62, P63]. Un point clef de l'intérêt de la méthode proposée sera de voir si il est possible de modéliser, ou tout au moins d'extraire, des structures spectrales que l'on peut obtenir dans la représentation de la formule (4.25). Pour l'instant, nous avançons dans cette direction en proposant d'extraire ces composantes dynamiques par factorisation de matrice positive [LS99, FI10] comme dans certains problèmes de séparation de source, avant de les reconstruire individuellement comme présenté ci-dessus [Ps67].

4.5 Développements et perspectives

Le lecteur aura pu discerner au fil des travaux qui sont présentés dans ce chapitre, une évolution de mon parcours, partant de l'analyse de systèmes complexes vus à travers leurs représentations comme des réseaux (contacts entre humains, Vélo'v) pour aller de plus en plus vers des approches méthodologiques remettant des méthodes de traitement statistique du signal dans ces sujets aux côtés des méthodes des sciences physiques et informatiques. Des premiers développements de ces travaux seront de croiser les différentes sections de ce chapitre : trouve-t-on aisément des communautés multi-échelle dans les réseaux de contacts entre humain ? Quid du suivi dynamique des réseaux Vélo'v (un peu abordé dans [P63, P64, Ps67]) sur lequel il reste beaucoup à faire ? Peut-on qualifier les sous-groupes de stations Vélo'v à l'aide des tests par bootstraps contraints ? D'une certaine manière, ce chapitre n'est que le début d'un programme de travail.

Plus généralement, les perspectives sont de contribuer à la fois au champ de l'étude des réseaux complexes et au domaine du traitement du signal sur graphes. En m'intéressant au traitement des signaux sur ou pour des graphes, je mets en avant la nuance entre les deux qui est de pouvoir à la fois analyser des signaux indexés par un graphe (ou dont le support est un graphe) mais aussi d'étudier des graphes qui seraient les données elles-mêmes pour y faire porter des opérations de traitement statistique du signal. Dans les deux cas, un graphe constitué de nœuds et d'arêtes code pour les relations entre les objets d'un réseau.

Je m'intéresse à ces deux situations possibles et aux cas où l'on doit mélanger les deux, dans le cas de signaux déployés sur un réseau qu'on peut dire complexe. Quelques nouvelles pistes sont déjà en cours et j'en livre ici quelques aspects.

Estimation des graphes de relations à partir des signaux. Dans les situations rencontrées ici, le graphe sous-jacent était soit donné par les mesures, soit construit par voisinage entre nœuds. Un enjeu plus général concerne l'estimation du graphe de relation entre des signaux quand il n'est pas donné à l'avance. En collaboration avec O. Michel, S. Achard et F. Châtelain du GIPSA-lab (Grenoble) qui abordaient le même type de question, nous étudions ce problème avec comme motivation l'étude des séries en IRM fonctionnelle. Il s'agit en particulier d'améliorer l'étude des dépendances entre les voxels et des réseaux de connexions fonctionnelles dans le cerveau [RABV13]. Une doctorant, Aude Costard (que je co-encadre à 25%), a ainsi commencé en septembre 2011 et nos premiers résultats portent sur l'estimation parcimonieuse (car on s'attend à ne devoir retenir comme pertinents qu'un nombre faible de liens par rapports à ceux qui sont possible) de la structure de dépendance entre nœuds d'un réseau. Pour cela, une combinaison des méthodes d'estimation bayésienne des supports de modèles graphiques [GG99, Rov02, MB06a] et des approches de *graphical lasso* [MB06b, FHT07] pour l'estimation des graphes de dépendances quand on a peu de points de mesure, a été étudiée. Un premier article de conférence a été publié [P58] mais la méthode était alors limitée à un très petit nombre de nœuds (moins de 10). Le travail en cours permet d'envisager d'aborder des tailles de réseaux plus raisonnables (quelques dizaines de nœuds serait déjà suffisant pour les applications en IRM fonctionnelle) et il devrait même être possible d'aller au-delà par des stratégies multi-résolution s'appuyant sur les ondelettes sur graphes.

Signaux sur graphes. Nous avons vu en 4.1.3 un l'état de l'art du traitement du signal sur graphes et, dans ce domaine très actif, il y a encore beaucoup de travaux à faire. Nous travaillons ainsi à importer les idées des décompositions non stationnaires en modes telles que l'EMD du chapitre 2, à l'étude des signaux sur graphes, en complément des représentations en ondelettes. Un autre sujet qui nous motive est l'agrégation conjointe des graphes et des signaux : les communautés sont une manière d'agréger les graphes, les ondelettes un autre pour agréger les signaux. Il est facile de voir qu'un rapprochement entre les deux opérations est possible à l'aide des outils que nous avons proposé. De manière générale, les travaux de ce chapitre peuvent servir à des études de signaux, y compris multi-variés comme ceux de réseaux de capteurs, pris en des points d'un graphe de terrain.

Travaux liés au chapitre 4

Journaux à comité de lecture

[Js31] R. Hamon, P. Borgnat, P. Flandrin, C. Robardet, "Following the structure of the graph to solve the cyclic bandwidth sum problem", submitted 12/2013.

[Js28] M. Vogel, R. Hamon, L. Merchez, G. Lozenguez, P. Abry, J. Barnier, P. Borgnat, P. Flandrin, I. Mallon, C. Robardet, "From Bikes to Bike Share Users in Velo'v Lyon : Use rates of Lyon's BBS and typology of users", submitted, 07/2013.

[Js27] N. Tremblay, P. Borgnat, "Graph Wavelets and Community Mining", submitted, 07/2013.

[Js25] J.-B. Rouquier, P. Borgnat, « Cartographie des pratiques du Vélo'v : le regard de physiciens et d'informaticiens », accepté à *RSL (Revue Sciences/Lettres)*, ENS éditions, numéro 2, 2014.

[**J24**] N. Tremblay, A. Barrat, C. Forest, M. Nornberg, J.F. Pinton, P. Borgnat, "Bootstrapping under constraint for the assessment of group behavior in human contact networks", *Physical Review E*, vol. 88 :5, p. 052812, Nov. 2013.

[**J19**] P. Borgnat, C. Robardet, J.-B. Rouquier, E. Fleury, P. Abry, P. Flandrin, "Shared Bicycles in a City : A Signal Processing and Data Analysis Perspective", *Advances in Complex Systems*, Vol. 14, No. 3, p. 415-438, 2011.

[**J8**] A. Scherrer, P. Borgnat, E. Fleury, J.-L. Guillaume, C. Robardet, "Description and simulation of mobility networks," *Computer Networks*, vol. 52, Issue 15, pp. 2842-2858, 23 October 2008.

Actes publiés dans des colloques avec actes à comité de lecture

[**Ps67**] R. Hamon, P. Borgnat, P. Flandrin, C. Robardet, "Nonnegative matrix factorization to find features in temporal networks", to be published in ICASSP 2014.

[**P66**] G. Michau, A. Nantes, E. Chung, P. Borgnat, P. Abry, "Retrieving Dynamic Origin-Destination Matrices from Bluetooth Data", *Transportation Research Board, 93rd Annual Meeting*, Washington DC, 12-16 January 2014.

[**P65**] N. Tremblay, P. Borgnat, "Community Mining in Large Networks using Graph Wavelet Transform of Random Vectors", *GlobalSIP 2013 : IEEE Global Conference on Signal and Information Processing (Symposium : Graph Signal Processing)*, Austin (TX, USA), December 3-5, 2013.

[**P64**] R. Hamon, P. Borgnat, P. Flandrin, C. Robardet, "Networks as Signals, with an Application to a Bike Sharing System", *GlobalSIP 2013 : IEEE Global Conference on Signal and Information Processing (Symposium : Information Processing over Networks)*, Austin (TX, USA), December 3-5, 2013.

[**P63**] N. Tremblay, P. Borgnat, "Multiscale Detection of Stable Communities Using Wavelets on Networks", *European Conference of Complex Systems, ECCS 2013*, Barcelona (Spain), September 2013.

[**P62**] 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 of Complex Systems, ECCS 2013*, Barcelona (Spain), September 2013.

[**P60**] N. Tremblay, P. Borgnat, "Multiscale community mining in networks using spectral graph wavelets", *21th European Signal Processing Conf. EUSIPCO-13*, Bucharest (RO), September 2013.

[**P59**] N. Tremblay, P. Borgnat, "Partitionnement multi-échelle d'un graphe en communautés : détection des échelles pertinentes", *24e Colloque sur le Traitement du Signal et des Images. GRETSI-2013*, id. 176, Brest (France), 3-6 septembre 2013.

[**P58**] A. Costard, S. Achard, O. Michel, P. Borgnat, P. Abry, "Encadrement du paramètre de pénalisation dans l'estimation bayésienne asymptotique de la structure d'un graphe initialisée par Graphical lasso", *24e Colloque sur le Traitement du Signal et des Images. GRETSI-2013*, id. 370, Brest (France), 3-6 septembre 2013.

[**P57**] R. Hamon, P. Borgnat, P. Flandrin, C. Robardet, "Transformation de graphes dynamiques en signaux non stationnaires", *24e Colloque sur le Traitement du Signal et des Images. GRETSI-2013*, id. 251, Brest (France), 3-6 septembre 2013.

[**P53**] N. Tremblay, P. Borgnat, J.-F. Pinton, A. Barrat, M. Nornberg, C. Forest, "Constrained graph resampling for group assessment in human social networks", *European Conference of Complex Systems, ECCS 2012*, Bruxelles (Belgique), September 2012.

[**P46**] G. Michau, C. Robardet, L. Merchez, P. Jensen, P. Abry, P. Flandrin, P. Borgnat,, "Peut-on attraper les utilisateurs de Vélo'v au Lasso ?", *23e Colloque sur le Traitement du Signal et des Images. GRETSI-2011*, id. 441, Bordeaux (France), 5-8 septembre 2011.

[**P39**] P. Borgnat, E. Fleury, J.-L. Guillaume, C. Robardet, "Characteristics of the Dynamic of Mobile Networks", *4th International Conference on Bio-Inspired Models of Network, Information, and Computing Systems*, 09/12/2009.

[**P37**] P. Borgnat, E. Fleury, C. Robardet, A. Scherrer, "Spatial analysis of dynamic movements of Vélo'v, Lyon's shared bicycle program", *European Conference on Complex Systems, ECCS'09*, Warwick University (UK), 21-25 September 2009.

[**P36**] P. Borgnat, P. Abry, P. Flandrin, J.-B. Rouquier, "Studying Lyon's Vélo ?V : A Statistical Cyclic Model", *European Conference on Complex Systems, ECCS'09*, Warwick University (UK), 21-25 September 2009.

[**P35**] P. Borgnat, P. Abry, P. Flandrin, "Modélisation statistique cyclique des locations Vélo'v à Lyon", *22e Colloque sur le Traitement du Signal et des Images. GRETSI-2009*, Dijon (France), 8-11 septembre 2009.

[**P26**] A. Scherrer, P. Borgnat, E. Fleury, J.-L. Guillaume, C. Robardet, "A Methodology to Identify Characteristics of the Dynamic of Mobile Networks," *AINTEC 2008 Asian Internet Engineering Conference)*, Bangkok, Thailand, 18-20 November 2008.

[**P25**] P. Borgnat, E. Fleury, J.-L. Guillaume, C. Magnien, C. Robardet, A. Scherrer "Evolving networks," *Selected Proceeding of the Mining Massive Data Sets for Security NATO Workshop 2007*, Gazzada (Italy), IOS Press, 2008.

Chapitre dans des ouvrages collectifs

[C6] P. Borgnat, C. Robardet, P. Abry, P. Flandrin, J.-B. Rouquier, N. Tremblay, "A Dynamical Network View of Lyon's Vélo'v Shared Bicycle System", in *Dynamics On and Of Complex Networks, Volume 2* (A. Mukherjee, M. Choudhury, N. Ganguly, F. Peruani, B. Mitra, eds.), Springer, 2013.

[C4] P. Borgnat, E. Fleury, J.-L. Guillaume, C. Magnien, C. Robardet, A. Scherrer "Evolving networks," in *Proceedings of NATO ASI 'Mining Massive Data Sets for Security'*, Françoise Fogelman-Soulié and Domenico Perrotta and Jakub Piskorski and Ralf Steinberg editors, *NATO Science for Peace and Security Series D : Information and Communication Security*, vol. 19, pp. 198-204, IOS Press, 2008.

Chapitre 5

Conclusion

5.1 Bilan sur les travaux effectués

Le mémoire qui s'achève bientôt a permis de présenter la plupart des travaux que j'ai menés en les structurant autour de trois domaines : le traitement du signal non stationnaire, le télétrafic sur les réseaux d'ordinateurs, l'étude des graphes complexes en lien avec le traitement du signal. Il est vraisemblablement clair maintenant que ces trois domaines ne sont pas séparés et peuvent être vus comme trois facettes d'une même activité. En effet, les études du télétrafic s'appuient sur des outils de traitement du signal pour les séries temporelles ; mes contributions en traitement du signal sont alimentées par les applications aux graphes complexes ou aux réseaux d'ordinateurs tandis que j'ai abordé l'étude des graphes complexes avec des idées venues du « signal » (les approches multi-échelle ou non stationnaires). Les trois domaines présents dans ce mémoire dialoguent donc en fait en permanence.

Des conclusions et perspectives ont été dressées après chacun des trois chapitres présentant ces travaux mais il me semble maintenant intéressant de compléter cela par un regard un plus synthétique qui ne cherche pas à séparer les contributions faites en trois domaines.

Une première conclusion à nos travaux est que le paradigme des méthodes « non stationnaires » a encore une grande efficacité, tant pour étudier des signaux que des réseaux et que de nombreux travaux restent à venir. Sur la question des tests de stationnarité, je pense que nous avons amené un cadre pragmatique original, assez large et surtout adaptable car il incorpore les notions de temps caractéristique d'observation et de représentation et fusionne dans un même cadre les signaux aléatoires et déterministes. Ce travail complète les travaux qui proposaient des tests paramétriques ou dans des situations spécifiques et les développements futurs attendus sur ce sujet concernent surtout des ajustements ou des améliorations algorithmiques du test dans des situations appliquées. C'est d'ailleurs dans les applications que l'on trouve sans cesse de nouveaux enjeux, pour cette question des tests de stationnarité, mais aussi pour d'autres questions sur le non stationnaire. Concernant les tests de stationnarité, ils retrouvent les résultats des tests paramétriques dans les situations où ces derniers sont valides, quoiqu'avec un peu moins de finesse – c'est le défaut général du non paramétrique. Face à des mesures expérimentales, on aimerait cependant guider l'expérimentateur vers la bonne approche, le bon test.

La même question se pose pour les diverses méthodes de représentations ou décompositions non stationnaires décrites dans ce mémoire (temps-fréquence, EMD, extraction de ten-

dances) auxquelles il faut ajouter celles de la littérature abondante sur ce sujet, comme par exemple les ondelettes (ou temps-échelle), la réallocation, le synchrosqueezing,.. Cependant, il reste cette fois beaucoup à faire aussi sur les aspects méthodologiques : les classes de solutions dans le domaine non stationnaire sont à comparer ; le passage à de nouvelles modalités de données comme des données multivariées, des images ou, plus compliqué encore, des données sur graphe, n'a pas toujours été fait. La combinaison entre dynamique et multi-échelle est aussi un point intéressant qui entre dans nos perspectives pour étudier les signaux sur réseaux complexes.

Concernant les applications, mon impression est que les contributions du traitement statistique du signal à l'étude du télétrafic informatique a eu ses heures de gloire et que l'on est actuellement dans une phase de consolidation des connaissances. Il reste des études précises à mener – j'ai mentionné celle sur la dynamique et les statistiques attendues à temps court et cette question de multifractalité – mais il faut bien reconnaître que les enjeux les plus visibles en réseau se trouvent maintenant ailleurs. La classification des paquets et des flots reste un sujet qui profite toujours des développements réalisés dans les méthodes d'apprentissage ou de classification, mais c'est à nouveau une même question qui est reprise avec de nouveaux outils, pas un nouvel enjeu.

Inversement, aborder des thématiques de transport (telles que les études sur Vélo'v) ou d'études de données liées aux comportements sociaux et humains, sont des sujets qui ne cessent de poser de nouvelles questions. Parfois, ce sont des questions anciennes de ces domaines qui peuvent être revisitées grâce aux nouvelles formes de données, et à leur abondance, que les développements techniques et informatiques amènent. Par exemple, pour le système Vélo'v, les études discutées ici ne sont que des étapes préliminaires pour parvenir à un modèle et une simulation des déplacements en Vélo'v qui devra se positionner par rapport au très classique modèle à quatre étapes en transport ; avoir à disposition des données exhaustives, des outils statistiques ou de fouille de données, des outils d'analyse temporelle ne dispense pas de s'appuyer sur la démarche usuelle en transport pour aller vers des modèles pertinents. Parfois, les questions peuvent être nouvelles. Le travail est cependant à faire pour voir comment les techniques élaborées d'analyse de données (par exemple, dans ce mémoire : bootstrap de graphes, classifications, représentations multi-échelles ou dynamiques) pourront contribuer à des études en sciences sociales que ce soit pour tracer le portrait d'utilisateurs de Vélo'v ou pour comprendre la dynamique des interactions et contacts entre humains.

Je reste enfin fasciné par les nombreuses questions que posent les sciences physiques aux méthodes de traitement du signal, en les poussant souvent dans leurs retranchements. Mon activité publiée sur des questions venues de la physique n'est finalement pas si grande que cela ces dernières années mais plusieurs projets existent et, pour ces études-là, il faut y passer autant de temps que pour les travaux multi-disciplinaires en transport ou en réseau. Le travail sur les simulations en mécanique des fluides a par exemple demandé un réel investissement pour finir par proposer une méthode plutôt simple techniquement, mais bien adaptée aux enjeux et aux contraintes. Je ne doute pas que les travaux d'analyse de signaux d'expériences de biophysique évoqués en fin de chapitre 2 conduiront à terme à des résultats intéressants.

Finalement, le travail pour lequel cette conclusion ne sera qu'un bilan préliminaire concerne les développements proposés en traitement du signal sur et pour les graphes. Ces travaux sont plus récents que les autres et doivent se confronter au corpus assez vaste de

travaux qui s'intéressent à l'étude de réseaux complexes. Comme il est très justement dit dans [SNF+13], "the bulk of the research prior to the past decade focused on analyzing the underlying graphs, as opposed to signals on graph." En quelque sorte, l'étude des réseaux complexes est déjà bien développée et se déploie maintenant vers leur dynamique, vers des réseaux multiplexés, vers des applications données. En revanche, il reste à faire mieux émerger une théorie du signal qui tiendrait pour les signaux sur graphes et se marierait sans heurt avec les analyses des réseaux, surtout quand ils se révèlent complexes.

Les travaux dont nous avons discuté dans le dernier chapitre sont des premiers pas qui veulent aller dans ces directions : comment répondre à des questions posées dans l'analyse des réseaux (détecter des communautés, étudier des réseaux dynamiques, tester des propriétés de sous-groupes de réseau) en s'appuyant sur une théorie du signal adaptée aux graphes de manière à pouvoir étudier de la même manière les graphes avec des signaux dessus. Notre relecture multi-échelle des communautés à l'aide des ondelettes est particulièrement intéressante pour cet objectif. Plus généralement, un graal est de savoir décrire conjointement graphes et signaux sur graphes à la fois en temps (avec leurs aspects dynamiques) et en espace (dans le domaine des nœuds) et il fallait bien commencer par des situations où certains aspects sont fixés ou simplifiés.

Positionnement scientifique. J'affiche bien volontiers deux points spécifiques au travail de recherche que j'ai mené et qui m'a permis l'écriture de ce mémoire.

Le premier est la démarche de s'intéresser à des domaines et des méthodes venus de disciplines différentes et de ne pas me poser en spécialiste d'une unique question, d'un seul outil. Bien qu'étant principalement en traitement du signal, les travaux vont selon le moment plutôt vers l'informatique (en particulier des réseaux d'ordinateurs) ou les sciences physiques, ou un mixte de tout cela quand il s'agit d'étudier des graphes et réseaux complexes. À ce propos, mon positionnement au sein d'un laboratoire de physique est pertinent pour ce thème car les sciences physiques sont, avec l'informatique, les disciplines qui ont le plus influé sur les méthodes d'étude des systèmes et réseaux complexes. Grâce à cette appartenance partielle aux sciences physiques, nous avons pu nous rapprocher du projet Sociopatterns et conduire des expériences de mesures des contacts face-à-face entre personne à l'aide de capteurs RFID. Les liens tissés dans le domaine de l'informatique des réseaux ont eux aussi conduit à des études intéressantes sur les réseaux de contacts entre humains ou sur les signaux non stationnaires de capteurs environnementaux ou de consommation d'énergie dans les bâtiments, et pas seulement à des études en télétrafic informatique. Plus avant, le travail que je fais trouve son prolongement dans l'IXXI, Institut Rhône-Alpin des Systèmes Complexes, qui regroupe des initiatives en étude des systèmes et réseaux complexes.

Le deuxième aspect spécifique à ma démarche est le choix d'apporter des approches pragmatiques pour étudier des problèmes et des situations données même quand une démarche fondée sur des considérations théoriques nous échappe un peu. On pourrait en effet souvent reprocher aux travaux de ne pas toujours avoir une base théorique complètement assise : nous n'avons pas de théorèmes qui prouvent la justesse des tests de stationnarité, la méthode EMD échappe encore à une analyse théorique complète, la méthode des médianes de sketch pour le trafic est elle aussi heuristique comme le sont en réalité les méthodes de bootstrap, de détection de communautés par ondelettes et de suivi dynamique de réseaux discutées dans le dernier chapitre. Tout ceci, malgré un goût personnel allant facilement vers les mathématiques... ! Dans tous ces cas où les théories nous font défaut, c'est par une

démarche expérimentale, appuyée sur des simulations ou des mesures réelles, que nous pouvons valider ces méthodes, les comparer à d'autres, jusqu'à acquérir la certitude qu'elles sont justes. Pragmatiquement, cette démarche nous permet d'avancer des méthodes bien fondées dans des cas où l'analyse théorique ne permet pas toujours d'en proposer.

Un nouveau regard sur les applications. Répondre aux enjeux des applications étudiées (comme le trafic Internet, Vélo'v, les LES,...) nécessite non seulement de faire preuve d'imagination et de méthode, de s'approprier des démarches venues de domaines voisins, mais surtout de travailler au plus près des chercheurs qui connaissent ces applications, ces domaines, leurs enjeux, les données et ce qui a déjà été fait. C'est ce qui a été fait en métrologie des réseaux ou en analyse des réseaux de contacts entre humains et c'est en cours pour les études en sociologie et en transport des déplacements en Vélo'v. Bien qu'il faille comprendre et s'approprier ces questions, il n'est pas possible de devenir pleinement chercheur dans tous ces domaines. Une conclusion des travaux, appuyée sur l'expérience de ceux effectués sur le télétrafic informatique et réalisés avec des chercheurs en pointe sur ces sujets, rejoint en fait les perspectives pour les travaux à réaliser sur les nouvelles applications en cours, en particulier en transport ou en étude de données liées à l'humain et au social : c'est en continuant à se rapprocher des chercheurs de ces domaines que l'on pourra contribuer de manière pertinente avec nos approches à ces questions.

5.2 Programme de travail futur

Quelques perspectives qui découlent directement des travaux présentés dans le mémoire ont été dressées en conclusion des différents chapitres. En adoptant une vision à plus long terme, je les complète ici par quelques éléments de programme de travail que je compte développer dans le futur.

Traitement du signal sur et pour des graphes. Le projet est de participer au développement de ce domaine, en particulier en se penchant sur les représentations multi-échelles dès lors que l'on veut qu'elles portent conjointement sur les graphes (en respectant leurs propriétés de réseaux complexes) et sur les signaux qu'ils portent. Les représentations adaptatives, éventuellement pilotées par les données telles que l'EMD, sont aussi un sujet d'étude que je compte regarder, de même que les approches décentralisées et distribuées qui, dans le contexte d'abondance de données qui est la donne actuelle, présentent bien des avantages pour aller vers des solutions effectives et efficaces en traitement du signal sur graphes. Bien entendu, un enjeu important reste la dynamique des réseaux et de leurs signaux ; nous avons déjà vu une proposition originale à ce sujet dans le mémoire et nous comptons continuer en ce sens, par exemple en combinant les approches classifications adaptées à des structures dynamiques à celles adaptées à une lecture multi-résolution.

Amener nos outils vers d'autres problématiques. Il a été suggéré en fin du chapitre 2 que les outils non stationnaires peuvent servir à de nouvelles applications mais c'est le cas d'autres méthodes discutées dans le mémoire. Par exemple, le domaine des transports est riche en questions pour le traitement du signal sur réseaux. Pour une thèse en cotutelle avec le QUT de Brisbane, nous commençons à étudier les méthodes d'estimation

sur graphe des flux origine-destination en partant de trajectoires de voitures mesurées par ces capteurs Bluetooth (voir une première communication [P66] présentée à la conférence *Transportation Research Board* en janvier 2014). Un autre domaine complètement différent qui sera peut-être l'objet d'application de l'approche multi-échelles de communautés dans les graphes est celui de l'étude du programme de réplication du génome en collaboration au sein du laboratoire avec B. Audit et sa doctorante R. Boulos, qui cherchent à regrouper des fragments du génome humain via les données de conformation et d'interaction entre eux. Pour cela, nous travaillons à adapter nos outils à cette nouvelle situation.

Analyse de données liées à l'humain et au social. Finalement, l'analyse de données des sciences humaines et sociales est un champ d'exploration encore plus difficile que ce que j'ai abordé jusqu'ici, à la fois par la nécessité de trouver un vocabulaire et des méthodes communes avec les chercheurs de ces domaines, et par le flou parfois obligatoire dans le traitement des données. J'en veux comme exemple le travail en cours [Js28] pour exhiber des classes d'usagers de Vélo'v qui ne sont pas forcément bien séparées au sens statistique de la classification, dans l'espace choisi de description des usagers. Privilégier l'idée que l'on cherche des classes qui font sens est finalement un peu différent de l'approche traditionnelle en classification de données qui, en général, ne sépare des groupes que si ils sont bien séparables. Un deuxième champ d'application, pour revenir aussi aux aspects dynamiques des réseaux complexes, pourrait être de porter une lecture plus historique sur des réseaux sociaux, ou des réseaux scientométriques que l'on sait maintenant extraire et caractériser à partir des bases de données de publications, afin de comprendre ce que nous apprennent les méthodes dynamiques d'analyse des réseaux. Est ainsi étudié par un groupe de travail au laboratoire, l'évolution du domaine scientifique des ondelettes qui a la chance d'avoir émergé il n'y a pas si longtemps et l'on peut porter dessus des regards croisés venus de la connaissance du domaine, des analyses automatisées par les méthodes pour les réseaux complexes et d'une démarche historique. Ce type de sujet ouvre de nouveaux horizons et présage que le point de vue et les méthodes proposés dans ce mémoire ne sont pas arrivés au bout du chemin.

Travaux antérieurs

Journaux à comité de lecture

[**J5**] P. Borgnat, P.O. Amblard, P. Flandrin, "Stochastic invariances and Lamperti transformations for stochastic processes," *J. of Physics A : Mathematical and General*, vol. 38(10), p. 2081-2101, février 2005.

[**J4**] P. Borgnat, P. Flandrin, "On the Chirp Decomposition of Weierstrass-Mandelbrot Functions, and their Time-Frequency Interpretation," *Applied and Computational Harmonic Analysis*, vol. **15**, p. 134-146, septembre 2003.

[**J3**] P. Borgnat, P. Flandrin et P.O. Amblard, "Stochastic Discrete Scale Invariance," *IEEE Signal Processing Letters*, **9**, n. 6, p. 181-184, juin 2002.

[**J2**] P. Borgnat, A. Lesage, S. Caldarelli et L. Emsley, "Narrowband Linear Selective Pulses for NMR," *J. of Nuclear Magnetic Resonance series A*, **119**, p. 289-294, 1996.

[**J1**] P. Borgnat, A. Lesage, S. Caldarelli et L. Emsley, "Improved Sensitivity in Selective NMR Correlation Spectroscopy and Applications to the Determination of Scalar Couplings in Peptides and Proteins," *J. of Am. Chemical Soc.*, vol. 118, **39**, p. 9320-9325, 1996.

Actes publiés dans des colloques avec actes à comité de lecture

[**P12**] P. Borgnat, « On Sampling Methods for Linear Scale-Invariant Systems » *IEEE Int. Conf. on Acoust., Speech and Signal Proc. ICASSP-06*, Toulouse (France), 14-19 mai 2006.

[**P9**] P.O. Amblard et P. Borgnat, "Stochastic discrete scale invariance : Renormalization group operators and iterated function systems", *IMA Conference on Mathematics and Signal Processing V*, p. 19-22, Cirencester (UK), 14-16 décembre 2004.

[**P8**] P. Borgnat, « Symétrie des champs bidimensionnels et générateurs stationnaires », Journées d'étude sur les *Méthodes pour les signaux complexes en traitement d'image* de l'ARC Thalweg, INRIA Rocquencourt, 9-10 décembre 2003.

[**P7**] P. Borgnat, P.O. Amblard et P. Flandrin, « Invariances d'échelles brisées et accroissements stationnaires », *Colloque GRETSI-03*, Paris (France), 8-11 septembre 2003.

[**P6**] P.O. Amblard, P. Borgnat et P. Flandrin, "Stochastic processes with finite size scale invariance", *SPIE Symposium on Fluctuations and Noise*, Sante Fe (USA), 1-4 juin 2003.

[**P5**] P. Borgnat, P. Flandrin et P.O. Amblard, "Lamperti transformation for finite size scale invariance", *Colloque PSIP 2003 : Physics in Signal and Image Processing*, p. 177-180, Grenoble (France), 29-31 janvier 2003.

[**P4**] P. Borgnat, P. Flandrin et P.O. Amblard, "Generalized Lamperti Transformation for Broken Scale Invariance", *36th Asilomar Conference on Signals, Systems, and Computers*, Pacific Groves (CA), 3-6 novembre 2002.

[**P3**] P. Borgnat, P. Flandrin et P.O. Amblard, « Une approche stochastique de l'invariance d'échelle discrète », *Colloque GRETSI-01*, Toulouse (France), septembre 2001.

[**P2**] P. Borgnat, P. Flandrin et P.O. Amblard, "Stochastic Discrete Scale Invariance and Lamperti Transformation", *IEEE Workshop on Statistical Signal Processing SSP-01*, p. 66-

69, Singapour, août 2001.

[P1] P. Borgnat, O. Michel, C. Baudet et P. Flandrin "From a Vortex Gas Model of Turbulence to Mellin Functions", *Advances in Turbulence VIII, proceedings of ETC-VIII*, p. 988, Barcelone (Espagne), juin 2000.

Chapitre dans des ouvrages collectifs

[C3] O. Teytaud, C. Antonini, P. Borgnat, A. Chateau, E. Lebeau, *Les maths pour l'agreg. Cours complet et synthétique.* Dunod, septembre 2007.

[C2] P. Borgnat, "Signal Processing Methods related to Models of Turbulence", dans *Harmonic Analysis and Rational Approximation : Their rôles in signals, control, and dynamical systems theory*, éd. J.D. Fournier, J. Grimm, J. Leblond et J.R. Partington, p. 277-301, Springer-Verlag, avril 2006.

[C1] P. Flandrin, P. Borgnat et P.O. Amblard "From stationarity to self-similarity, and back : Variations on the Lamperti transformation", dans *Processes with Long-Range Correlations : Theory and applications*, editeurs G. Rangarajan et M. Ding, vol. 26 des *Lecture Notes in Physics*, Springer-Verlag, juin 2003.

Bibliographie

[ABA13] Angeletti (F.), Bertin (E.) et Abry (P.), « Random vector and time series definition and synthesis from matrix product representations : From statistical physics to hidden markov models », *IEEE Transactions on Signal Processing*, vol. 61, 2013, p. 5389 – 5400.

[ABD+08] Alexandrov (T.), Bianconcini (S.), Dagum (E. B.), Maass (P.) et McElroy (T.), « A review of some modern approaches to the problem of trend extraction ». Rapport technique n° RRS2008/03, U.S. Census Bureau, Washington, DC, 2008.

[ABF+02] Abry (P.), Barananiuk, Flandrin (P.), Riedi (R.) et Veitch (D.), « Multiscale network traffic analysis, modeling, and inference using wavelets, multifractals, and cascades », *IEEE Signal Processing Magazine*, vol. 3, n° 19, May 2002, p. 28–46.

[AC07] Acosta (W.) et Chandra (S.), « Trace driven analysis of the long-term evolution of gnutella peer-to-peer traffic », dans *PAM'07*, p. 42–51, 2007.

[Ach68] Achenbach (E.), « Distribution of local pressure and skin friction around a circular cylinder in cross-flow up to R$e = 5 \times 10^6$ », *Journal of Fluid Mechanics*, vol. 34, 1968, p. 625–639.

[ACMF12] Auger (F.), Chassande-Mottin (E.) et Flandrin (P.), « Making reassignment adjustable : The Levenberg-Marquardt approach », dans *IEEE Proc. ICASSP 2012*, Kyoto, Japan, mars 2012.

[AF95] Auger (F.) et Flandrin (P.), « Improving the readability of time-frequency and time-scale representations by the reassignment method », *IEEE Trans. Signal Process.*, vol. 43, n° 5, 1995, p. 1068–1089.

[AFG08] Arenas (A.), Fernandez (A.) et Gomez (S.), « Analysis of the structure of complex networks at different resolution levels », *New Journal of Physics*, vol. 10, n° 5, 2008, p. 053039.

[AFT98] Adler (R. J.), Feldman (R. E.) et Taqqu (M. S.), *A Practical Guide To Heavy Tails*. Chapman and Hall, New York, 1998.

[AFTV00] Abry (P.), Flandrin (P.), Taqqu (M.) et Veitch (D.), « Wavelets for the analysis, estimation and synthesis of scaling data », dans Park (K.) et Willinger (W.), éditeurs, *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons, Inc., 2000.

[AGCO06] Aujol (J.-F.), Gilboa (G.), Chan (T.) et Osher (S.), « Structure-texture image decomposition - modeling, algorithms, and parameter selection », *Int. J. Comp. Vis.*, vol. 67, n° 1, Apr. 2006, p. 111–136.

[AGF95] Abry (P.), Gonçalves (P.) et Flandrin (P.), « Wavelets, spectrum analysis and $1/f$ processes », dans *Wavelets and Statistics, Lecture Notes in Statistics*, vol. 103, p. 15–29, 1995.

[AN98] Andersen (A.) et Nielsen (B.), « A Markovian approach for modelling packet traffic with long range dependence », *IEEE journal on Selected Areas in Communications*, vol. 5, n° 16, 1998, p. 719–732.

[APT07] Allman (M.), Paxson (V.) et Terrell (J.), « A brief history of scanning », dans *IMC'07*, p. 77–82, 2007.

[AV98] Abry (P.) et Veitch (D.), « Wavelet analysis of long-range dependent traffic », *IEEE Trans. on Info. Theory*, vol. 44, n° 1, janvier 1998, p. 2–15.

[BA99] Barabási (A.-L.) et Albert (R.), « Emergence of scaling in random networks », *Science*, vol. 286, 1999, p. 509–512.

[BACPP11] Bricenõ-Arias (L. M.), Combettes (P. L.), Pesquet (J.-C.) et Pustelnik (N.), « Proximal algorithms for multicomponent image processing », *J. Math. Imag. Vis.*, vol. 41, n° 1, Sep. 2011, p. 3–22.

[Bas89] Basseville (M.), « Distances measures for signal processing and pattern recognition », *Signal Processing*, vol. 18, n° 4, 1989, p. 349–369.

[BB00] Bayram (M.) et Baraniuk (R.), « Multiple window time-varying spectrum estimation », dans et al. (W. F.), éditeur, *Nonlinear and Nonstationary Signal Processing*. Cambridge Univ. Press, 2000.

[BBN+11] Bajardi (P.), Barrat (A.), Natala (F.), Savini (L.) et Colizza (V.), « Dynamical patterns of cattle trade movements », *PLoS ONE*, vol. 6, n° 5, 2011, p. e19869.

[BBV08] Barrat (A.), Barthélemy (M.) et Vespignani (A.), *Dynamical processes on complex networks*. Cambridge University Press, Cambridge, 2008.

[BCNV08] Baraniuk (R.), Candès (E.), Nowak (R.) et Vetterli (M.), « Special section on compressive sampling », *IEEE Sig. Proc. Mag.*, vol. 25, n° 2, 2008.

[BCSL07] Boudet (J.), Caro (J.), Shao (L.) et Lévêque (E.), « Numerical studies towards practical large-eddy simulation », *Journal of Thermal Science*, vol. 16, n° 4, 2007, p. 328–336.

[BD08] Blumensath (T.) et Davies (M. E.), « Iterative thresholding for sparse approximations », *J. Fourier Anal. Appl.*, vol. 14, n° 5, 2008.

[Ber94] Beran (J.), *Statistics for Long-memory processes*. Chapman & Hall, New York, 1994.

[BG05] Borg (I.) et Groenen (P.), *Modern multidimensional scaling : Theory and applications*. Springer, 2005.

[BGLL08] Blondel (V. D.), Guillaume (J.-L.), Lambiotte (R.) et Lefebvre (E.), « Fast unfolding of communities in large networks », *J.STAT.MECH.*, 2008.

[BGV92] Boser (B.), Guyon (I.) et Vapnik (V.), « A training algorithm for optimal margin classifiers », dans *Proc. Fifth Annual Workshop on Computational Learning Theory*, p. 144–152, 1992.

[BI06] Brcich (R.) et Iskander (D.), « Testing for stationarity in the frequency domain using a sphericity statistic », dans *Proceedings of IEEE ICASSP-06*, vol. III, p. 464–467, Toulouse (France), 2006.

[BKPR02] Barford (P.), Kline (J.), Plonka (D.) et Ron (A.), « A signal analysis of network traffic anomalies », dans *ACM/SIGCOMM Internet Measurement Workshop*, Marseille, France, novembre 2002.

[BM95] Brooks (S. P.) et Morgan (B. J. T.), « Optimization using simulated annealing », *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 44, n° 2, 1995, p. pp. 241–257.

[BMA+11] Bellala (G.), Marwah (M.), Arlitt (M.), Lyon (G.) et Bash (C. E.), « Towards an understanding of campus-scale power consumption », dans *Buildsys'11*, p. 6, Seattle, WA, Nov. 1, 2011.

[Boa02] Boashash (B.), éditeur, *Time-Frequency Signal Analysis and Processing*. Prentice Hall, Englewood Cliffs (NJ), 2002.

[Bol98] Bollobas (B.), *Modern Graph Theory*. Springer Verlag, New York, USA, 1998.

[Bru00] Brutlag (J.), « Aberrant behavior detection in time series for network monitoring », dans *USENIX System Administration Conference*, New Orleans, decembre 2000.

[BTI+02] Barakat (C.), Thiran (P.), Iannaccone (G.), Diot (C.) et Owezarski (P.), « A flow-based model for internet backbone traffic », dans *ACM/SIGCOMM Internet Measurement Workshop*, p. 35–47, New York, NY, USA, 2002. ACM Press.

[BTS06] Bernaille (L.), Teixeira (R.) et Salamatian (K.), « Early Application Identification », *ACM CoNEXT 2006*, 2006, p. 12.

[CAI] CAIDA. http ://www.caida.org/.

[CB96] Crovella (M. E.) et Bestavros (A.), « Self-similarity in World Wide Web traffic : Evidence and possible causes », dans *Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, p. 160–169, mai 1996.

[CB97] Crovella (M. E.) et Bestavros (A.), « Self-similarity in world wide web traffic : Evidence and possible causes », *IEEE/ACM Trans. Network.*, vol. 5, n° 6, 1997, p. 835–846.

[CBB+10] Cattuto (C.), Van den Broeck (W.), Barrat (A.), Colizza (V.), Pinton (J.) et Vespignani (A.), « Dynamics of person-to-person interactions from distributed rfid sensor networks », *PloS one*, vol. 5, n° 7, 2010, p. e11596.

[CBBV06] Colizza (V.), Barrat (A.), Barthélemy (M.) et Vespignani (A.), « The role of the airline transportation network in the prediction and predictability of global epidemics », *Proc. Natl. Acad. Sci. USA*, vol. 103, 2006, p. 2015.

[CBBV07] Colizza (V.), Barrat (A.), Barthélemy (M.) et Vespignani (A.), « Modeling the worldwide spread of pandemic influenza : Baseline case and containment interventions », *PLoS Med.*, vol. 4(1), 2007, p. e13.

[CC11] Chen (C.) et Cook (D. J.), « Energy outlier detection in smart environments. », dans *Artificial Intelligence and Smarter Living*, vol. WS-11-07 (coll. *AAAI Workshops*). AAAI, 2011.

[CCLS02] Cao (J.), Cleveland (W. S.), Lin (D.) et Sun (D. X.), « Internet traffic tends toward poisson independent as the load increases », dans Holmes (C.) et et al., éditeurs, *Nonlinear estimation and classification*, 2002.

[CDG+00] Cleary (J.), Donnelly (S.), Graham (I.), McGregor (A.) et Pearson (M.), « Design principles for accurate passive measurement », dans *Passive and Active Measurements*, Hamilton, New Zealand, avril 2000.

[CFQS12] Casteigts (A.), Flocchini (P.), Quattrociocchi (W.) et Santoro (N.), « Time-varying graphs and dynamic networks », *International Journal of Parallel, Emergent and Distributed Systems*, vol. 27, n° 5, octobre 2012, p. 387–408.

[CHT98] Carmona (R.), Hwang (W.) et TorrÃ©sani (B.), *Practical Time-Frequency analysis*. Academic Press, 1998.

[Chu88] Chung (F.), « Labelings of graphs », *Selected topics in graph theory*, vol. 3, 1988, p. 151–168.

[Chu97] Chung (F.), *Spectral graph theory*, n° 92. Amer Mathematical Society, 1997.

[CK03] Crovella (M.) et Kolaczyk (E.), « Graph wavelets for spatial traffic analysis », dans *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 3, p. 1848–1857, 2003.

[CKR+07] Cha (M.), Kwak (H.), Rodriguez (P.), Ahn (Y.-Y.) et Moon (S.), « I tube, you tube, everybody tubes : Analyzing the world's largest user generated video system », dans *IMC'07*, p. 1–14, 2007.

[CL02] Chung (F.) et Lu (L.), « The average distances in random graphs with given expected degrees », *Proceedings of the National Academy of Sciences*, vol. 99, n° 25, 2002, p. 15879.

[CL06] Chung (F.) et Lu (L.), *Complex graphs and networks*, n° 107. Amer Mathematical Society, 2006.

[CM05] Cormode (G.) et Muthukrishnan (S.), « What's hot and what's not : Tracking most frequent items dynamically », *ACM Transaction on Database Systems*, vol. 30, n° 1, March 2005, p. 249–278.

[CM06] Coifman (R.) et Maggioni (M.), « Diffusion wavelets », *Applied and Computational Harmonic Analysis*, vol. 21, n° 1, 2006, p. 53–94.

[CMDAF97] Chassande-Mottin (E.), Daubechies (I.), Auger (F.) et Flandrin (P.), « Differential reassignment », *IEEE Signal Process. Lett.*, vol. 4, n° 10, 1997, p. 293–294.

[CMK00] Cho (K.), Mitsuya (K.) et Kato (A.), « Traffic data repository at the WIDE project », dans *USENIX 2000 Annual Technical Conference : FREENIX Track*, p. 263–270, 2000.

[CMM+08] Clementi (A.), Macci (C.), Monti (A.), Pasquale (F.) et Silvestri (R.), « Flooding time in edge-markovian dynamic graphs », dans *27th ACM SIGACT-SIGOPS Symp. on Principles of Distr. Compt. (PODC'08)*, p. 213–222. ACM Press, 2008.

[CMN08] Clauset (A.), Moore (C.) et Newman (M.), « Hierarchical structure and the prediction of missing links in networks », *Nature*, vol. 453, n° 7191, 2008.

[CMP+02] Cappé (O.), Moulines (E.), Pesquet (J.-C.), Petropulu (A.) et Yang (X.), « Long-range dependence and heavy-tail modeling for teletraffic data », *IEEE Signal Processing Magazine*, vol. 5, n° 1, May 2002, p. 14–27.

[CMPS09] Clementi (A.), Monti (A.), Pasquale (F.) et Silvestri (R.), « Information spreading in stationary markovian evolving graphs », dans *23rd IEEE Int. Parallel and Ditr. Proc. Symp. (IPDPS'09)*, 2009.

[CO13] Côme (E.) et Oukhellou (L.). « Model-based count series clustering for bike-sharing system usage mining, a case study with the vélib' system of paris ». Submited to ACM TIST, 2013.

[Coh95] Cohen (L.), *Time-Frequency analysis*. Prentice Hall, New Jersey, 1995.

[Cor] « CoralReef », *http ://www.caida.org/tools/measurement/coralreef/*.

[CP10] Combettes (P. L.) et Pesquet (J.-C.), « Proximal splitting methods in signal processing », dans Bauschke (H. H.), Burachik (R.), Combettes (P. L.), Elser (V.), Luke (D. R.) et Wolkowicz (H.), éditeurs, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, p. 185–212. Springer-Verlag, New York, 2010.

[CP11] Combettes (P. L.) et Pesquet (J.-C.), « Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators », *Set-Valued Var. Anal.*, 2011.

[cPB94] kc claffy, Polyzos (G. C.) et Braun (H.-W.), « Tracking long-term growth of the nsfnet », *Communications of the ACM*, vol. 37, n° 8, 1994, p. 34–45.

[CR05] Candès (E.) et Romberg (J.). « ℓ_1-magic : Recovery of sparse signals via convex programming, user's guide of the ℓ_1-magic toolbox for matlab ». `http://www.acm.caltech.edu/l1magic/`, 2005.

[CRT06] Candès (E.), Romberg (J.) et Tao (T.), « Robust uncertainty principles : Exact signal reconstruction from highly incomplete frequency information », *IEEE Trans. on Info. Theory*, vol. 52, n° 2, 2006.

[CSM+11] Campanharo (A.), Sirer (M.), Malmgren (R.), Ramos (F.) et Amaral (L.), « Duality between time series and networks », *PloS one*, vol. 6, n° 8, 2011, p. e23378.

[CSR] « Compressed sensing resources at Rice university ». `http://www.dsp.ece.rice.edu/cs/`.

[CT06] Candès (E.) et Tao (T.), « Near-optimal signal recovery from random projections : Universal encoding strategies ? », *IEEE Trans. on Info. Theory*, vol. 52, n° 12, 2006.

[CVB+10] Cattuto (C.), Van den Broeck (W.), Barrat (A.), Colizza (V.), Pinton (J.-F.) et Vespignani (A.), « Dynamics of person-to-person interactions from distributed rfid sensor networks », *PLoS ONE*, vol. 5, 2010, p. e11596.

[Dau92] Daubechies (I.), *Ten lectures on wavelets*. SIAM, 1992.

[DF79] Dickey (D. A.) et Fuller (W. A.), « Distribution of the estimators for autoregressive time series with a unit root », *Journal of the American Statistical Association*, vol. 74, n° 366, 1979.

[DFL08] Daubechies (I.), Fornasier (M.) et Loris (I.), « Accelerated projected gradient method for linear inverse problems with sparsity constraints », *Journal of Fourier Analysis and Applications*, vol. 14, n° 5-6, 2008, p. 764–792.

[DFVN12] Dong (X.), Frossard (P.), Vandergheynst (P.) et Nefedov (N.), « Clustering with multi-layer graphs : A spectral perspective », *IEEE Transactions on Signal Processing*, vol. 60, n° 11, 2012.

[DFVN13] Dong (X.), Frossard (P.), Vandergheynst (P.) et Nefedov (N.), « Clustering on multi-layer graphs via subspace analysis on grassmann manifolds », dans *IEEE Global Conference on Signal and Information Processing*, 2013.

[DFVN14] Dong (X.), Frossard (P.), Vandergheynst (P.) et Nefedov (N.), « Clustering on multi-layer graphs via subspace analysis on grassmann manifolds », *IEEE Transactions on Signal Processing*, vol. 62, n° 4, 2014.

[DG02] Davy (M.) et Godsill (S.), « Detection of abrupt signal changes using support vector machines : An application to audio signal segmentation », dans *Proceedings of IEEE ICASSP-02*, Orlando (FL, USA), 2002.

[DLW11] Daubechies (I.), Lu (J.) et Wu (H.-T.), « Synchrosqueezed wavelet transforms : an Empirical Mode Decomposition-like tool », *Appl. and Comp. Harm. Analysis*, vol. 30, n° 2, 2011, p. 243–261.

[DM96] Daubechies (I.) et Maès (S.), « A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models », dans Aldroubi (A.) et Unser (M.), éditeurs, *Wavelets in Medicine and Biology*, p. 527–546. CRC Press, Boca Raton, FL, 1996.

[DM03] Dorogovtsev (S. N.) et Mendes (J. F. F.), *Evolution of networks : From biological nets to the Internet and WWW*. Oxford University Press, Oxford, 2003.

[DM12] Dali (L.) et Mladenic (D.), « BICIKELJ : Environmental data mining on the bicycle », dans *Information Technology Interfaces (ITI), Proceedings of the ITI 2012 34th International Conference on*, p. 331–336, June 2012.

[DMS01] Dorogovtsev (S. N.), Mendes (J. F. F.) et Samukhin (A. N.), « Size-dependent degree distribution of a scale-free growing network », *Phys. Rev. E*, vol. 63, 2001, p. 062101.

[DN93] Dietrich (C.) et Newsam (G.), « A fast and exact method for multidimensional gaussian stochastic simulations », *Water Resourc. Res.*, 1993, p. 2861–2869.

[Don06] Donoho (D. L.), « Compressed sensing », *IEEE Trans. on Info. Theory*, vol. 52, n° 4, 2006.

[Doo67] Doob (J.), *Stochastic Processes*. Wiley and Sons, 1967.

[DOT03a] Doukhan (P.), Oppenheim (G.) et Taqqu (M.), *Long-Range Dependence : Theory and Applications*. Birkhäuser, Boston, 2003.

[DOT03b] Doukhan (P.), Oppenheim (G.) et Taqqu (M.), *Long-Range Dependence : Theory and Applications*. Birkhäuser, Boston, 2003.

[DS98] Draper (N. R.) et Smith (H.), *Applied Regression Analysis*. Wiley Series in Probability and Statistics, 1998.

[EAM06] Erman (J.), Arlitt (M.) et Mahanti (A.), « Traffic Classification Using Clustering Algorithms », *ACM SIGCOMM'06 MINENET*, 2006, p. 281–286.

[Efr82] Efron (B.), *The jackknife, the bootstrap, and other resampling plans*, vol. 38. Society for Industrial and Applied Mathematics Philadelphia, 1982.

[EHJT04] Efron (B.), Hastie (T.), Johnstone (I.) et Tibshirani (R.), « Least angle regression », *The Annals of Statistics*, vol. 32, n° 2, 2004, p. 407–499.

[EHP00] Evans (M.), Hastings (N.) et Peacock (B.), *Statistical Distributions*. Wiley (Interscience Division), juin 2000.

[EK10] Easley (D.) et Kleinberg (J.), *Networks, Crowds, and Markets : Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[EMVR06] Ehrhardt (G. C. M. A.), Marsili (M.) et Vega-Redondo (F.), « Phenomenological models of socioeconomic network dynamics », *Phys. Rev. E*, vol. 74, 2006, p. 036106.

[ENW96] Erramilli (A.), Narayan (O.) et Willinger (W.), « Experimental queueing analysis with long-range dependent packet traffic », *ACM/IEEE transactions on Networking*, vol. 4, n° 2, 1996, p. 209–223.

[EP06] Eagle (N.) et Pentland (A.), « Reality mining : Sensing complex social systems », *Personal and Ubiquitous Computing*, vol. 10, 2006, p. 255–268.

[ERS+10] Espanol (V. C.), Ros (P. B.), Simó (M. S.), Dainotti (A.), Donato (W. D.) et Pescapé (A.), « K-dimensional trees for continuous traffic classification », *TMA 2010*, 2010, p. 14.

[FACM03] Flandrin (P.), Auger (F.) et Chassande-Mottin (E.), « Time-Frequency reassignment — From principles to algorithms », dans Papandreou-Suppappola (A.), éditeur, *Applications in Time-Frequency Signal Processing*, chap. 5, p. 179–203. CRC Press, Boca Raton, FL, 2003.

[FB07] Fortunato (S.) et Barthelemy (M.), « Resolution limit in community detection », *PNAS*, vol. 104, n° 1, 2007, p. 36.

[FCF+11] Friggeri (A.), Chelius (G.), Fleury (E.), Fraboulet (A.), Mentré (F.) et Lucet (J.-C.), « Reconstructing Social Interactions Using an unreliable Wireless Sensor Network », *Computer Communications*, vol. 34, n° 5, avril 2011, p. 609–618.

[FF11] Fontugne (R.) et Fukuda (K.), « A Hough-transform-based anomaly detector with an adaptive time interval », *ACM SAC '11*, 2011.

[FG13] Fricker (C.) et Gast (N.). « Incentives and regulations in bike-sharing systems with stations of finite capacity ». Submitted in European Journal of Transportation and Logistics, arXiv :1201.1178v1, avril 2013.

[FGW98] Feldmann (A.), Gilbert (A.) et Willinger (W.), « Data networks as cascades : Investigating the multifractal nature of internet wan traffic », dans *SIGCOMM*, 1998.

[FHT07] Friedman (J.), Hastie (T.) et Tibshirani (R.), « Sparse inverse covariance estimation with the graphical lasso », *Biostatistics*, vol. 9, 2007, p. 432–441.

[FI10] Févotte (C.) et Idier (J.), « Algorithms for nonnegative matrix factorization with the beta-divergence », *CoRR*, vol. abs/1010.1763, 2010.

[FKMc04] Fomenkov (M.), Keys (K.), Moore (D.) et k claffy, « Longitudinal study of internet traffic in 1998-2003 », dans *WISICT'04*, 2004.

[Fla99] Flandrin (P.), *Time-Frequency / Time-Scale Analysis*. Academic Press, 1999.

[FMG12] Fricker (C.), Mohamed (H.) et Gast (N.), « Mean field analysis for inhomogeneous bike sharing systems », dans *DMTCS Proceedings, Proceedings of AofA'12*, juin 2012.

[FNO08] Froehlich (J.), Neumann (J.) et Oliver (N.), « Measuring the pulse of the city through shared bicycle programs », dans *International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems - UrbanSense08*, novembre 2008.

[FNO09] Froehlich (J.), Neumann (J.) et Oliver (N.), « Sensing and predicting the pulse of the city through shared bicycling », dans *Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, juillet 2009.

[FNW07] Figueiredo (M.), Nowak (R.) et Wright (S.), « Gradient projection for sparse reconstruction : Application to compressed sensing and other inverse problems », *IEEE J. of Select. Topics in Signal Proc. : Special Issue on Convex Optimization Methods for Signal Processing*, vol. 1, n° 4, 2007.

[For10] Fortunato (S.), « Community detection in graphs », *Physics Reports*, vol. 486, n° 3-5, 2010, p. 75–174.

[Fue05] Fuentes (M.), « A formal test for nonstationarity of spatial stochastic processes », *J. Multivariate Analysis*, vol. 96, 2005, p. 20–54.

[GBB09] Gautreau (A.), Barrat (A.) et Barthélemy (M.), « Microdynamics in stationary complex networks », *Proc. Natl. Acad. Sci. USA*, vol. 106, 2009, p. 8847.

[GCHM06] Grikschat (S.), Costa (J. A.), and. Hero (A. O.) et Michel (O.), « Dual rooted-diffusions for clustering and classification on manifolds », *IEEE ICASSP'06*, May 2006.

[GG84] Geman (S.) et Geman (D.), « Stochastic relaxation, gibbs distributions, and the bayesian restoration of images », *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-6, n° 6, Nov 1984, p. 721–741.

[GG99] Giudici (P.) et Green (P.), « Decomposable graphical gaussian model determination », *Biometrika*, vol. 86, 1999, p. 785–801.

[GGF13] Girault (B.), Gonçalves (P.) et Fleury (E.), « Graphe de contacts et ondelettes : étude d'une diffusion bactérienne », dans *Proceedings of the 24th Colloquium GRETSI*, 2013.

[GGF14] Girault (B.), Gonçalves (P.) et Fleury (E.). « Apprentissage semi-supervisé pour la transformation graphe vers signal : Interprétation en termes de filtrage de signaux sur graphe ». Journées Rescom 2013, 2014.

[GH10] Grindrod (P.) et Higham (D. J.), « Evolving graphs : dynamical models, inverse problems and propagation », *Proc. of the Royal Society A : Math., Phys. and Eng. Sc.*, vol. 466, n° 2115, 2010, p. 753–770.

[Gir08] Girardin (F.). « Revealing paris through velib' data ». http ://liftlab.com/think/fabien/2008/02/27/revealing-paris-through-velib-data/, 2008.

[GMC+09] Galluccio (L.), Michel (O.), Comon (P.), Hero (A. O.) et Kliger (M.), « Combining multiple partitions created with a graph-based construction for data clustering », *IEEE Int. Workshop on Machine Learning for Signal Processing*, September 2009.

[GMTA05] Guimerà (R.), Mossa (S.), Turtschi (A.) et Amaral (L. A. N.), « The worldwide air transportation network : Anomalous centrality, community structure, and cities' global roles », *Proc. Natl. Acad. Sci. USA*, vol. 102, n° 22, mai 2005, p. 7794–7799.

[GNP06] Gardner (W. A.), Napolitano (A.) et Paura (L.), « Cyclostationarity : Half a century of research », *Signal Processing*, vol. 86, n° 4, 2006, p. 639 – 697.

[GO09] Goldstein (T.) et Osher (S.), « The split bregman method for ℓ_1 regularized problems », *SIAM J. Imaging Sci.*, vol. 2, 2009.

[GP11] Grindrod (P.) et Parsons (M. C.), « Social networks : evolving graphs with memory dependent edges », *Physica A*, vol. 390, 2011, p. 3970.

[GPS89] Greig (D.), Porteous (B.) et Seheult (A.), « Exact maximum a posteriori estimation for binary images », *Journal of the Royal Statistical Society Series B*, vol. 51, 1989.

[GS08] Gross (T.) et Sayama (H.), éditeurs, *Adaptive Networks : Theory, Models and Applications*, coll. « Springer/NECSI Studies on Complexity Series ». Spreinger, 2008.

[GUT] GUTP. « Green University of Tokyo Project ». `http://www.gutp.jp/`.

[HA85] Hubert (L.) et Arabie (P.), « Comparing partitions », *Journal of classification*, vol. 2, n° 1, 1985, p. 193–218.

[Har89] Harvey (A.), *Forecasting, Structural time series model and the Kalman filter*. Cambridge University Press, 1989.

[HB10] Hill (S.) et Braha (D.), « Dynamic model of time-dependent complex networks », *Phys. Rev. E*, vol. 82, 2010, p. 046105.

[HBFR13] Hamon (R.), Borgnat (P.), Flandrin (P.) et Robardet (C.), « Relabeling nodes according to the structure of the graph ». Rapport technique, ENS de Lyon, INSA de Lyon, 2013. http ://perso.ens-lyon.fr/ronan.hamon/files/relabeling.pdf.

[HCS+05] Hui (P.), Chaintreau (A.), Scott (J.), Gass (R.), Crowcroft (J.) et Diot (C.), « Pocket switched networks and human mobility in conference environments », dans *ACM SIGCOMM workshop on Delay-tolerant networking*, p. 244 – 251, 2005.

[HF97] Hlawatsch (F.) et Flandrin (P.), « The interference structure of wigner distribution and related time-frequency representations », dans Mecklenbräuker (W.) et Hlawatsch (F.), éditeurs, *The Wigner Distribution — Theory and Applications in Signal Processing*. Elsevier, 1997.

[HFO04] Hobijn (B.), Franses (P.) et Ooms (M.), « Generalization of the kpss-test for stationarity », *Statistica Neerlandica*, vol. 58, 2004, p. 482–502.

[HHP03] Hussain (A.), Heidemann (J.) et Papadopoulos (C.), « A framework for classifying denial of service attacks », dans *SIGCOMM*, Karlsruhe, Germany, 2003.

[HK02] Holme (P.) et Kim (B. J.), « Growing scale-free networks with tunable clustering », *Phys. Rev. E*, vol. 65, 2002, p. 026107.

[HN06] Holme (P.) et Newman (M.), « Nonequilibrium phase transition in the coevolution of networks and opinions », *Phys. Rev. E*, vol. 74, 2006, p. 056108.

[HP97] Hodrick (R. J.) et Prescott (E. C.), « Postwar U.S. business cycles : An empirical investigation », *Journal of Money, Credit, and Banking*, vol. 29, n° 1, 1997, p. 1–16.

[HPA11a] Helgason (H.), Pipiras (V.) et Abry (P.), « Fast and exact synthesis of stationary multivariate Gaussian time series using circulant embedding », *Signal Processing*, vol. 95, n° 5, 2011, p. 1123–1133.

[HPA11b] Helgason (H.), Pipiras (V.) et Abry (P.), « Synthesis of multivariate stationary series with prescribed marginal distributions and covariance using circulant embedding », *Signal Processing*, vol. 91, n° 8, 2011, p. 1741–1758.

[HR09] Huber (P. J.) et Ronchetti (E. M.), *Robust Statistics*, coll. « Wiley Series in Probability and Statistics ». Wiley, 2009.

[HS05] Huang (N. E.) et Shen (S. S.), *Hilbert-Huang transform and its applications*, vol. 5. World Scientific, Singapore, 2005.

[HS12] Holme (P.) et Saramäki (J.), « Temporal networks », *Physics Reports*, vol. 519, n° 3, 2012, p. 97–125.

[HS13] Holme (P.) et Saramäki (J.), éditeurs, *Temporal Networks*. Springer, 2013.

[HSL+98] Huang (N. E.), Shen (Z.), Long (S. R.), Wu (M. L.), Shih (H. H.), Zheng (Q.), Yen (N. C.), Tung (C. C.) et Liu (H. H.), « The Empirical Mode Decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis », *Proc. Roy. Soc. London A*, vol. 454, 1998, p. 903–995.

[HTF+01] Hastie (T.), Tibshirani (R.), Friedman (J.) et others. « The elements of statistical learning : data mining, inference, and prediction », 2001.

[HVA03] Hohn (N.), Veitch (D.) et Abry (P.), « Cluster processes, a natural language for network traffic », *IEEE Transactions on Signal Processing Special Issue on Signal Processing in Networking*, vol. 8, n° 51, octobre 2003, p. 2229–2244.

[HVA05] Hohn (N.), Veitch (D.) et Abry (P.), « Multifractality in TCP/IP traffic : the case against », *Journal Comp. Networks*, vol. 48, n° 3, 2005, p. 293–313.

[HVG11] Hammond (D.), Vandergheynst (P.) et Gribonval (R.), « Wavelets on graphs via spectral graph theory », *Applied and Computational Harmonic Analysis*, vol. 30, n° 2, 2011, p. 129–150.

[IFM09] Iliofotou (M.), Faloutsos (M.) et Mitzenmacher (M.), « Exploiting Dynamicity in Graph-based Traffic Analysis : Techniques and Applications », *ACM CoNEXT 2009*, 2009, p. 241–252.

[IGER+10] Iliofotou (M.), Gallagher (B.), E.-Rad (T.), Xie (G.) et Faloutsos (M. V.), « Profiling-by-Association : A Resilient Traffic Profiling Solution for the Internet Backbone », *ACM CoNEXT 2010*, 2010, p. 12.

[IRB+11] Isella (L.), Romano (M.), Barrat (A.), Cattuto (C.), Colizza (V.), den Broeck (W. V.), Gesualdo (F.), Pandolfi (E.), Ravà (L.), Rizzo (C.) et Tozzi (A.), « Close encounters in a pediatric ward : measuring face-to-face proximity and mixing patterns with wearable sensors », *PLoS ONE*, vol. 6, 2011, p. e17144.

[ISB+11] Isella (L.), Stehlé (J.), Barrat (A.), Cattuto (C.), Pinton (J.) et Van den Broeck (W.), « What's in a crowd ? analysis of face-to-face behavioral networks », *Journal of theoretical biology*, vol. 271, n° 1, 2011, p. 166–180.

[JHVB12] Jacomy (M.), Heymann (S.), Venturini (T.) et Bastian (M.). « Algorithm for handy network visualization ». www.gephi.org, août 2012.

[JKR02] Jung (J.), Krishnamurthy (B.) et Rabinovich (M.), « Flash Crowds and Denial of Service Attacks : Characterization and Implications for CDNs and Web Sites », dans *International WWW Conference*, Honolulu, HI, mai 2002.

[JROR10] Jensen (P.), Rouquier (J.-B.), Ovtracht (N.) et Robardet (C.), « Characterizing the speed and paths of shared bicycles in lyon », *Transportation Research Part D : Transport and Environment*, vol. 15, n° 8, 2010, p. 522–524.

[JSZ09] Jin (Y.), Sharafuddin (E.) et Zhang (Z. L.), « Unveiling Core Network-Wide Communication Patterns through Application Traffic Activity Graph Decomposition », *ACM SIGMETRICS'09*, 2009, p. 49–60.

[JTH09a] Jung (A.), Tauböck (G.) et Hlawatsch (F.), « Compressive nonstationary spectral estimation using parsimonious random sampling of the ambiguity function », dans *Proc. IEEE Statistical Signal Processing Workshop SSP-09*, 2009.

[JTH09b] Jung (A.), Tauböck (G.) et Hlawatsch (F.), « Compressive spectral estimation for nonstationary random processes », dans *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Proc. ICASSP-09*, 2009.

[JTH13] Jung (A.), Tauböck (G.) et Hlawatsch (F.), « Compressive spectral estimation for nonstationary random processes », *IEEE Trans. Inform. Theory*, vol. 59, n° 5, mai 2013.

[Kay08] Kay (S.), « A new nonstationary detector », *IEEE Trans. Signal Process.*, vol. 56, n° 4, 2008, p. 1440–1451.

[KB05a] Katipamula (S.) et Brambley (M. R.), « Methods for fault detection, diagnostics, and prognostics for building systems, a review, part I », *HVAC&R Research*, vol. 11, n° 1, 2005, p. 3–25.

[KB05b] Katipamula (S.) et Brambley (M. R.), « Methods for fault detection, diagnostics, and prognostics for building systems, a review, part II », *HVAC&R Research*, vol. 11, n° 2, 2005, p. 169–187.

[KB08] Kozma (B.) et Barrat (A.), « Consensus formation on adaptive networks », *Phys. Rev. E*, vol. 77, 2008, p. 016102.

[KBB+04] Karagiannis (T.), Broido (A.), Brownlee (N.), kc claffy et Faloutsos (M.), « Is P2P dying or just hiding ? », dans *IEEE GLOBECOM'04*, 2004.

[KcF+08] Kim (H.), kc claffy, Fomenkov (M.), Barman (D.), Faloutsos (M.) et Lee (K. Y.), « Internet Traffic Classification Demystified : Myths, Caveats, and the Best Practices », *ACM CoNEXT 2008*, 2008, p. 12.

[Key06] Keylock (C.), « Constrained surrogate time series with preservation of the mean and variance structure », *Phys. Rev. E*, vol. 73, 2006, p. 030767.1–030767.4.

[KG11] Kepner (J.) et Gilbert (J.), éditeurs, *Graph algorithms in the language of linear algebra*. SIAM, Philadelphia, 2011.

[KKB+12] Krings (G.), Karsai (M.), Bernhardsson (S.), Blondel (V.) et Saramäki (J.), « Effects of time window size and placement on the structure of an aggregated communication network », *EPJ Data Science*, vol. 1, n° 4, 2012, p. 1–16.

[KKBG09] Kim (S. J.), Koh (K.), Boyd (S.) et Gorinevsky (D.), « l_1 trend filtering », *SIAM Review*, vol. 51, n° 2, 2009, p. 339–360.

[KKK+11] Kovanen (L.), Karsai (M.), Kaski (K.), Kertész (J.) et Saramäki (J.), « Temporal motifs in time-dependent networks », *eprint*, 2011, p. arXiv :1107.5646.

[KKP+11] Karsai (M.), Kivelä (M.), Pan (R. K.), Kaski (K.), Kertész (J.), Barabási (A.) et Saramäki (J.), « Small But Slow World : How Network Topology and Burstiness Slow Down Spreading », *Phys Rev E*, vol. 83, 2011, p. 025102(R).

[KKR+99] Kleinberg (J. M.), Kumar (R.), Raghavan (P.), Rajagopalan (S.) et Tomkins (A. S.), « The Web as a graph : Measurements, models and methods », *Lecture Notes in Computer Science*, vol. 1627, 1999, p. 1–18.

[KMF04] Karagiannis (T.), Molle (M.) et Faloutsos (M.), « Long-range dependence - ten years of internet traffic modeling », *IEEE Internet Computing*, septembre 2004.

[KMFB04] Karagiannis (T.), Molle (M.), Faloutsos (M.) et Broido (A.), « A nonstationary poisson view of internet traffic », dans *IEEE INFOCOM'04*, 2004.

[Kol09] Kolaczyk (E.), *Statistical Analysis of Network Data : Methods and Models*. Springer, 2009.

[KPF05] Karagiannis (T.), Papagiannaki (K.) et Faloutsos (M.), « BLINC : Multilevel Traffic Classification in the Dark », *ACM SIGCOMM'05*, 2005, p. 229–240.

[KPSS92] Kwiatkowski (D.), Phllips (P.), Schmidt (P.) et Shin (Y.), « Testing the null hypothesis of stationarity against the alternative of a unit root », *J. of Econometrics*, vol. 54, 1992, p. 159–178.

[KRR+00] Kumar (R.), Raghavan (P.), Rajagopalan (S.), Sivakumar (D.), Tomkins (A.) et Upfal (E.), « Stochastic models for the Web graph », *Proceedings of the 41th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2000, p. 57 – 65.

[KSKK07] Kumpula (J.), Saramäki (J.), Kaski (K.) et Kertesz (J.), « Limited resolution in complex network community detection with Potts model approach », *Eur. Phys. J. B*, vol. 56, n° 1, 2007, p. 41–45.

[KSZC03] Krishnamurty (B.), Sen (S.), Zhang (Y.) et Chen (Y.), « Sketch-based change detection : Methods, evaluation, and applications », dans *ACM IMC*, octobre 2003.

[l7-] « l7-filter ». http ://l7-filter.sourceforge.net.

[LAC12] Lathia (N.), Ahmed (S.) et Capra (L.), « "measuring the impact of opening the london shared bicycle scheme to casual users », *Transportation Research Part C : Emerging Technologies*, vol. 22, 2012.

[Lah99] Lahiri (S.), « Theoretical comparisons of block bootstrap methods », *The Annals of Statistics*, vol. 27, n° 1, 1999, p. 386–404.

[Lam10] Lambiotte (R.), « Multi-scale modularity in complex networks », dans *Proc. 8th Int. Symp. WiOpt*, p. 546–553, 2010.

[Lat07] Latapy (M.). « Grands graphes de terrain – mesure et métrologie, analyse, modélisation, algorithmique ». Habilitation à Diriger des Recherches. Université Pierre et Marie Curie., 2007.

[LBC+06] Li (X.), Bian (F.), Crovella (M.), Diot (C.), Govindan (R.), Iannaccone (G.) et Lakhina— (A.), « Detection and identification of network anomalies using sketch subspaces », dans *ACM IMC*, octobre 2006.

[LCD04] Lakhina (A.), Crovella (M.) et Diot (C.), « Diagnosing network-wide traffic anomalies », dans *SIGCOMM*, août 2004.

[LCD05] Lakhina (A.), Crovella (M.) et Diot (C.), « Mining Anomalies Using Traffic Feature Distributions », *ACM SIGCOMM'05*, 2005, p. 217–228.

[LCL+11] Laouénan (C.), Chelius (G.), Lepelletier (D.), Fleury (E.), Mentré (F.) et Lucet (J.-C.), « Mesurer les contacts entre soignants et patients au moyen de capteurs électroniques : le cas de la tuberculose », *Revue d'Epidémiologie et de Santé Publique*, vol. S1, avril 2011, p. 33. Doi : 10.1016/j.respe.2011.02.055.

[LD98] Laurent (H.) et Doncarli (C.), « Stationarity index for abrupt changes detection in the time-frequency plane », *IEEE Signal Proc. Lett.*, vol. 24, n° 1, 1998, p. 43–45.

[LF09] Lancichinetti (A.) et Fortunato (S.), « Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities », *Physical Review E*, n° 1, 2009.

[LG12] Lezoray (O.) et Grady (L.), éditeurs, *Image Processing and Analysis with Graphs. Theory and Practice*. CRC Press, 2012.

[liv09] Homepage of the Live E ! Project : `http://www.live-e.org/`, 2009.

[Lju99] Ljung (L.), *System identification : theory for the user*, chap. 10.2. PTR Prentice Hall, 1999.

[LKJ+10] Lim (Y.), Kim (H.), Jeong (J.), Kim (C.), Kwon (T. T.) et Choi (Y.), « Internet Traffic Classification Demystified : On the Sources of the Discriminative Power », *ACM CoNEXT 2010*, 2010, p. 12.

[LLC+12] Lucet (J.-C.), Laouenan (C.), Chelius (G.), Veziris (N.), Lepelletier (D.), Friggeri (A.), Abiteboul (D.), Bouvet (E.), Mentré (F.) et Fleury (E.), « Electronic Sensors for Assessing Interactions between Healthcare Workers and Patients under Airborne Precautions », *PLoS ONE*, vol. 7, n° 5, mai 2012.

[LM02] Latora (V.) et Marchiori (M.), « Is the boston subway a small-world network ? », *Physica A*, vol. 314, 2002, p. 109–113.

[LM08] Latapy (M.) et Magnien (C.), « Complex network measurements : Estimating the relevance of observed properties », dans *Infocom'08*, Phoenix, USA, 2008.

[LMC05] Lesieur (M.), Métais (O.) et Comte (P.), *Large-Eddy Simulations of Turbulence*. Cambridge University Press, Cambridge (UK), 2005.

[Loè62] Loève (M.), *Probability Theory*. D. von Nostrand Comp., 1962.

[LS99] Lee (D. D.) et Seung (H. S.), « Learning the parts of objects by non-negative matrix factorization », *Nature*, vol. 401, n° 6755, octobre 1999, p. 788–791.

[LTSB07] Lévêque (E.), Toschi (F.), Shao (L.) et Bertoglio (J.-P.), « Shear-improved smagorinsky model for large-eddy simulation of wall-bounded turbulent flows », *Journal of Fluid Mechanics*, vol. 570, 2007, p. 491–502.

[LTWW93] Leland (W. E.), Taqqu (M. S.), Willinger (W.) et Wilson (D. V.), « On the self-similar nature of ethernet traffic », dans *SIGCOMM '93 : Conference proceedings on Communications architectures, protocols and applications*, p. 183–193, New York, NY, USA, 1993. ACM Press.

[LTWW94a] Leland (W. E.), Taqqu (M. S.), Willinger (W.) et Wilson (D. V.), « On the self-similar nature of ethernet traffic (extended version) », *ACM/IEEE Transactions on Networking*, vol. 2, n° 1, février 1994, p. 1–15.

[LTWW94b] Leland (W. E.), Taqqu (M. S.), Willinger (W.) et Wilson (D. V.), « On the self-similar nature of ethernet traffic (extended version) », *ACM/IEEE transactions on Networking*, vol. 2, n° 1, février 1994, p. 1–15.

[LV13] Leonardi (N.) et Ville (D. V. D.), « Tight wavelet frames on multisclice graphs », *IEEE Trans. on Signal Processing*, vol. 61, n° 13, 2013, p. 3357–3367.

[Mag10] Magnien (C.). « Intégrer mesure, métrologie et analyse pour l'étude des graphes de terrain dynamiques ». Habilitation à Diriger des Recherches. Université Pierre et Marie Curie., 2010.

[Mal99] Mallat (S.), *A Wavelet tour of signal processing*. Academic Press, 1999.

[Mar84] Martin (W.), « Measuring the degree of non-stationarity by using the Wigner-Ville spectrum », dans *Proc. IEEE ICASSP-84*, p. 41B.3.1–41B.3.4, San Diego (CA), 1984.

[MAW] MAWI. « Mawi traffic archive ». WIDE Project, http ://mawi.wide.ad.jp/mawi/.

[MB06a] Marrelec (G.) et Benali (H.), « Asymptotic bayesian structure learning using graph supports for gaussian graphical models », *Journal of Multivariate Analysis*, vol. 97, 2006, p. 1451–1466.

[MB06b] Meinshausen (N.) et Buhlmann (P.), « High-dimensional graphs and variable selection with the lasso », *The Annals of Statistics*, vol. 34, 2006, p. 1435–1462.

[Mel93] Melamed (B.), « An overview of TES processes and modeling methodology », dans *Performance/SIGMETRICS Tutorials*, p. 359–393, 1993.

[Mer04] Mercklé (P.), *Les réseaux sociaux, les origines de l'analyse des réseaux sociaux*. CNED, ens-lsh, 2004.

[Mey90] Meyer (Y.), *Ondelettes et opérateurs*. Hermann, 1990.

[MF85] Martin (W.) et Flandrin (P.), « Detection of changes of signal structure by using the Wigner-Ville spectrum », *Signal Proc.*, vol. 8, 1985, p. 215–233.

[MH97] Mecklenbräuker (W.) et Hlawatsch (F.), éditeurs, *The Wigner Distribution – Theory and Applications in Signal Processing*. Elsevier, Amsterdam (The Netherlands), 1997.

[MH11] Miller (J.) et Hagberg (A.), « Efficient generation of networks with given expected degrees », *Algorithms and Models for the Web Graph*, 2011, p. 115–126.

[Mie11] Mieghem (P. V.), *Graph Spectra for Complex Networks*. Cambridge University Press, New York, NY, USA, 2011.

[MIO+07] Matsuura (S.), Ishizuka (H.), Ochiai (H.), Doi (S.), Ishida (S.), Nakayama (M.), Esaki (H.) et Sunahara. (H.), « Live e! project : Establishment of infrastructure sharing environmental information », dans *Applications and the Internet Workshops, 2007. SAINT Workshops 2007. International Symposium on*, p. pp.67–67, Jan. 2007.

[MKH07] Maraun (D.), Kurths (J.) et Holschneider (M.), « Nonstationary gaussian processes in the wavelet domain : Synthesis, estimation and significance testing », *Phys. Rev. E*, vol. 75, 2007, p. 016707.1–016707.14.

[MML11] Miritello (G.), Moro (E.) et Lara (R.), « Dynamical strength of social ties in information spreading », *Phys. Rev. E*, vol. 83, n° 4, 2011, p. 045102.

[MMN08] McHugh (J.), McLeod (R.) et Nagaonkar (V.), « Passive network forensics : behavioural classification of network hosts based on connection patterns », *ACM SIGOPS Operating Systems Review, Vol.42, No.3*, 2008, p. 99–111.

[MR98] Molloy (M.) et Reed (B.), « The size of the giant component of a random graph with a given degree distribution », *Combinatorics, Probab. Comput.*, vol. 7, 1998, p. 295.

[MR10] Merchez (L.) et Rouquier (J.-B.), « L'usage des vélos en libre service (VLS) comme révélateur des rythmes urbains : le cas des stations de vélo'v à lyon », *Données Urbaines*, 2010.

[Mut03] Muthukrishnan (S.), « Data streams : Algorithms and applications », dans *ACM SIAM SODA*, janvier 2003.

[MVS01] Moore (D.), Voelker (G.) et Savage (S.), « Inferring internet denial-of-service activity », dans *Usenix Security Symposium*, 2001.

[MZ05] Moore (A. W.) et Zuev (D.), « Internet traffic classification using bayesian analysis techniques », *ACM SIGMETRICS'05*, 2005, p. 50–60.

[NCB09] Nguyen (K. N. T.), Cerf (L.) et Boulicaut (J.), « Discovering relevant cross-graph cliques in dynamic networks », dans *Proc. ISMIS'09. Springer LNCS*, vol. 5722, p. 513–522, 2009.

[NCPB11] Nguyen (K. N. T.), Cerf (L.), Plantevit (M.) et Boulicaut (J.), « Multidimensional association rules in boolean tensors », dans *Proc. SIAM Int. Conf. on Data Mining SDM'11*, p. 570–581, Phoenix, USA, avril 2011.

[NCPB13] Nguyen (K. N. T.), Cerf (L.), Plantevit (M.) et Boulicaut (J.), « Discovering descriptive rules in relational dynamic graphs », *Intelligent Data Analysis*, vol. 17, n° 1, 2013, p. 49–69.

[New03] Newman (M. E. J.), « The structure and function of complex networks », *SIAM Review*, vol. 45, 2003, p. 167–256.

[New06] Newman (M.), « Modularity and community structure in networks », *PNAS*, vol. 103, n° 23, 2006, p. 8577.

[New10] Newman (M.), *Networks : An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[NKB08] Nardini (C.), Kozma (B.) et Barrat (A.), « Who's talking first ? consensus or lack thereof in coevolving opinion formation models », *Phys. Rev. Lett.*, vol. 100, 2008, p. 158701.

[NMHHB13] Nair (R.), Miller-Hooks (E.), Hampshire (R.) et Busic (A.), « Large-scale bicycle sharing systems : analysis of vélib », *International Journal of Sustainable Transportation*, vol. 7, n° 1, 2013.

[NO09] Narang (S. K.) et Ortega (A.), « Lifting based wavelet transforms on graphs », dans *Proc. APSIPA ASC*, octobre 2009.

[NSW13] Nakatsukasa (Y.), Saito (N.) et Woei (E.), « Mysteries around the graph laplacian eigenvalue 4 », *Linear Algebra and its Applications*, vol. 438, n° 8, 2013, p. 3231 – 3246.

[NT09] Needell (D.) et Tropp (J. A.), « Cosamp : Iterative signal recovery from incomplete and inaccurate samples », *Appl. Comp. Harm. Anal.*, vol. 26, n° 3, 2009.

[OLB+07] Owezarski (P.), Larrieu (N.), Bernaille (L.), Saddi (W.), Guillemin (F.), Soule (A.) et Salamatian (K.), « Distribution of traffic among applications as measured in the french metropolis project », *Annals of Telecommunication, Special issue on Analysis of traffic and usage traces on the Internet - From network engineering to sociology of uses*, 2007.

[ope] « OpenDPI ». http ://opendpi.org/.

[P.11] P. (M.), « Bicycle'sharing schemes : Enhancing sustainable mobility in urban areas », *Commission on Sustainable Development, Department of Economic and Social Affairs, United Nations, New-York, Background Paper*, vol. 8, mai 2011.

[PCUKEN09] Pietrzyk (M.), Costeux (J. L.), Urvoy-Keller (G.) et En-Najjary (T.), « Challenging statistical classification for operational usage : the ADSL case », *ACM IMC'09*, 2009, p. 122–135.

[PDC+10] Parshani (R.), Dickison (M.), Cohen (R.), Stanley (H. E.) et Havlin (S.), « Dynamic networks and directed percolation », *Europhys. Lett.*, vol. 90, 2010, p. 38004.

[Pen08] Pentland (A.), *Honest Signals : how they shape our world*. MIT Press, 2008.

[PF95] Paxson (V.) et Floyd (S.), « Wide-area traffic : The failure of Poisson modeling », *ACM/IEEE transactions on Networking*, vol. 3, n° 3, juin 1995, p. 226–244.

[PKC96] Park (K.), Kim (G.) et Crovella (M.), « On the relationship between file sizes, transport protocols, and self-similar network traffic », dans *International Conference on Network Protocols*, p. 171, Washington, DC, USA, 1996. IEEE Computer Society.

[Pol03] Politis (D.), « The impact of bootstrap methods on time series analysis », *Statistical Science*, vol. 18, n° 2, 2003, p. 219–230.

[Pon06] Pons (P.). « Post-processing hierarchical community structures : Quality improvements and multi-scale view ». eprint arXiv :cs/0608050, 2006.

[Pri58] Price (R.), « A useful theorem for nonlinear devices having Gaussian inputs », *IRE Trans. Inform. Theory*, vol. IT-4, 1958, p. 69–72.

[Pri81] Priestley (M.), *Spectral analysis and times series*. Academic Press, San Diego, 1981.

[PS69] Priestley (M. B.) et Subba Rao (T.), « A test for non-stationarity of time-series », *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 31, n° 1, 1969, p. 140–149.

[PSV04] Pastor-Satorras (R.) et Vespignani (A.), *Evolution and structure of the Internet : A statistical physics approach*. Cambridge University Press, Cambridge, 2004.

[PT94] Prichard (D.) et Theiler (J.), « Generating surrogate data for time series with several simultaneously measured variables », *Physical Review Letters*, vol. 73, n° 7, 1994, p. 951–954.

[PTB+02] Papagiannaki (K.), Taft (N.), Bhattacharyya (S.), Thiran (P.), Salamatian (K.) et Diot (C.), « A Pragmatic Definition of Elephants in Internet Backbone Traffic », *ACM SIGCOMM IMW'02*, November 2002, p. 175–176.

[PW00] Park (K.) et Willinger (W.), « Self-similar network traffic : An overview », dans Park (K.) et Willinger (W.), éditeurs, *Self-Similar Network Traffic and Performance Evaluation*, p. 1–38. Wiley (Interscience Division), 2000.

[PWg00] Park (K.) et Willinger (W.), *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons, Inc., New York, NY, USA, 2000.

[RABV13] Richiardi (J.), Achard (S.), Bunke (H.) et Van De Ville (D.), « Machine learning with brain graphs : Predictive modeling approaches for functional imaging in systems neuroscience », *IEEE Signal Process. Mag.*, vol. 30, n° 3, 2013, p. 58–70.

[RB06] Reichardt (J.) et Bornholdt (S.), « Statistical mechanics of community detection », *Physical Review E*, vol. 74, n° 1, 2006, p. 016110.

[RB10] Rosvall (M.) et Bergstrom (C.), « Mapping change in large networks », *PloS one*, vol. 5, n° 1, 2010, p. e8694.

[RKBB11] Roth (C.), Kang (S. M.), Batty (M.) et Barthelemy (M.), « Structure of urban movements : Polycentric activity and entangled hierarchy flows », *PLoS One*, vol. 6, n° 1, janvier 2011, p. e15923.

[Roe99] Roesch (M.), « Snort - Lightweight Intrusion Detection for Networks », *USENIX LISA'99*, 1999, p. 229–238.

[Rov02] Roverato (A.), « Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models », *Scandinavian Journal of Statistics*, vol. 29, n° 3, 2002, p. 391–411.

[RPDB11] Rabin (J.), Peyré (G.), Delon (J.) et Bernot (M.), « Wasserstein barycenter and its application to texture mixing », dans *Proc. SSVM'11*, 2011.

[RZMD05] Ribeiro (V.), Zhang (Z.-L.), Moon (S.) et Diot (C.), « Small-time scaling behavior of internet backbone traffic », *Comp. Networks*, vol. 37, 2005, p. 315–334.

[SA09] Scherrer (A.) et Abry (P.), « Synthèse de processus bivariés non Gaussiens à mémoires longues », dans *22nd GRETSI Symposium on Signal and Image Processing*, Dijon, 2009.

[Sag06] Sagaut (P.), *Large eddy simulation for incompressible flows : An introduction*. Springer-Verlag, Berlin Heidelberg, 3e édition, 2006.

[SB05] Serrano (M. A.) et Boguñá (M.), « Tuning clustering in random networks with arbitrary degree distributions », *Phys. Rev. E*, vol. 72, 2005, p. 036133.

[SBB10] Stehlé (J.), Barrat (A.) et Bianconi (G.), « Dynamical and bursty interactions in social networks », *Phys. Rev. E*, vol. 81, 2010, p. 035101(R).

[SDYB12] Schaub (M.), Delvenne (J.), Yaliraki (S.) et Barahona (M.), « Markov dynamics as a zooming lens for multiscale community detection : non clique-like communities and the field-of-view limit », *PloS one*, vol. 7, n° 2, 2012, p. e32210.

[See07] Seem (J. E.), « Using intelligent data analysis to detect abnormal energy consumption in buildings », *Energy and Buildings*, vol. 39, n° 1, 2007, p. 52 – 58.

[SIS12] Shimada (Y.), Ikeguchi (T.) et Shigehara (T.), « From networks to time series », *Phys. Rev. Lett.*, vol. 109, n° 15, octobre 2012, p. 158701.

[SKL+10] Salathé (M.), Kazandjieva (M.), Lee (J. W.), Levis (P.), Feldman (M. W.) et Jones (J. H.), « A high-resolution human contact network for infectious disease transmission », *Proc. Natl. Acad. Sci. (USA)*, n° 107, decembre 2010.

[SM13] Sandryhaila (A.) et Moura (J.), « Discrete signal processing on graphs », *IEEE Trans. on Signal Processing*, vol. 361, n° 7, 2013.

[SNF+13] Shuman (D. I.), Narang (S. K.), Frossard (P.), Ortega (A.) et Vandergheynst (P.), « The emerging field of signal processing on graphs : Extending high-dimensional data analysis to networks and other irregular domains », *IEEE Signal Processing Magazine*, vol. 30, n° 3, 2013, p. 83–98.

[Soc] Sociopatterns. « www.sociopatterns.org ».

[SPGF08] Suissa (M.), Place (C.), Goillot (E.) et Freyssingeas (E.), « Internal dynamics of a living cell nucleus investigated by dynamic light scattering », *The European Physical Journal E*, vol. 26, n° 4, septembre 2008.

[SPGF09] Suissa (M.), Place (C.), Goillot (E.) et Freyssingeas (E.), « Evolution of the global internal dynamics of a living cell nucleus during interphase », *Biophysical Journal*, vol. 97, n° 2, août 2009.

[SPGMA07] Sales-Pardo (M.), Guimera (R.), Moreira (A.) et Amaral (L.), « Extracting the hierarchical organization of complex systems », *PNAS*, vol. 104, n° 39, 2007, p. 15224–15229.

[SPM13] « Special section – adaptation and learning over complex networks », *Signal Processing Magazine*, vol. 30, n° 3, 2013.

[SPSG05] Serpedin (E.), Panduru (F.), Sari (I.) et Giannakis (G.), « Bibliography on cyclostationarity », *Signal Proc.*, vol. 85, n° 12, 2005, p. 2233–2303.

[SPST+01] Schölkopf (B.), Platt (J. C.), Shawe-Taylor (J.), Smola (A. J.) et Williamson (R. C.), « Estimating the support of a high-dimensional distribution », *Neural Computation*, vol. 13, n° 7, 2001, p. 1443–1471.

[SRV13] Shuman (D.), Ricaud (B.) et Vandergheynst (P.), « Vertex-Frequency Analysis on Graphs », *Applied and Computational Harmonic Analysis*, 2013.

[SS96] Schreiber (T.) et Schmitz (A.), « Improved surrogate data for nonlinearity tests », *Phys. Rev. Lett.*, vol. 77, n° 4, 1996, p. 635–638.

[SS00] Schreiber (T.) et Schmitz (A.), « Surrogate time series », *Physica D*, vol. 142, n° 3-4, 2000, p. 346–382.

[SSM98] Schölkopf (B.), Smola (A. J.) et Müller (K.), « Nonlinear component analysis as a kernel eigenvalue problem », *Neural Computation*, vol. 10, n° 5, 1998, p. 1299–1319.

[ST94] Samorodnitsky (G.) et Taqqu (M.), *Stable Non-Gaussian Random Processes*. Chapman&Hall, 1994.

[STC04] Shawe-Taylor (J.) et Cristianini (N.), *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[SVB+11] Stehle (J.), Voirin (N.), Barrat (A.), Cattuto (C.), Isella (L.), Pinton (J.), Quaggiotto (M.), Van Den Broeck (W.), Regis (C.), Lina (B.) et Vanhems (P.), « High-resolution measurements of face-to-face contact patterns in a primary school », *PloS one*, vol. 6, n° 8, 2011, p. e23176.

[SVF11] Shuman (D.), Vandergheynst (P.) et Frossard (P.), « Chebyshev polynomial approximation for distributed signal processing », dans *International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS) 2011*, 2011.

[SWHV14] Shuman (D.), Wiesmeyr (C.), Holighaus (N.) et Vandergheynst (P.), « Spectrum-Adapted Tight Graph Wavelet and Vertex-Frequency Frames », *IEEE Transactions on Signal Processing*, 2014.

[TC84] Trussell (H. J.) et Civanlar (M. R.), « The feasible solution in signal restoration », *IEEE Trans. Acous., Speech Signal Process.*, vol. 32, n° 2, Apr. 1984, p. 201–212.

[TCR12] Tabatabaei (S.), Coates (M.) et Rabbat (M.), « Ganc : Greedy agglomerative normalized cut », *Pattern Recognition*, vol. 45, n° 2, février 2012.

[TCSF11] Torres (M. E.), Colominas (M. A.), Schlotthauer (G.) et Flandrin (P.), « A complete ensemble empirical mode decomposition with adaptive noise », dans *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, p. 4144–4147. IEEE, 2011.

[TD04] Tax (D. M. J.) et Duin (R. P. W.), « Support vector data description », *Machine Learning*, vol. 54, n° 1, 2004, p. 45–66.

[TEL+92] Theiler (J.), Eubank (S.), Longtin (A.), Galdrikian (B.) et Farmer (J. D.), « Testing for nonlinearity in time series : the method of surrogate data », *Physica D*, vol. 58, n° 1–4, 1992, p. 77–94.

[TG07] Tropp (J. A.) et Gilbert (A.), « Signal recovery from random measurements via orthogonal matching pursuit », *IEEE Trans. on Info. Theory*, vol. 53, n° 12, 2007.

[Tho82] Thomson (D. J.), « Spectrum estimation and harmonic analysis », *Proceedings of the IEEE*, vol. 70, 1982, p. 1055–1096.

[Tib96] Tibshirani. (R.), « Regression shrinkage and selection via the lasso », *J. Roy. Statist. Soc. Ser. B*, vol. 58, 1996, p. 267–288.

[TLB+09] Tournoux (P.-H.), Leguay (J.), Benbadis (F.), Conan (V.), de Amorim (M. D.) et Whitbeck (J.), « The accordion phenomenon : Analysis, characterization, and impact on DTN routing », *IEEE INFOCOM 2009*, avril 2009.

[TPGK03] Tan (G.), Poletto (M.), Guttag (J.) et Kaashoek (F.), « Role Classification of Hosts within Enterprise Networks Based on Connection Patterns », *2003 USENIX Annual Technical Conference*, 2003, p. 15–28.

[TRKN08] Trestian (I.), Ranjan (S.), Kuzmanovic (A.) et Nucci (A.), « Unconstrained Endpoint Profiling (Googling the Internet) », *ACM SIGCOMM'08*, 2008, p. 279–290.

[TSM+10] Tang (J.), Scellato (S.), Musolesi (M.), Mascolo (C.) et Latora (V.), « Small-world behavior in time-varying graphs », *Phys Rev E*, vol. 81, 2010, p. 055101.

[TTW97] Taqqu (M.), Teverosky (V.) et Willinger (W.), « Is network traffic self-similar or multifractal ? », *Fractals*, vol. 5, n° 1, 1997, p. 63–73.

[TWS97] Taqqu (M. S.), Willinger (W.) et Sherman (R.), « Proof of a fundamental result in self-similar traffic modeling », *SIGCOMM Comput. Commun. Rev.*, vol. 27, n° 2, 1997, p. 5–23.

[TZ04] Thorup (M.) et Zhang (Y.), « Tabulation based 4-universal hashing with applications to second moment estimation », dans *Proc. ACM-SIAM SODA*, janvier 2004.

[VA99] Veitch (D.) et Abry (P.), « A wavelet based joint estimator of the parameters of long-range dependence », *IEEE Trans. on Info. Theory special issue on "Multiscale Statistical Signal Analysis and its Applications"*, vol. 45, n° 3, avril 1999, p. 878–897.

[VA01] Veitch (D.) et Abry (P.), « A statistical test for the time constancy of scaling exponents », *IEEE Transactions on Signal Processing*, vol. 49, n° 10, octobre 2001, p. 2325–2334.

[Vap95] Vapnik (V. N.), *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[Vat98] Vaton (S.), « A new test of stationarity and its application to teletraffic data », dans *Proceedings of IEEE ICASSP-98*, vol. 6, p. 3449–3452, Seattle (USA), 1998.

[VEM08] Vazquez (F.), Eguíluz (V. M.) et Miguel (M. S.), « Generic absorbing transition in coevolution dynamics », *Phys. Rev. Lett.*, vol. 100, 2008, p. 108702.

[Ver06] Vert (R.), *Theoretical insights on density level set estimation, application to anomaly detection*. PhD thesis, Paris 11 - Paris Sud, 2006.

[Vil48] Ville (J.), « Théorie et applications de la notion de signal analytique », *Câbles et Transmissions*, vol. 2A, 1948.

[VL07] Von Luxburg (U.), « A tutorial on spectral clustering », *Statistics and computing*, vol. 17, n° 4, 2007, p. 395–416.

[Vél] Vélo'V. « http ://www.velov.grandlyon.com/ ».

[WAJ07] Wendt (H.), Abry (P.) et Jaffard (S.), « Bootstrap for empirical multifractal analysis », *IEEE Signal Proc. Mag.*, vol. 24, n° 4, 2007, p. 38–48.

[WC94a] Wood (A.) et Chan (G.), « Simulation of stationary Gaussian processes in $[0, 1]^d$ », *J. of Comput. and Graph. Stat.*, vol. 3, n° 4, 1994, p. 409–432.

[WC94b] Wood (A.) et Chan (G.), « Simulation of stationary process in $[01]^d$ », *J. of Computational and Graphical Stat.*, vol. 3, 1994, p. 409–432.

[WEFW12] Wrinch (M.), EL-Fouly (T. H.) et Wong (S.), « Anomaly detection of building systems using energy demand frequency domain anlaysis », dans *IEEE Power & Energy Society General Meeting*, San-Diego, CA, USA, 2012.

[WgTSW97] Willinger (W.), Taqqu (M. S.), Sherman (R.) et Wilson (D. V.), « Self-similarity through high-variability : statistical analysis of ethernet lan traffic at the source level », *IEEE/ACM Trans. Netw.*, vol. 5, n° 1, 1997, p. 71–86.

[WH09] Wu (Z.) et Huang (N. E.), « Ensemble empirical mode decomposition : A noise-assisted data analysis method », *Advances in Adaptive Data Analysis*, vol. 1, n° 01, 2009, p. 1–41.

[WS98] Watts (D. J.) et Strogatz (D. H.), « Collective dynamics of "small-world" networks », *Nature*, vol. 393, 1998, p. 440–442.

[XF07] Xiao (J.) et Flandrin (P.), « Multitaper time-frequency reassignment for nonstatinary spectrum estimation and chirp enhancement », *IEEE Trans. on Signal Proc.*, vol. 55, n° 6 (Part 2), 2007, p. 2851–2860.

[XKH11] Xu (K. S.), Kliger (M.) et Hero (A. O.), « Tracking communities in dynamic social networks », dans *Conf on Social Computing, Behavioral-Cultural Modeling, and Prediction*, March 2011.

[XKH13] Xu (K. S.), Kliger (M.) et Hero (A. O.), « A regularized graph layout framework for dynamic network visualization », *J. Data Mining and Knowledge Discovery*, vol. 27, n° 1, July 2013.

[XKH14] Xu (K. S.), Kliger (M.) et Hero (A. O.), « Adaptive evolutionary clustering », *J. Data Mining and Knowledge Discovery*, vol. 28, n° 2, march 2014.

[XZB05] Xu (K.), Zhang (Z. L.) et Bhattacharyya (S.), « Profiling Internet Backbone Traffic : Behavior Models and Applications », *ACM SIGCOMM'05*, 2005, p. 169–180.

[Yag87] Yaglom (A.), *Correlation Theory of Stationary and Related Random Functions*. Springer-Verlag, 1987.

[YR90] Yokuda (S.) et Ramaprian (B. R.), « The dynamics of flow around a cylinder at subcritical reynolds numbers », *Phys. Fluids A*, vol. 2, n° 5, may 1990, p. 784–791.

[Zdr02] Zdravkovich (M. M.), *Flow Around Circular Cylinders*. Oxford University Press, 2002.

[ZES04] Zimmermann (M.), Eguíluz (V.) et San Miguel (M.), « Coevolution of dynamical states and interactions in dynamic networks », *Phys. Rev. E*, vol. 69, 2004, p. 065102.

[ZI04] Zoubir (A.) et Iskander (D.), *Bootstrap techniques for signal processing*. Cambridge University Press, 2004.

[ZSBB11] Zhao (K.), Stehlé (J.), Bianconi (G.) et Barrat (A.), « Social network dynamics of face-to-face interactions », *Phys. Rev. E*, vol. 83, 2011, p. 056109.

Table des matières