Motifs séquentiels pour la description de séries temporelles d'images satellitaires et la prévision d'événements

Nicolas Méger

Université de Savoie

Habilitation à diriger des recherches
Spécialité : sciences et technologies de l'information

Motifs séquentiels pour la description de séries temporelles d'images satellitaires et la prévision d'événements

Nicolas MÉGER

Soutenue à Annecy-le-Vieux le 29 mars 2013

Jury

Atilla BASKURT	Rapporteur	Professeur à l'INSA de Lyon
Marie-Odile CORDIER	Rapportrice	Professeur à l'Université Rennes 1
Bruno CRÉMILLEUX	Examinateur	Professeur à l'Université de Caen
Sylvie GALICHET	Examinatrice	Professeur à Polytech Annecy-Chambéry
Pascal PONCELET	Rapporteur	Professeur à l'IUT de Béziers
Christophe RIGOTTI	Examinateur	Maître de Conférences HDR à l'INSA de Lyon

LISTIC (EA 3703), Polytech Annecy-Chambéry
B.P. 80439, F-74944 Annecy-le-Vieux Cedex

Table des matières

Table des figures

Première partie

Synthèse

Chapitre 1

Curriculum vitae

1.1 État civil

Nom :	MÉGER
Prénoms :	Nicolas, Pierre, Christian
Date de naissance :	20/02/1977
Lieu de naissance :	Bagnols/Cèze (Gard)
Nationalité :	Française
Situation familiale :	Marié, 2 enfants
Adresse :	315 route de la Fougère, F-73100 Grésy/Aix
Téléphone :	+33 (0) 610 748 466

1.2 Statut

Fonction :	Maître de conférences, 27$^{\text{ème}}$ section Université de Savoie
Courrier électronique :	nicolas.meger@univ-savoie.fr
Coordonnées enseignement :	IUT Annecy - Département INFO 9 rue de l'Arc-en-Ciel, F-74940 Annecy Tél. : +33 (0) 450 092 353
Coordonnées recherche :	LISTIC - Polytech Annecy-Chambéry B.P. 80439, F-74944 Annecy-le-Vieux Cedex Tél. : +33 (0) 450 096 520
Date d'installation :	28/09/2005
Date de titularisation :	18/09/2006

1.3 Expérience professionnelle

Depuis septembre 2005 Maître de conférences au département INFO de l'IUT d'Annecy, Université de Savoie. Laboratoire d'accueil : Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (LISTIC, EA 3703).

Oct. 2004/août 2005 ATER mi-temps. Enseignement à l'Institut National des Sciences Appliquées (INSA) de Lyon, département Informatique. Laboratoire d'accueil : Laboratoire d'InfoRmatique en Images et Systèmes d'information (LIRIS, UMR 5205).

Oct. 2002/août 2005 Vacataire en EURINSA, premier cycle européen de l'INSA Lyon.

Oct. 2001/ sept. 2004 Chercheur au sein du projet européen AEGIS (IST-2000-26450), employé par l'INSA Lyon. Laboratoire d'accueil : Laboratoire d'InfoRmatique en Images et Systèmes d'information (LIRIS, UMR 5205).

1999/2000/2001 Stagiaire à la COGEMA (analyste-programmeur, 2 mois), à la SNCF (analyste-programmeur pour la cellule de veille technologique, 6 mois) et chez Cesame [3] (création d'un méthode d'analyse du système d'information par l'analyse de la valeur, 6 mois).

1.4 Formation

16 décembre 2004 **Grade de Docteur**, INSA de Lyon, spécialité Extraction de Connaissances à partir des Données. **Directeurs de thèse** : Pr. Jean-François Boulicaut et Christophe Rigotti (maître de conférences HdR). **Rapporteurs** : Pr. Dominique Laurent et Pr. Pascal Poncelet. **Membres du jury** : Pr. Marie-Odile Cordier, Pr. Dominique Laurent, Pr. Pascal Poncelet, Pr. Bruno Crémilleux, Pr. Jean-François Boulicaut et Christophe Rigotti (maître de conférences HdR). **Titre de la thèse** : Recherche automatique des fenêtres temporelles optimales des motifs séquentiels.

29 octobre 2002 **D.E.A. en Extraction de Connaissances à partir des Données**, INSA de Lyon, Mention Très Bien. **Directeur du stage** : Christophe Rigotti (maître de conférences HdR). **Sujet du stage** : Résumés de collections de règles d'association, application aux tâches de monitoring.

21 novembre 2001 **Diplôme d'Ingénieur, Spécialité Informatique**, INSA de Lyon. **Responsable du projet de fin d'études** : Régis Aubry (maître de conférences). **Sujet de projet de fin d'études** : Conception des systèmes d'information et de communication par la méthode d'analyse de la valeur.

25 septembre 1995 **Diplôme du Baccalauréat Général**, série Scientifique, Mention Très Bien, délivré à Aix-En-Provence.

1.5 Langues

Anglais :	Lu, parlé, écrit.
Allemand :	Lu, parlé, écrit. Stage ouvrier chez LURGI A.G. Frankfurt am Main, Allemagne.
Espagnol :	Lu, parlé, écrit. Dernière année de formation ingénieur de l'INSA de Lyon à l'université Jaume I, Erasmus, Castellón de la Plana, Espagne.

Chapitre 2

Enseignement

2.1 Panorama

J'ai commencé à enseigner en tant que vacataire au sein du premier cycle européen de l'INSA Lyon, EURINSA, entre 2002 et 2005. Mes premiers cours/TD/TP, 90 heures éq. TD, traitaient d'algorithmie, de programmation et des systèmes d'exploitation. J'ai également enseigné en tant qu'ATER mi-temps au second cycle de l'INSA Lyon, pour le département Informatique, en 2004/2005. J'ai ainsi accompagné les élèves ingénieurs sur 72 heures éq. TD de TP de système embarqué et de système temps réel/multitâches. Enfin, en 2003, en tant que chercheur du projet européen AEGIS (IST-2000-26450), j'ai formé en 40 heures des doctorants et post-doctorants à l'extraction de connaissances dans les données, à la fouille de données, au C/C++, aux bases de données et aux accès web aux bases de données.

Recruté à la création du département INFO de l'IUT d'Annecy comme maître de conférences en 27$^{\text{ème}}$ section à l'Université de Savoie en septembre 2005, j'y effectue depuis lors l'essentiel de mes enseignements. La création d'un département est un contexte singulier et motivant qui a nécessité une forte implication de ma part, à la fois au niveau de l'enseignement et des charges administratives. La figure 2.1 donne, par année scolaire, un aperçu des volumes horaires effectués à l'IUT d'Annecy. Après l'ouverture du département en 2005, la montée en régime de 2006 et le pic d'activité de 2008 et 2009, j'ai engagé en 2010 un effort de baisse du volume d'enseignement afin de pouvoir développer mes activités de recherche. Il est à noter que ces volumes incluent les charges administratives **à l'exception** de la direction de le licence professionnelle *Chargé de projet informatique* (670 heures éq. TD entre 2007 et 2011) et des enseignements dispensés dans cette même licence et dans la licence professionnelle *Bases de données* (110 heures éq. TD entre 2007 et 2011). Pour la période 2007-2011, ceci représente un volume supplémentaire de 156 heures éq. TD en moyenne par année qu'il faut ajouter aux volumes ici rapportés. Les années 2008 et 2009, avec environ 500 heures éq. TD chacune, ont ainsi été les plus chargées.

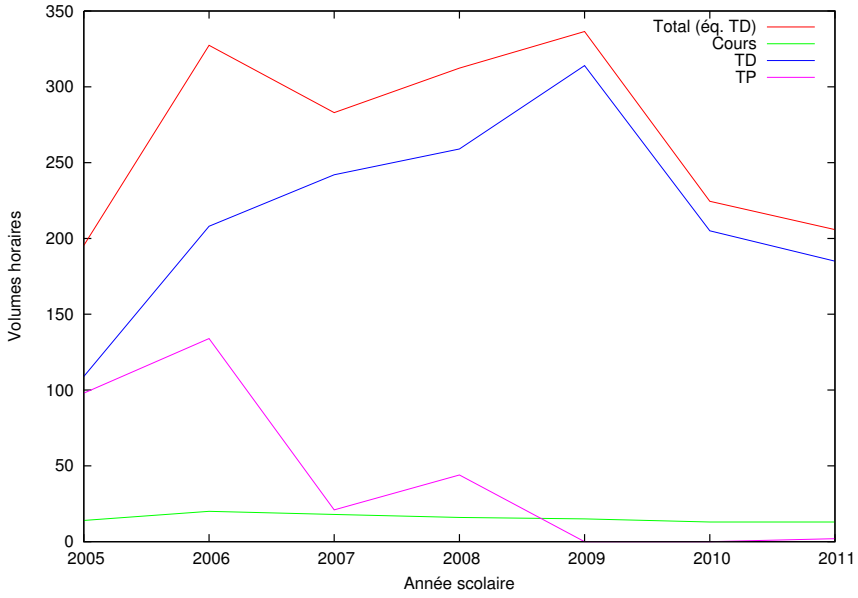


FIGURE 2.1 – Volumes horaires à l'IUT d'Annecy (hors licences professionnelles).

2.2 Contenus des enseignements et charges d'enseignement

Le/la lecteur/lectrice trouvera ci-après les détails relatifs aux contenus des enseignements et aux charges d'enseignement. Les années rapportées sont des années scolaires.

Analyse et conception des systèmes d'information (934 heures éq. TD de 2005 à 2011, 109 h. cours, 719 h. TD, 72 h. TP)

La responsabilité de ces cours de première année et de deuxième année m'a été confiée à mon arrivée. Entre 2007 et 2010, j'ai également enseigné ce type de cours en licence professionnelle *Chargé de projet informatique* afin de procéder à la conception d'applications destinées à des associations 1901. Quel que soit le niveau, j'ai assuré la création des différents supports pour les cours, les TD et les TP associés. J'ai donné la totalité des cours tout en intervenant en TD et TP. L'objectif de ces cours est de faire découvrir et utiliser les outils de modélisation des systèmes d'information pour analyser et concevoir un système d'information, le plus souvent réduit au système informatique. Au niveau des méthodes et des modèles, le choix a été fait d'enseigner MERISE et de proposer également une approche objet avec UML 2.0 comme formalisme. Un accent tout particulier a été mis sur l'utilisation de cas d'étude réels en TP comme le *Festival du cinéma d'animation d'Annecy*. L'atelier de génie logiciel utilisé lors des TP est PowerAMC. Les cours dispensés en première année sont obligatoires. Des TP de découverte du SQL pour la création, la mise à jour et l'interrogation des bases de données illustrent l'intérêt d'une modélisation correcte des données. La conception d'interfaces basées sur l'utilisation de requêtes SQL

est également abordée sous Access. Les cours proposés en deuxième année font l'objet d'un module complémentaire mettant en œuvre les concepts vus en première année et en deuxième année sur un projet couvrant l'ensemble des étapes d'un développment logiciel, de la définition du cahier des charges à la recette de l'application. La réalisation de cette dernière s'effectue sur la base d'une architecture MVC/MVC2, en utilisant des technologies récentes (PHP5.3, Oracle 11g) et en prêtant attention à la documentation associée (cahier des charges, dossier de spécification, manuel utilisateur, manuel technique, cahiers de tests et de recette). L'utilisation de frameworks de type Symfony, Zend ou Yii est encouragée.

Architecture et programmation assembleur (198 heures éq. TD de 2005 à 2009, 112 h. TD, 129 h. TP)

Pour ces cours, je suis venu en soutien de mes collègues au niveau des TD et TP. Au travers de la découverte de l'architecture des processeurs et de l'assembleur, l'objectif est ici de démystifier le fonctionnement des ordinateurs et de montrer en quoi les sécurités d'un programme peuvent être facilement contournées et en quoi peut consister l'optimisation d'un programme. Après avoir fonctionné avec un simulateur de processeur de type M68000, nous avons utilisé des processeurs Intel en programmant en assembleur x86. Afin de rendre le cours attrayant, un accent particulier a été mis sur la programmation de jeux dits *vintage* de type *Pong* ou *Asteroids*.

Algorithmie et programmation (150 heures éq. TD de 2005 à 2008, 77 h. TD, 94 h. TP)

En ce qui concerne ces cours, je suis également venu en soutien de mes collègues au niveau des TD et TP. Ces cours traitent d'algorithmie et de programmation procédurale en C (première année), de programmation par objets en C++ et Java (première et deuxième année).

Fouille de données (63 heures éq. TD de 2009 à 2011)

Que ce soit au niveau DUT ou au niveau licence professionnelle, j'ai enseigné la fouille de données (ou data mining), et l'extraction de connaissances dans les données. Les modèles globaux (cluterings, classifieurs) ainsi que les motifs locaux (règles d'association, motifs séquentiels, épisodes) sont présentés. Une mise en pratique grâce à l'atelier Knime vient en support de l'assimilation de ces concepts. Ce cours, d'une quinzaine d'heures en moyenne, constitue une partie d'un module complémentaire de deuxième année de DUT et est obligatoire en licence professionnelle *Bases de données*. Pour ce qui est des aspects théoriques, je m'appuie sur le livre *Introduction to Data Mining* de Tan, Steinbach et Kumar [TSK05]. J'utilise également une version modifiée de leurs transparents. Pour ce qui est des TP sous Knime, je me suis inspiré des TP construits par Christophe Rigotti pour l'INSA Lyon. J'ai également dispensé ce type de cours en Espagne, à l'université Jaume I, au niveau master, en deux sessions de 10 heures chacune. Ces sessions ne sont pas comptées dans les volumes horaires ici rapportés.

Projets tutorés et stages (290 heures éq. TD de 2005 à 2011)

Nos étudiants découvrent le monde de l'entreprise au travers d'un projet tutoré et d'un stage en entreprise. Les projets tutorés sont des missions facturées aux entreprises et réalisées par groupe de 5/6 étudiants dans les locaux de l'IUT. En dehors d'une mise en œuvre technique des cours dispensés, ces projets permettent de comprendre quelles sont les attentes et les exigences d'un client, d'une entreprise. Les projets démarrent en fin de première année et finissent en milieu de deuxième année. S'en suivent alors les stages en entreprise permettant de quitter le cocon de l'IUT et d'apprendre à s'intégrer au sein même des entreprises. Tout ceci représente un apprentissage, tant en savoir-être qu'en compétences techniques. Afin d'aider les étudiants dans cette démarche, nous les accompagnons avec des heures dédiées aux projets tutorés et aux stages, heures auxquelles j'ai participé. Ces heures incluent les visites de stages, visites particulièrement intéressantes en ce qui concerne la relation à l'entreprise. Il est ainsi possible de présenter et d'expliquer nos différentes formations (DUT, licences professionnelles) tout en recensant les besoins des entreprises.

Actions de promotions, Recrutement, Projet Personnel et Professionnel (PPP) (136 heures éq. TD de 2006 à 2011)

Afin d'atteindre le meilleur taux de réussite possible, nous nous assurons que les élèves recrutés aient choisi une formation en phase avec leurs aspirations. À cet effet, mes collègues et moi-même assurons des actions de promotion en étant présents dans les salons étudiants et procédons à des entretiens de recrutement. De même, que ce soit en groupe ou individuellement, j'ai accompagné les élèves dans la définition et la mise en œuvre de leur Projet Personnel et Professionnel (PPP), en première et en deuxième année.

Projets de communication (10 heures éq. TD en 2005)

J'ai encadré deux projets de communication. Le premier était un projet audiovisuel dont le but était de produire des films présentant le département INFO. L'autre projet concernait quant à lui l'animation du département (organisation de sorties raquettes, restaurant, LAN parties, laser game, etc.).

Bureautique et modélisation des systèmes d'information (56 h. TD, de 2005 à 2006)

Lors de ces cours en département GEA (Gestion des Entreprises et des Administrations), j'ai enseigné l'utilisation de Word, PowerPoint et Publisher. J'ai également enseigné la modélisation des données d'un système d'information (MERISE, MCD/MLD/MPD). Mis à part le fait de compléter mon service en 2005, ce fut l'occasion de rencontrer un public d'élèves très différent et de mettre en perspective les choix pédagogiques du département GEA avec ceux faits lors de la mise en place du département INFO.

Direction de la licence professionnelle *Chargé de projet informatique* (CPINFO) (670 heures éq. TD de 2007 à 2011)

Après avoir porté et défendu le projet de cette licence pendant un an et demi, j'en ai pris la direction dès son ouverture en 2007. Cette licence, en alternance, s'effectue

en partenariat avec Tétras (association 1901 de l'Université de Savoie et de la Chambre Syndicale de la Métallurgie) et le lycée Saint Michel d'Annecy. Tétras facilite la gestion des contrats de professionnalisation tandis que le lycée Saint Michel nous apporte un retour d'expérience d'une quinzaine d'année sur ce type de formation. La responsabilité de cette formation est intéressante car elle requiert la gestion d'une maquette pédagogique, d'une équipe d'enseignants et de professionnels, des notes, et du planning. Le suivi des étudiants, tant au niveau de la formation que de l'entreprise fait également partie intégrante de la fonction. Cela permet de côtoyer les entreprises et comme pour les stages, de présenter et d'expliquer nos différentes formations tout en recensant les besoins des entreprises.

Relations internationales, Poursuites d'études, Anciens et Emplois, 95 heures éq. TD de 2005 à 2010)

Afin d'assurer le bon fonctionnement du département, dès sa création, j'ai pris en charge les relations internationales. Ce fut un travail, qui malgré le nombre d'heures affichées (45 heures éq. TD), a été particulièrement chronophage : mise en place d'accord Socrates, Erasmus ou bi-latéraux, sélection des candidats au départ, accompagnement des candidats tant au niveau des procédures universitaires, régionales, européennes que des procédures des universités d'accueil. En 2008, à l'occasion du renouvellement du chef de département, j'ai changé de charge administrative en prenant les poursuites d'études (organisation de forums d'information, préparation du jury de poursuites d'études, conseils aux étudiants, participation à des jurys de recrutement), les anciens et les emplois. Enfin, en 2009, à l'occasion du recrutement d'un collègue, je me suis concentré sur les seules poursuites d'études.

Chapitre 3

Recherche

3.1 Panorama

Recruté en 2005 à l'IUT d'Annecy pour l'ouverture du département INFO, j'ai été accueilli au sein du laboratoire LISTIC. J'ai donc effectué à la fois une mobilité géographique et une mobilité thématique. En effet, j'ai fait évoluer mon travail de recherche démarré en D.E.A. à l'INSA de Lyon, au sein du laboratoire LIRIS (UMR 5205). Ce travail avait trait à l'Extraction de Connaissances à partir des Données (ECD ou KDD, Knowledge Discovery in Databases) et portait sur les motifs locaux, des règles d'associations aux épisodes en passant par les motifs séquentiels. Les publications concernées ([5], [22], [23], [24], [25] et [27]) ont toutes été publiées avant 2005 et ne seront pas mentionnées ni comptabilisées dans cette partie. Les références utilisées dans cette partie sont définies au chapitre 7 de cette première partie.

Mon intégration au sein du LISTIC en tant qu'unique « data miner » m'a permis d'avoir accès à une large palette de données (données satellitaires, sismiques, mécaniques, médicales), d'applications (observation de la Terre, pilotage de systèmes complexes) et de compétences théoriques (théorie des ensembles flous, théorie des possibilités, traitement du signal, traitement d'images, télédétection, fusion de données/d'informations) à partir desquelles j'ai orienté ma recherche. C'est ainsi que j'ai dédié mes travaux à la description spatio-temporelle non supervisée de séries temporelles d'images satellitaires et à la prévision d'événements dans un flot de données pour l'aide au pilotage de systèmes complexes. Plus généralement, le premier aspect de mes travaux concerne la question de la représentation des connaissances extraites et de son adéquation par rapport aux domaines d'applications envisagés. Un second aspect est alors la création d'algorithmes justes et complets permettant l'extraction des représentations définies. Ces algorithmes sont ensuite implantés dans des prototypes qui sont validés par des expériences sur des jeux de données réels et/ou synthétiques. Quel que soit l'axe de recherche considéré, et afin de garantir les meilleurs résultats possibles, cette démarche est instanciée deux fois, sur deux étapes différentes. La première étape correspond à une étude de *faisabilité et d'opportunité*. À ce niveau, le besoin utilisateur est recensé, détaillé, et les outils théoriques et techniques disponibles sont réutilisés le plus directement possible afin de valider ou non la

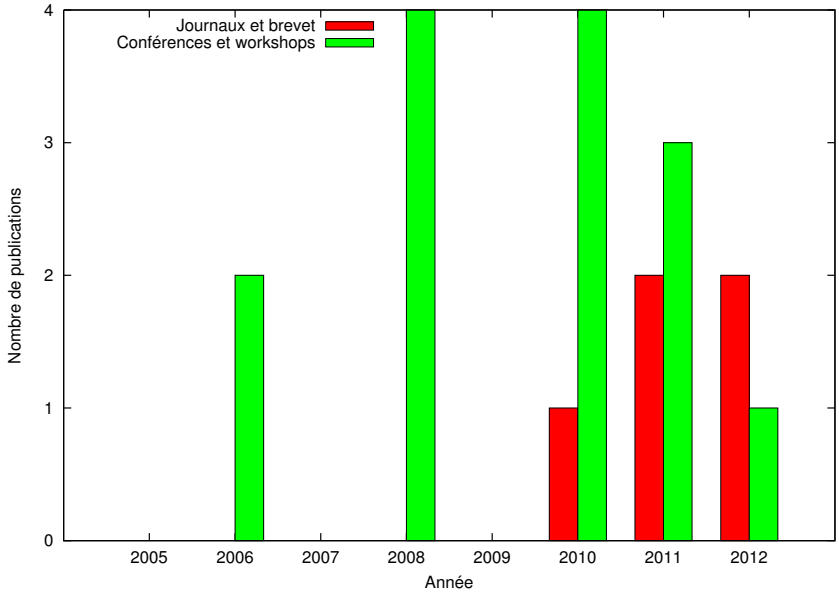


FIGURE 3.1 – Nombre de publications par année civile, de 2005 à 2012.

présence d'une opportunité quant à de futurs développements. Cette étape est en partie confiée, sous mon encadrement, à des stagiaires de master recherche ou à des stagiaires en dernière année d'école d'ingénieur. Si cette étape se révèle être concluante, alors une deuxième étape, dite de *proposition*, est engagée et vise à la création de solutions originales permettant d'améliorer les résultats obtenus lors de la première étape. Pour cette dernière étape, appel est fait à des doctorants ou post-doctorants pour engager le travail de fond nécessaire. Mis à part le fait d'adopter une démarche construite et ordonnée, ce mode de gestion m'a permis de gérer de front la création du département INFO et mes premiers pas au sein de LISTIC, et ce au travers des études de faisabilité. Comme le montre la figure 3.1, la traduction de cette démarche au niveau des publications est directe : 2006 et 2008 sont les années où ont été publiés les articles issus des études de faisabilité tandis que les apports majeurs provenant des deuxièmes étapes ont été publiés entre 2010 et 2012. Toutes les publications ici rapportées ont été validées par des comités de lecture.

3.2 Description spatio-temporelle non supervisée de Séries Temporelles d'Images Satellitaires (STIS)

Un premier volet de mes activités concerne donc la description spatio-temporelle de séries d'images satellitaires à l'aide de méthodes de fouille de données (ou data mining), en mode non supervisé. Afin d'évaluer la généricité des développements ici présentés, ceux-ci ont été testés à la fois sur des données optiques (METEOSAT, SPOT ou LANDSAT) et sur des données radar (ENVISAT ou ERS). Ces travaux furent pour moi l'occasion d'encadrer Andreea Maria Julea en stage de master (2006), puis de co-encadrer sa thèse (2007-2011) avec Philippe Bolon. Suivant la démarche présentée dans la par-

tie 3.1, nous avons dans un premier temps appliqué des méthodes existantes de fouille de données, l'extraction de motifs séquentiels fréquents, afin d'évaluer leurs potentiels et leurs limites. Ceci a fait l'objet de collaborations avec Telecom ParisTech (ENST) et de publications dans une conférence liée au data mining [21] (workshop ECML/PKDD) et dans des conférences liées à l'analyse d'images satellitaires (European Space Agency [20], [19], [18] et IGARSS [17]). À ce niveau, les motifs séquentiels fréquents ont été retenus. Chaque motif décrit une évolution/sous-évolution de pixels dans le temps. Une évolution est proposée à l'utilisateur si celle-ci concerne suffisamment de pixels. Cette contrainte de surface minimum peut être directement traduite à l'aide du support minimum d'un motif séquentiel. Le caractère anti-monotone de cette dernière assure des extractions dans des conditions standards du point du vue temps et espace (mémoire). Puis, en collaboration avec le LIRIS, nous nous sommes orientés vers la conception d'algorithmes originaux de fouille de données permettant d'extraire des motifs séquentiels fréquents qui, en plus de la notion de surface, intègrent une contrainte de connexité spatiale, que ce soit de façon active ou non. Une contrainte active permet, ici associée à la contrainte de fréquence, d'élaguer des portions entières de l'espace des solutions. D'un point de vue applicatif, cela permet également de ne proposer que des motifs qui ont du sens à la fois temporellement et spatialement. Ces derniers développements ont fait l'objet de 2 publications dans des revues de chacun des domaines, MLDM [2] et TGRS [3], dans la revue nationale ISI [4] et dans des conférences internationales touchant également au data mining et à l'analyse d'images satellitaires (IGARSS [10], [12], Living Planet - European Space Agency [14], ou bien encore ICDM [9]). Ces travaux, ainsi que la diffusion du logiciel SPATPAM [33] s'inscrivent en tant que résultats du projet ANR EFIDIR auquel je participe (responsable de workpackage). Ils ont servi d'appui à l'obtention du projet ANR FOSTER dont je suis actuellement responsable scientifique pour le LISTIC et pour le compte duquel j'encadre Felicity Lodge, post-doctorante, depuis janvier 2012. Dans ce contexte, et toujours en collaboration avec le LIRIS, nous avons travaillé à l'établissement d'une méthode permettant de sélectionner les motifs les plus singuliers, c'est-à-dire les motifs qui n'apparaissent que très peu dans des jeux aléatoires où les fréquences des symboles sont préservées. Pour ce faire, nous nous sommes appuyés sur les localisations spatio-temporelles des motifs, sur la mesure d'information mutuelle et sur les techniques de swap-randomization. Ce travail a fait l'objet d'un workshop dédié à l'analyse de séries temporelles d'images satellitaires [11] (Multitemp 2011) et d'une conférence internationale [8] (European Space Agency). Nous continuons aujourd'hui à travailler sur les aspects théoriques de cette approche (atteignabilité, uniformité, convergence).

En perspective, le calcul/la caractérisation de champs de déplacements à partir de séries d'images nous semble prometteur : de nombreux experts utilisent ces représentations pour observer et simuler des phénomènes en particulier géophysiques et géomécaniques. Dans ce cadre, de telles techniques peuvent être envisagées comme support à l'inversion de modèles. Aidés de collègues géophysiciens, géomécaniciens, traiteurs de signal et fouilleurs de données, nous rédigeons un projet ANR en ce sens que je serai amené à diriger s'il aboutit.

3.3 Prévision d'événements dans un flot de données : application à l'aide au pilotage de systèmes complexes

Un deuxième volet de mes activités concerne l'utilisation automatique de motifs locaux, en particulier les épisodes (*FLM-règles*), dans le but de prévoir, en nature et dans le temps, des événements pouvant apparaître dans un flot de données tels que les défaillances d'un système complexe. L'objectif sous-jacent est ici d'aider au pilotage d'un tel système. À nouveau, suivant la démarche présentée dans la partie 3.1, j'ai initié ces travaux en co-encadrant en 2008 (avec Lionel Valet et Julien Boissière) Nicolas Le Normand en stage d'ingénieur CNAM, au sein du projet BQR « Chaîne logistique ». Comme son nom l'indique, la problématique alors traitée concerne le pilotage d'un chaîne logistique. Quelles sont les quantités de produits à déplacer et où déplacer ces produits tout en assurant un coût minimal et une satisfaction des clients maximale, telles sont, en partie, les questions posées par cette notion de pilotage. Dans ce contexte, nous avons par exemple cherché à prévoir que des clients seraient insatisfaits (dépassement du délai de livraison annoncé). Ce type de prévision permet de tirer le signal d'alarme et de mettre en œuvre, à temps, les actions correctrices adéquates. Des résultats très encourageants ont été obtenus sur des jeux de données synthétiques générés avec le logiciel ARENA. Cela a fait l'objet d'une publication (workshop ECML/PKDD [16]). La méthode retenue ne prenait cependant pas en compte toutes les propriétés temporelles des FLM-règles. De plus, elle ne permettait pas de sélectionner les règles les plus génériques produisant le moins possible de fausses alarmes. Ce type d'approche a donc été raffiné et validé sur des jeux de données réels grâce à Florent Martin, en thèse CIFRE chez ADIXEN (Alcatel-Lucent, 2007-2011, avec contrat d'accompagnement) et que j'ai co-encadrée avec Sylvie Galichet. Plus précisément, dans ce contexte, nous avons cherché à prédire les grippages, c'est-à-dire les blocages, de pompes à vides. Prévoir un grippage permet aux opérateurs d'effectuer une maintenance avant même qu'une pompe ne tombe en panne, ce qui génère des économies substentielles. Ces pompes à vide sont utilisées par des clients du marché des semi-conducteurs et sont soumises, chez un même client, à des conditions d'utilisation extrêmement sévères et variables (température, gaz corrosifs, dépôts de particules). L'expertise des mécaniciens n'étant pas suffisante pour prévoir un grippage dans ces conditions, une méthode non supervisée était à construire. Les résultats obtenus ont validé l'approche retenue et ont même permis de découvrir de nouvelles connaissances liées à la cinématique des pompes. Un brevet mondial est aujourd'hui accepté [6]. Ce travail a également fait l'objet de communications dans des conférences internationales (ICDM [13] et IAE/AIE [15]), dans une conférence nationale (workshop EGC [26]) et dans la revue MLDM [1].

En perspective, l'introduction de la gradualité dans les alarmes, la définition de motifs dédiés à la prévision et la prise en compte de la nature dynamique de certains flots de données nous semblent constituer des pistes prometteuses.

Chapitre 4

Encadrement

Outre l'encadrement de stagiaires de niveau DUT pour le compte du laboratoire, j'ai également encadré et co-encadré des stages master, ingénieur ainsi que des thèses. Le détail est produit ci-après. Les références qui sont utilisées dans ce chapitre sont définies au chapitre 7.

4.1 Stages master et ingénieur (niveau M2)

Andreea Maria Julea

Période	mars 2006 - juin 2006.
Formation	Master de l'Université Politehnica din Bucaresti Faculté d'Électronique, Télécommunications et Technologies de l'Information.
Spécialisation	Images, Formes et Intelligence Artificielle.
Titre	Frequent sequential patterns in satellite imagery.
Encadrement	100 %.
Publications	1 article en conférence [20] et 1 article en workshop [21].
Situation actuelle	chercheuse à l'Institut des Sciences Spatiales de Roumanie.

Nicolas Le Normand

Période	septembre 2007 - juillet 2008.
Formation	Conservatoire National des Arts et Métiers (école d'ingénieur).
Spécialisation	Management des Systèmes d'Information.
Titre	Optimiser une chaîne logistique par traitement de l'information.
Encadrement	33%, en collaboration avec Julien Boissière et Lionel Valet.
Publications	1 article en workshop [16].
Situation actuelle	CDI chez AUSY.

4.2 Thèses

Andreea Maria Julea

Période	mars 2007 - septembre 2011.
Contexte	co-tutelle entre l'Université de Savoie et l'Université Politehnica din Bucaresti.
Titre	Extraction de motifs spatio-temporels dans des séries d'images de télédétection : application à des données optiques et radar.
Encadrement	50 % avec Philippe Bolon.
Soutenance	le 20 Septembre 2011 devant Teodor Petrescu (président), Jean-François Boulicaut (rapporteur), Alexandru Badea (rapporteur), Yannick Berthoumieu (examinateur), Mihai Datcu (examinateur), Philippe Bolon (directeur de thèse), Nicolas Méger (co-directeur), Vasile Lazarescu (directeur de thèse).
Publications	2 articles en revue, [2] et [3]. 8 articles en conférences/workshops : [9],[10], [11],[12],[14],[17],[18], et [19].
Situation actuelle	chercheuse à l'Institut des Sciences Spatiales de Roumanie.

Florent Martin

Période	octobre 2007 - juin 2011.
Contexte	CIFRE entre ADIXEN (Alcatel-Lucent) et l'Université de Savoie.
Titre	Pronostic de défaillances de pompes à vide - Exploitation automatique de règles extraites par fouille de données.
Encadrement	50 % avec Sylvie Galichet.
Soutenance	le 29 Juin 2011 devant Laurent Foulloy (président du jury), Marie-Odile Cordier (rapporteur), Christophe Rigotti (rapporteur), Nicolas Bécourt (examinateur), Sylvie Galichet (directeur de thèse), Nicolas Méger (co-directeur de thèse).
Publications	1 article en revue [1]. 1 brevet [6]. 2 articles en conférences, [13] et [15].
Situation actuelle	CDI chez Alpha 3i.

Chapitre 5

Projets et contrats

Outre trois projets BQR de l'Université de Savoie auxquels j'ai participé, d'autres projets et un contrat ont également servi et servent de support financier au travail de recherche mis en œuvre. En voici le détail.

Projet ANR EFIDIR - Extraction et Fusion d'Informations pour la mesure des Déplacements en Imagerie Radar (2007-2012)

- Co-responsable (avec Cécile Lasserre, laboratoire ISTerre) du workpackage *Détection de perturbations atmosphériques*.
- Participation à la rédaction du projet.
- Responsable de la production de 3 livrables : prototype SPATPAM (version bêta [34] et version finale [33]), rapport sur l'élimination d'artefacts atmosphériques [32] (pour ce dernier, responsabilité partagée avec C. Lasserre).
- Responsable du cours *Data mining and its application to SITS analysis* de *l'école de printemps 2011 EFIDIR Spring School, Ecole de Physique des Houches, mai 2011.*
- Membre du comité d'organisation de *l'école de printemps 2011 EFIDIR Spring School, Ecole de Physique des Houches, mai 2011.*
- Axe de recherche concerné : description spatio-temporelle non supervisée de séries d'images satellitaires.
- Site web : http ://www.efidir.fr/

Projet ANR FOSTER - FOuille de données Spatio-Temporelles : application à la compréhension et à la surveillance de l'Erosion (2011-2013)

- Responsable scientifique pour le laboratoire LISTIC.
- Participation à la rédaction du projet.
- Responsable de la production de 3 livrables sur la fouille de données multi-étapes : état de l'art [31], rapport de recherche [28], expérimentation et validation (à venir).
- Encadrant d'une post-doctorante, Felicity Lodge, recrutée depuis Janvier 2012.
- Axe de recherche concerné : description spatio-temporelle non supervisée de séries d'images satellitaires.
- Site web : http ://foster.univ-nc.nc

Projet ANR pFlower - Reconnaissance de flot applicatif par processeur Multicœurs (2010-2013)

- Co-responsable (avec Kavé Salamation et Flavien Vernier, laboratoire LISTIC) du workpackage *Behavioral Modeling and Data Mining*.
- Participation à la rédaction du projet.
- Reponsable de la production de 2 livrables sur l'extration d'episodes : état de l'art [30] et algorithme parallèle pour l'extraction de FLM-règles [29].
- Axe de recherche concerné : parallélisation d'algorithmes d'extraction d'épisodes.
- Site web : http ://www.agence-nationale-recherche.fr

Contrat d'accompagnement de la thèse CIFRE de Florent Martin - ADIXEN (Alcatel-Lucent) (2007-2010)

- Participation à la rédaction du contrat.
- Axe de recherche concerné : prévision d'événements dans un flot de données.
- Site web : http ://www.adixen.fr

Chapitre 6

Activités d'intérêt collectif

Ce chapitre présente la liste des activités d'intérêt collectif que j'ai assurées depuis mon recrutement à l'Université de Savoie en septembre 2005.

6.1 Enseignement

Les responsabilités listées dans cette partie sont détaillées au chapitre 2 et sont toutes relatives au département INFO de l'IUT Annecy.

- **2005 - 2008** responsables des relations internationales.
- **Depuis 2007** responsable pédagogique de la licence professionnelle CPINFO.
- **2008 - 2009** responsable des poursuites des études, des anciens et de l'emploi.
- **2009 - 2011** responsable des poursuites des études.

6.2 Recherche

Membre du comité d'organisation

- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2009).
- Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2011).
- The 7$^{\text{th}}$ conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2011).
- 2011 EFIDIR Spring School, École de Physique des Houches, Chamonix.

Relecteur pour les revues internationales

- IEEE Transactions on Geoscience and Remote Sensing (TGRS).
- IEEE Geoscience and Remote Sensing Letters (GRSL).
- Knowledge and Information Systems (KAIS).
- Knowledge Discovery.
- Data Mining & Knowledge Discovery (DMKD).
- Data & Knowledge Engineering (DKE).

Relecteur pour les ateliers et conférences internationaux

- The 19$^{\text{th}}$ International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2006).
- The 2011 IEEE Workshop on *Evolving and Adaptive Intelligent Systems* organisé au sein du 2011 IEEE Symposium Series in Computational Intelligence.
- The 7$^{\text{th}}$ Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2011).

Membre du comité de programme

- Le workshop *Data Mining, Applications, Cas d'études et Success Stories* organisé au sein de la 11$^{\text{ème}}$ Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2011).
- Le workshop *FOuille de données Spatio-Temporelles et Applications* organisé au sein de la 13$^{\text{ème}}$ Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2013).

Membre du comité de rédaction

- Revue d'Intelligence Artificielle (RIA), numéro spécial « Intelligence Artificielle et Agronomie/Environnement ».

Diffusion de logiciels

- SPATPAM (une évolution de DMT4SP) [33] : extraction de motifs séquentiels fréquents groupés dans des séries d'images satellitaires. Christophe Rigotti, Nicolas Méger, Andreea Julea. Depuis Juillet 2009.
- WinMiner : extraction, dans une longue séquence d'événements, de règles d'épisodes et de leurs fenêtres temporelles optimales respectives (FLM-règles). Nicolas Méger, Christophe Rigotti. Depuis Janvier 2005.

Divers

- Membre élu du conseil de laboratoire du LISTIC (EA 3703) de 2006 à 2010.
- Membre de la commission *locaux* du laboratoire du LISTIC (EA 3703) depuis 2010.
- Expertise de dossiers de thèses CIFRE pour l'ANRT.
- Membre de 2 comités de sélection pour le recrutement de maîtres de conférences en 2012 (INSA Lyon/LIRIS et Université de Strasbourg/LSIIT).
- Président de la session *Data mining and its applications* à la 19$^{\text{th}}$ International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2006).
- Président de la session *Monitoring of the environment and natural hazards* au 2010 IEEE International Geoscience And Remote Sensing Symposium.

Chapitre 7

Publications, livrables de projets ANR et séminaires

Toutes les publications listées ci-après ont été visées par un comité de lecture.

7.1 Journaux internationaux

[1] Martin F., Méger N., Galichet S., Bécourt N., *Forecasting Failures in a Data Stream Context : Application to Vacuum Pumping System Prognosis.* In journal : Transactions on Machine Learning and Data Mining, volume 5, number 2, pp 87-116, October 2012, ISSN : 1865-6781.

[2] Julea A., Méger N., Rigotti C., Trouvé E., Jolivet R., Bolon P., *Efficient Spatiotemporal Mining of Satellite Image Time Series for Agricultural Monitoring.* In journal : Transactions on Machine Learning and Data Mining, volume 5, number 1, pp 23-44, July 2012, ISSN 1865-6781.

[3] Julea A., Méger N., Bolon P., Rigotti C., Doin M.-P., Lasserre C., Trouvé E., Lazarescu V., *Unsupervised Spatiotemporal Mining of Satellite Image Time Series using Grouped Frequent Sequential Patterns.* In journal : IEEE Transactions on Geoscience and Remote Sensing, volume 49, issue 4, pp. 1417 - 1430, April 2011, doi :10.1109/TGRS.2010.208-1372.

7.2 Journaux nationaux

[4] Trouvé E., Nicolas J-M., Ferro-Famil L., Gay M., Pinel, Doin M-P., Méger N., Lasserre C., Mauris G., Vernier F., Fallourd R., Yan Y., Harant O., Jolivet R., *EFIDIR : extraction et fusion d'informations pour la mesure de déplacements par imagerie radar.* In journal : Revue Traitement du Signal (TS), vol. 3-4, numéro 28, pp. 375-416, 2011, doi :10.3166/ts.28.375-416.

[5] Leleu M., Méger N., Rigotti C., *Extraction de Motifs Séquentiels Fréquents sous Contraintes dans des Données Contenant des Répétitions Consécutives*. In journal : Revue Ingénierie des Systèmes d'Information (ISI), vol. 9/3-4, pp. 133-159, 2004, doi :10.3166/isi.9-.3-4.133-159.

7.3 Brevet

[6] Nicolas Bécourt, Florent Martin, Cécile Pariset, Sylvie Galichet, Nicolas Méger, *Method for predicting a rotation fault in the rotor of a vacuum pump and associated pumping device*. Patent # WO 210/149 738 (29/12/2010), Alcatel-Lucent.

7.4 Chapitre d'ouvrage

[7] Bykowski A., Daurel T., Méger N., Rigotti C., *Integrity Constraints over Association Rules*. In Database Support for Data Mining Applications, Eds. Rosa Meo, Pier Luca Lanzi, Mika Klemettinen, Springer-Verlag, 2004, pp. 306-325.

7.5 Conférences internationales

[8] Méger N. Rigotti C., Gueguen L., Lodge F., Pothier C., Andréoli R. , Datcu M., *Normalized Mutual Information-based Ranking of Spatio-temporal Maps*. In proc. of the 8th Conf. on Image Information Mining : Knowledge Discovery from Earth Observation Data (ESA-EUSC 2012), CD-ROM, German Aerospace Centre (DLR), Oberpfaffenhofen, Germany, 4 pages, October 2012.

[9] Julea A., Méger N., Rigotti C., Trouvé E., Bolon Ph., Lazarescu V., *Mining Pixel Evolutions in Satellite Image Time Series for Agricultural Monitoring*. In proc. of Advances in Data Mining : Applications and Theoretical Aspects - 11th Industrial Conference on Data Mining (ICDM 2011), New-York, USA, August 2011, pp 189-203, ISBN : 978-3-642-23183-4.

[10] Julea A., Ledo F., Méger N., Trouvé E., Bolon Ph., Rigotti C., Fallourd R., Nicolas J-M., Vasile G., Harant O., Ferro-Famil L., Lodge F., *Polsar Radarsat-2 Satellite Image Time Series Mining Over the Chamonix Mont-Blanc Test Site*. In proc. of IEEE Int. Geoscience And Remote Sensing Symposium (IGARSS 11), Vancouver, Canada, July 2011, pp 1191-1194, doi :10.1109/IGARSS.2011.6049411.

[11] Méger N., Jolivet R., Lasserre C., Trouvé E., Rigotti C., Lodge F., Doin M-P., Guillaso S., Julea A. and Bolon Ph., *Spatio-Temporal Mining of ENVISAT SAR Interferogram Time Series over the Haiyuan Fault in China*. In proc. of the Sixth Int. Workshop on the Analysis of Multitemporal Remote Sensing Images (MULTITEMP'2011), Trento, Italy, July 2011, 4 pages, doi :10.1109/Multi-temp.2011.6005067.10.

[12] Julea A., Méger N., Rigotti C., Doin M.-P., Lasserre C., Trouvé E., Bolon P., Lazarescu V., *Extraction of Frequent Grouped Sequential Patterns from Satellite Image Time Series*. In proc of IEEE Int. Geoscience And Remote Sensing Symposium (IGARSS 10), Honolulu, HI, USA, July 2010, pp. 3434 - 3437, doi :10.1109/IGARSS.2010.5654127.

[13] Martin F., Méger N., Galichet S., Bécourt N., *Episode Rule-Based Prognosis Applied to Complex Vacuum Pumping Systems Using Vibratory Data*. In proc. of Advances in Data Mining : Applications and Theoretical Aspects - 10th Industrial Conference on Data Mining (ICDM 2010), Berlin, Germany, July 2010, pp. 376-389, ISBN :3-642-14399-7.

[14] Julea A., Méger N., Trouvé E., Bolon P., Rigotti C., Fallourd R., Nicolas J.-M., Vasile G., Gay M., Harrant O. et al, *Spatio-Temporal Mining of PolSAR Satellite Image Time Series*. In proc. of the 2010 European Space Agency (ESA) Living Planet Symposium, CD-ROM, Bergen, Norway, July, 6 pages.

[15] Martin F., Méger N., Galichet S., Bécourt N., *Data-Driven Prognosis Applied to Complex Vacuum Pumping Systems*. In proc. of Trends in Applied Intelligent Systems - 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2010), Cordoba, Spain, July 2010 pp. 468-477, doi :10.1007/978-3-642-13022-9_47.

[16] Le-Normand N., Boissiere J., Méger N., Valet L., *Supply chain management by means of FLM-rules*. In proc. of the 12th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'08), Induction of Process Models Workshop., Antwerpen, Belgium, September 2008, pp. 29-36.

[17] Julea A., Méger N., Trouvé E., Bolon Ph., *On extracting evolutions from satellite image time series*. In proc. of IEEE Int. Geoscience And Remote Sensing Symposium (IGARSS 08), Boston, MA, USA, July 2008, pp. 228- 231, doi :10.1109/IGARSS.2008.4780-069.

[18] Julea A., Méger N., Bolon Ph., *On mining pixel based evolution classes in satellite image time series*. In proc. of the 5th Conf. on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008), CD-ROM , ESRIN ESA Centre - Frascati, Italy, March 2008, 6 pages.

[19] Le Men C., Julea A., Méger N., Datcu M., Bolon Ph., Maître H., *Radiometric evolution classification in high resolution satellite image time series SITS*. In proc. of the 5th Conf. on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008), CD-ROM , ESRIN ESA Centre - Frascati, Italy, March 2008, 5 pages.

[20] Julea A., Méger N., Trouvé E., *On mining METEOSAT and ERS Multitemporal Images*. In proc. of the 4th Conf. on Image Information Mining for Security and Intelligence (ESA-EUSC 2006), CD-ROM , Torrejon Air Base - Madrid, Spain, November 2006, 6 pages.

[21] Julea A., Méger N., Trouvé E., *Sequential Patterns Extraction in Multitempo-*

ral Satellite Images. In proc. of the 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'06), Practical Data Mining Workshop : Applications, Experiences and Challenges, Berlin, Germany, September 2006, pp. 94-97.

[22] Méger N., Rigotti C., *Constraint-based Mining of Episode Rules and Optimal Window Sizes*. In proc. of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04), Pise, Italie, September 2004, pp. 313-324, doi :10.1007/978-3-540-30116-5_30.

[23] Méger N., Leschi C., Lucas N., Rigotti C., *Mining episode rules in STULONG dataset*. In proc. of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04), Discovery Challenge, Pise, Italy, September 2004, pp. 33-45.

[24] Bykowski A., Daurel T., Méger N., Rigotti C., *Finding Interesting Association Rules Using Confidence Variations*. In proc. of of the International Conference on Computer, Communication and Control Technologies (CCCT'03), Orlando, USA, August 2003, pp. 72-84.

[25] Bykowski A., Daurel T., Méger N., Rigotti C., *Association Maps as Integrity Constraints in Inductive Databases*. In proc. of the First International Workshop on Knowledge Discovery in Inductive Databases (KDID'02), Helsinki, Finland, August 2002, pp. 61-75.

7.6 Conférence nationale

[26] Martin F., Méger N., Galichet S., Bécourt N., *FLM-rule-based prognosis*. In proc of Extraction et Gestion de Connaissances (EGC 2011), "Data Mining, Applications, Cas d'Etudes et Success Stories" workshop, Brest, France, January 2011, 6 pages, online.

7.7 Thèse

[27] Méger N., *Recherche automatique des fenêtres optimales des motifs séquentiels*, INSA Lyon, Lyon, France, Décembre 2004.

7.8 Livrables de projets ANR

[28] Lodge F., Rigotti C., Méger N., *Multi-stage data mining processes for satellite images : research report*, FOSTER ANR project, 22 pages, December 2012.

[29] Méger N., *A parallel algorithm for mining FLM-rules on shared-memory machines*, pFLOWER ANR project, 14 pages, September 2012.

[30] Méger N., *State of the art on episode mining*, pFLOWER ANR project, 18 pages, September 2012.

[31] Lodge F., Rigotti C., Méger N., *Multi-stage data mining processes for satellite images : the state of the art*, FOSTER ANR project, 27 pages, July 2012.

[32] Julea A., Méger N., Rigotti C., Doin M-P., Lasserre C., Trouvé E., Bolon Ph., Lazarescu V., *EFIDIR report on local pattern-based atmospheric artefacts removal*, EFIDIR ANR project, 5 pages, January 2011.

[33] Rigotti C., A. Julea, Méger N., *SPATPAM (SPAtio-TemPorAl Mining) prototype, final version*, EFIDIR ANR project, July 2010.

[34] Rigotti C., A. Julea, Méger N., *SPATPAM (SPAtio-TemPorAl Mining) prototype, beta version*, EFIDIR ANR project, July 2009.

7.9 Séminaires

[35] Méger N., Julea A., Rigotti C., Bolon Ph., *Description spatio-temporelle non supervisée de séries d'images satellitaires par extraction de motifs séquentiels fréquents groupés*, GdR ISIS, journée *Techniques et algorithmes pour les séries d'images multi-temporelles*, Télécom ParisTech, Paris, France, June 2012.

[36] Méger N., *Fouille de données pour la mesure de déplacement par imagerie radar*, Institut de Physique du Globe (IPGS), Strasbourg, France, May 2012.

Deuxième partie

Travaux de recherche

Chapitre 8

Introduction

« *Computers have promised us a fountain of wisdom but delivered a flood of data.* »

—A frustrated management information system executive.

Cette citation, extraite de [FPSM91] pose clairement le problème de la valorisation des énormes quantités de données accumulées au sein des systèmes d'information. Quel que soit le domaine, la question n'est plus d'acquérir ou de stocker d'énormes volumes de données mais bien de les exploiter. Les données constituent en effet un réservoir de connaissances pouvant être mobilisées en vue de créer un avantage compétitif, que ce soit dans le monde de l'entreprise ou de la recherche scientifique. L'*Extraction de Connaissances dans les Données (ECD)* ou *Knowledge Discovery in Databases (KDD)* vient en réponse à ce défi et a été définie dans [FPSM91] comme « l'extraction non triviale, à partir des données, d'informations implicites, auparavant inconnues et potentiellement utiles ». En terme d'exploitation des données, cela amène à un changement de paradigme où le test d'hypothèse est exclu afin d'aider à la formation d'hyptohèses et de susciter ainsi la découverte de connaissances. L'ECD est un processus dans lequel l'utilisateur final joue un rôle central. En effet, il lui est demandé de participer à chacune des étapes de ce dernier que sont (i) la sélection des données, (ii) le nettoyage des données (gestion des valeurs manquantes et/ou aberrantes), (iii) la transformation des données (mise en forme compatible avec l'état de l'art technique), (iv) la *fouille des données* (ou *data mining*), c'est-à-dire l'extraction de descriptions des relations entre les données et (v) l'interprétation des descriptions précédemment extraites. Cette définition opérationnelle est détaillée dans [FPSS96]. À l'interactivé avec l'utilisateur final s'ajoute également l'itérativité au sein de ce processus : chaque résultat d'une étape peut mettre en cause les étapes et résultats précédents. Par extension et abus de langage, l'ECD est également appelé data mining. Cette étape est en effet cruciale et a suscité la grande majorité des travaux aujourd'hui publiés. Cela tient au fait que le volume des données, la dimension des données (le nombre d'attributs caractérisant les objets enregistrés), l'explosion combinatoire des relations à explorer et la nature hétérogène et répartie des données rendent l'extraction de descriptions de relations entre les données particulièrement ardue. Selon [HMS01], ces descriptions peuvent se scinder en deux familles : les *modèles* et les *motifs*. Un modèle décrit à lui seul un très large échantillon d'un jeu de données, il résume ce dernier. Les classifieurs, les clusterings, les modèles de régression linéaire en sont des exemples. À l'opposé, un motif, ou *motif*

local, décrit une relation entre quelques variables et/ou enregistrements. Autrement dit, le jeu de données ne peut être décrit dans sa globalité que par une collection de tels motifs. L'avantage de l'un est l'inconvénient de l'autre : une synthèse sera plus facilement accessible au travers d'un modèle tandis que plus de précisions sur les relations présentes au sein des données seront apportées par les motifs. Toujours selon [HMS01], quel que soit le type des descriptions utilisées, celles-ci peuvent être également construites à des fins d'inférence. Les classifieurs peuvent ainsi être utilisés afin de déterminer, sur des données autres que celles de l'apprentissage, la valeur d'une variable de classe en fonction d'autres variables.

Accueilli au laboratoire LISTIC à l'occasion de la création du département INFO de l'IUT d'Annecy, mes travaux de recherche se situent dans le cadre de l'ECD et concernent les motifs locaux, en particulier les motifs *séquentiels* tels que définis dans [AS95] ou dans [MTIV97]. Un motif séquentiel est un motif de la forme $\ll A \rightarrow B \gg$, ce qui est interprété comme « la propriété/l'objet B est observé quelques temps ou immédiatement après A ». Bien qu'unique data miner au sein du laboratoire, j'ai pu dérouler, en collaboration avec les experts des domaines concernés, le processus ECD tout en produisant des propositions originales relatives à l'étape de fouille de données. Au niveau thématique, ma recherche s'est organisée autour de deux axes (i) la description spatio-temporelle non supervisée de séries temporelles d'images satellitaires et (ii) la prévision d'événements dans un flot de données pour l'aide au pilotage de systèmes complexes. Le premier axe envisage l'extraction de motifs séquentiels à des fins de description tandis que le deuxième vise à l'inférence. Quel que ce soit l'axe considéré, et comme déjà indiqué au chapitre 3, mon travail de recherche s'articule autour de deux étapes, l'étude d'opportunité et l'étape de proposition. La première étape, l'étude d'opportunité, voit le besoin utilisateur recensé et détaillé. Les outils théoriques et techniques disponibles sont réutilisés le plus directement possible afin de valider ou non la présence d'une opportunité quant à de futurs développements. Si cette étape se révèle être concluante, alors une deuxième étape, dite de proposition, est engagée et vise à la création de solutions originales permettant d'améliorer les résultats obtenus lors de la première étape.

À un premier niveau, cette partie s'organise naturellement autour de chacune des thématiques de recherche développées. Le chapitre 9 est ainsi dédié à la description spatio-temporelle non supervisée de séries temporelles d'images satellitaires tandis que le chapitre 10 détaille les travaux relatifs à la prévision d'événements dans un flot de données pour l'aide au pilotage de systèmes complexes. À un deuxième niveau, au sein de chacun de ces chapitres, le lecteur retrouvera trois sections. La première, appelée *Contexte*, permet de situer le contexte thématique, de préciser les objectifs et de référencer les principales contributions existantes. La deuxième section, nommée *Étude d'opportunité*, présente les travaux relatifs à l'étude d'opportunité et donne les définitions de base. Enfin la troisième et dernière section, dont le titre est *Propositions*, détaille les propositions avancées et retenues à ce jour. Cette partie et ce document s'achèvent par le chapitre 11. Ce dernier offre une synthèse des travaux et les perspectives qui y sont associées.

Chapitre 9

Description non supervisée de Séries Temporelles d'Images Satellitaires (STIS)

Ce chapitre présente les travaux relatifs à la description spatio-temporelle non supervisée de séries d'images satellitaires. Après avoir détaillé le contexte thématique, les objectifs associés et les principales contributions existantes dans le section 9.1, présentation est faite de notre étude d'opportunité dans la section 9.2. Cette étude, qui met en avant l'utilisation de *motifs séquentiels fréquents* tels que définis dans [AS95], est reprise et complétée par la section 9.3 avec l'introduction de *motifs séquentiels fréquents groupés* et d'une mesure permettant d'exhiber ceux dont l'apparition est peu probable dans un jeu de données aléatoires où les fréquences des symboles sont préservées.

9.1 Contexte

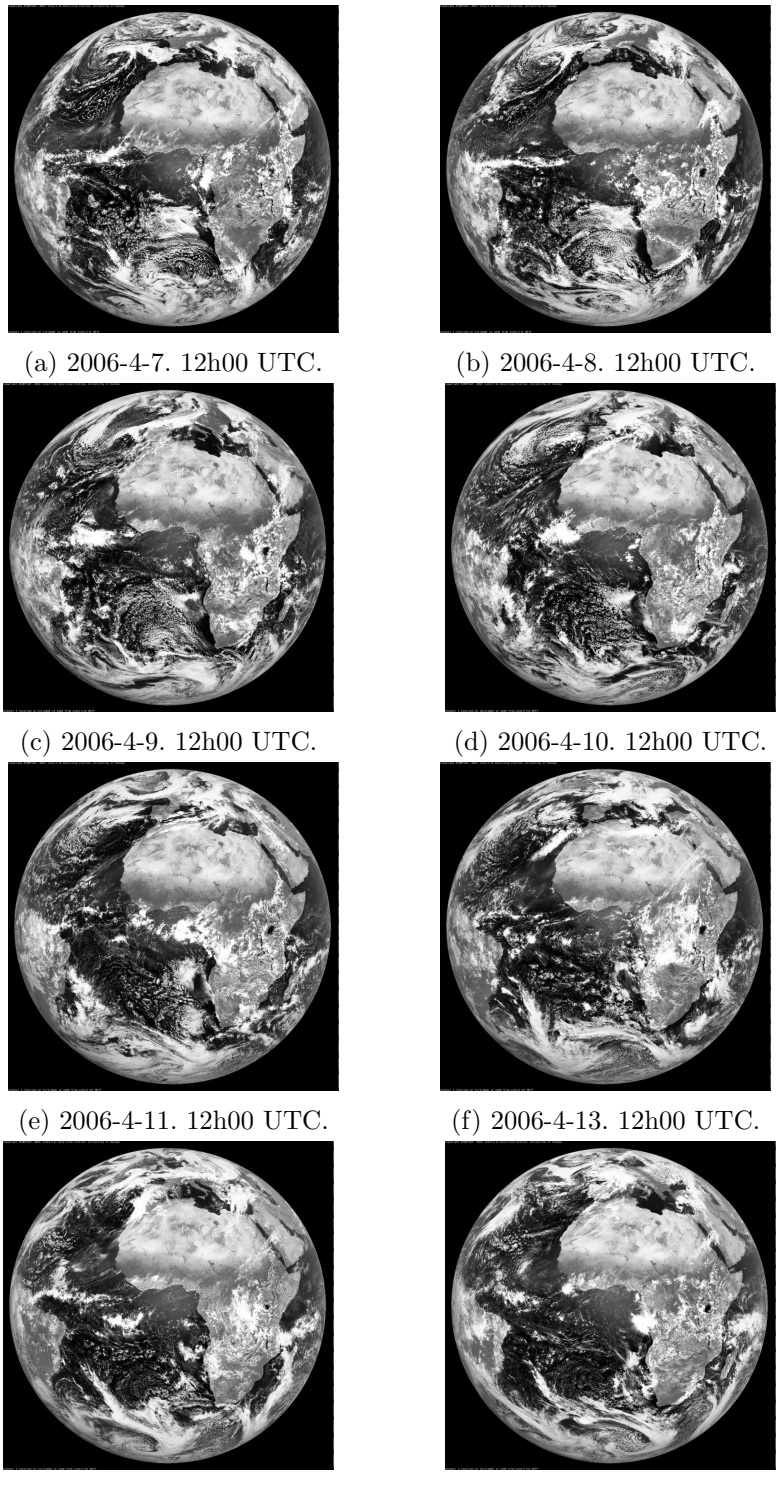
Le développement continu des techniques de télédétection permet l'accumulation de volumes de données toujours plus importants. En effet, les résolutions spatiales et temporelles sont sans cesse améliorées tandis que le nombre de canaux d'acquisitions est toujours plus élevé. À titre d'exemple, le satellite TerraSAR-X, équipé d'un radar et lancé en juin 2007, peut, à partir de son orbite polaire située à 514 km d'altitude, faire des acquisitions sous le même angle d'une même zone tous les 11 jours, cette zone pouvant mesurer 10×15 km, et ce à une résolution spatiale de 1 m. Il est ainsi possible de constituer des Séries Temporelles d'Images Satellitaires (STIS) couvrant une même zone géographique. Les STIS sont très utiles pour l'étude de phénomènes dont l'évolution est graduelle. La surveillance de cultures ou de la déformation de la croûte terrestre en sont des exemples typiques, exemples qui seront développés dans la suite de ce chapitre. Appliquer un processus ECD se révèle être très prometteur pour des utilisateurs voulant consolider et/ou enrichir leur connnaissance de la zone observée.

De par leur nature spatio-temporelle et les volumes de données concernés, les STIS constituent un véritable défi en terme d'ECD. Afin d'appréhender la difficulté de la tâche, la figure 9.1 présente une série d'acquisitions fournies par Meteosat-7 à 12h00 UTC les

7,8,9,10,11,13,14 et 15 avril 2006. Bien que les acquisitions de ce satellite géostationnaire situé à environ 36000 km d'altitude ne soient pas d'une grande résolution et que la connaissance nécessaire à l'analyse de ces images puisse être assez rudimentaire, force est de constater la difficulté à extraire manuellement une caractérisation spatio-temporelle des évolutions présentes dans cette STIS.

Les STIS peuvent être analysées à un autre niveau que le niveau des pixels, et ce après avoir identifié des objets ou des groupes de pixels formant des régions d'intérêt. Par exemple, dans [HD05], des caractéristiques spatiales, de texture ou bien encore spectrales sont extraites à l'aide de modèles stochastiques pour ensuite classifier, au sens du clustering, les données. Sur la base des clusters découverts, des graphes spatio-temporels sont inférés et proposés aux utilisateurs finaux. Quelques hypothèses doivent être posées : des modèles statistiques des images comme les champs aléatoires de Gibbs-Markov sont introduits, les clusters doivent suivre des formes gaussiènes et les graphes sont construits en prenant en compte des contraintes spatiales additionnelles. Comme proposé dans [HK01], des motifs spatio-temporels peuvent aussi être extraits d'une STIS au niveau des images elles-mêmes en fouillant une séquence de signatures. Tout d'abord, des Self-Organizing Maps (SOM) sont utilisées pour extraire la signature de chacune des images. La STIS est alors transformée en une séquence de signatures. Cette séquence est par la suite fouillée, sous des contraintes temporelles et une contrainte de fréquence, à la recherche de motifs séquentiels que sont les règles d'*épisodes sériels* tels que définis dans [MTIV97]. Une définition formelle de ces épisodes est également disponible dans la section 10.2.1. En revanche, la définition des occurrences utilisée diffère de celles proposées dans [MTIV97]. Ainsi, la règle $\ll A \Rightarrow B \gg$ est lue comme « si la signature A est observée une fois ou plus alors, immédiatement après ou quelques temps plus tard, la signature B est observée une fois ou plus ». À nouveau des hypothèses doivent être formulées et ont trait à l'échelle temporelle et spatiale des phénomènes observés. Ces techniques, bien que non supervisées, nécessitent néanmoins en entrée de formuler des hypothèses sur les caractéristiques/objets/régions qui doivent être identifiés et étudiés. Cela ne constitue en rien une tâche facile puisque les groupes de pixels ne forment pas toujours des objets dans une même image. Des perturbations atmosphériques ou des phénomènes d'occultation peuvent en effet altérer ces objets. De plus, un même objet peut voir son aspect modifié d'une image à l'autre de par un changement intrinsèque comme la fonte d'un glacier.

L'analyse des STIS au niveau pixel a ainsi retenu l'attention des chercheurs puisqu'elle ne requiert pas l'identification des objets à priori. Autrement dit, plutôt que de raisonner spatialement puis temporellement, ces techniques fonctionnent directement dans le temps et visent à former des objets définis par leur évolution temporelle. Ce type d'analyse s'appuie essentiellement sur les techniques de clustering. Dans ce cadre, un vecteur de caractéristiques est associé à chaque pixel et est utilisé pour comparer les pixels entre eux. Les caractéristiques utilisées peuvent être la moyenne, le minimum ou le maximum des valeurs que prend un pixel dans le temps. Cette proposition peut ainsi être retrouvée dans [NGSR96]. Néanmoins, l'utilisateur doit avoir une idée à priori des caractéristiques à prendre en compte. Si tel n'est pas le cas, alors il est également possible d'utiliser l'ensemble des valeurs que prennent les pixels dans le temps. Le clustering doit alors être effectué dans un espace à haute dimension. Un tel clustering peut être difficile à interpréter et nécessite un réglage attentionné des paramètres comme dans [GMC08] ainsi que des mesures de distances sophistiquées telle que l'adaptation de la distance d'édition de Levenshtein proposée dans [PIG12]. Dans ce dernier cas, la distance entre les séquences de valeurs des pixels pouvant être de taille variable est mesurée, ce qui permet de prendre en



(a) 2006-4-7. 12h00 UTC. (b) 2006-4-8. 12h00 UTC.

(c) 2006-4-9. 12h00 UTC. (d) 2006-4-10. 12h00 UTC.

(e) 2006-4-11. 12h00 UTC. (f) 2006-4-13. 12h00 UTC.

(g) 2006-4-14. 12h00 UTC. (h) 2006-4-15. 12h00 UTC.

FIGURE 9.1 – Acquisitions panchromatiques ($0.45 - 1.00\ \mu m$) fournies par Meteosat-7.

compte l'ordre induit par le temps tout en écartant des valeurs jugées comme traduisant la présence de perturbations atmosphériques et/ou de défauts d'acquisition. Ces approches effectuent donc une analyse au niveau pixel sans connaissance à priori des objets ou régions à surveiller. Cependant, elles nécessitent l'incorporation d'une connaissance du domaine exprimée au travers des définitions des caractéristiques à étudier et des mesures de distances à considérer. De plus, celles-ci ne permettent pas de trouver des zones qui se recouvrent, partiellement ou non.

D'autres approches, dites de *détection de changement*, génèrent une seule image où les changements sont représentés : une carte de changements est produite. Les techniques de détection de changement requièrent une connaissance à priori sur le type de changement qui doit être pris en compte et sont dédiées à des applications spécifiques. Soit l'utilisateur s'intéresse à des changements soudains tels que des inondations, des tremblements de terre, ou des catastrophes d'origine humaine (cf. [IFY$^+$03]), soit des changements graduels sont à prendre compte comme l'accumulation de biomasse (cf [VERR04]). Les techniques de détection de changement peuvent être appliquées au niveau du pixel (cf. [CJN$^+$04] ou [LMBM04]), au niveau des textures commme proposé dans [LL01] ou bien encore au niveau des objets comme détaillé dans [BBTD08].

Certains travaux en data mining tels que [CMC05, CMC07, HZZ08, ASBF$^+$12] s'appuient sur les *motifs séquentiels* tels que définis dans [AS95] pour analyser des jeux de données spatio-temporels. Pour rappel, un motif séquentiel est un motif de la forme « $A \rightarrow B$ », ce qui peut interprété comme « la propriété/l'objet B est observé quelques temps ou immédiatement après A ». Une définition formelle de ces épisodes est également disponible dans la section 9.2.1. Dans [CMC05], les motifs séquentiels fréquents représentants des sous-trajectoires d'objets, c'est-à-dire des séquences de localisations spatiales, sont extraites. Dans [CMC07], un unique objet est considéré et sa trajectoire est représentée comme une longue séquence à partir de laquelle les sous-trajectoires périodiques et fréquentes sont extraites. Une extension des motifs séquentiels, les motifs *spatio-séquentiels* est proposée dans [ASBF$^+$12]. Dans ce cas, plusieurs objets/régions sont considérés. Pour chaque objet/région, son évolution est décrite à la fois en fonction de ses propres attributs et des attributs des objets/régions qui lui sont connectés. Les évolutions les plus fréquentes spatialement et temporellement (une nouvelle mesure est proposée à cet effet) sont alors retenues. L'extraction de trajectoires peut aussi être envisagée avec d'autres motifs et critères. Par exemple, dans [GKS07], un motif est un groupe d'objets partageant un mouvement, une trajectoire commune en direction et en vitesse, le tout à une même date et à l'intérieur d'une même portion de l'espace. Des modèles globaux peuvent aussi être utilisés pour fouiller les trajectoires comme proposé dans [NP06]. L'algorithme utilisé est un algorithme de clustering basé sur la densité. Toutes ces techniques pourraient être adaptées au STIS pour analyser les STIS, après avoir identifié les objets d'intérêt. Dans [HZZ08], les motifs séquentiels fréquents sont utilisés pour représenter des relations spatio-temporelles dans des voisinages spatio-temporels de « data points », d'objets. Par exemple, si un motif séquentiel « $A \rightarrow B$ » est trouvé, alors il est interprété comme « les objets de type B tendent à apparaître autour et après les objets de type B ». Sur le principe, et avec des valeurs de paramètres appropriées (les voisinages spatio-temporels doivent être définis), cette approche pourrait être adaptée au STIS. Néanmoins, l'extraction est spatialement et temporellement contrainte et, à notre connaissance, aucune application à l'analyse de STIS n'a été rapportée.

En résumé, les méthodes existantes requièrent des connaissances à priori sur les objets/régions et/ou le type des évolutions contenues dans les STIS. Afin de susciter la

découverte de nouvelles connaissances, notre travail vise à réduire à sa plus simple expression l'appel à des connaissances du domaine et à exprimer les relations spatio-temporelles extraites d'une STIS le plus simplement possible. Enfin, lorsque cela est souhaité, un autre objectif est d'écarter de façon automatique les données affectées par des perturbations atmosphériques et/ou des défauts de capteur.

9.2 Étude d'opportunité

Cette étude a été menée dans le cadre du stage de master d'Andreea Maria Julea, de mars à juin 2006 (cf. chapitre 4). Ce master de l'Université Politehnica din Bucaresti est mis en œuvre par la faculté d'Électronique, Télécommunications et Technologies de l'Information et a pour intitulé *Images, Formes et Intelligence Artificielle*. Les financements utilisés proviennent à la fois de fonds propres du laboratoire et du projet ACI *Masses de Données* MEGATOR [1] porté par Emmanuel Trouvé.

9.2.1 Définitions préliminaires

Les objectifs mentionnés à la fin de la section 9.1 nous ont conduits à étudier l'utilisation des *motifs séquentiels* tels que définis dans [AS95]. En effet, une STIS peut être vue comme un ensemble de séquences temporelles. Plus précisément, considérons une STIS couvrant une même zone géographique à des dates différentes. Le postulat sur lequel nous nous appuyons pose que les images sont recalées, c'est-à-dire qu'elles sont directement superposables, et que les valeurs présentes au sein d'une même image sont correctement calibrées. À l'intérieur de chaque image, chaque pixel est associé à une valeur représentant une variable quantitative, par exemple l'intensité de rétrodiffusion de la portion de la zone géographique qu'il représente. Transformons ces valeurs de pixels en valeurs appartenant à une variable catégorique. Nous passons alors d'une représentation numérique à une représentation symbolique où chaque symbole/label permet de caractériser les états des pixels. Ces labels peuvent correspondre à des intervalles obtenus par quantification, à des classes de pixels résultants d'une classification non supervisée (par exemple en utilisant des clusterings de type K-means ou densité) ou à des classes fournies par l'utilisateur.

Définition 1 ***(label et état pixellaire)*** Soit $L = \{i_1, i_2, \ldots, i_s\}$ un ensemble contenant s symboles distincts appelés *labels* et utilisés pour encoder les valeurs associées aux pixels. Un *état pixellaire* est un couple (e, t) où $e \in L$ et $t \in \mathbb{N}$ tel que t est la date d'occurrence de e. La date t est simplement la date de l'image (ou le numéro d'ordre de l'image) dans laquelle la valeur correspondant à e a été observée.

Une *STIS symbolique* est alors un ensemble de *séquences d'évolution pixellaire*, chaque séquence décrivant les états d'un pixel dans le temps.

1. Mesure de l'Évolution des Glaciers Alpins par Télédétection Optique et Radar, ACI Masses de Données 2004-2007, http ://www.megator.fr.

Définition 2 ***(séquence d'évolution pixellaire et STIS symbolique)*** Soit un pixel p. Sa *séquence d'évolution pixellaire* est un couple $((x, y), seq)$ où (x, y) sont les coordonnées de p et seq est un n-uplet d'états pixellaires $seq = \langle(e_1, t_1), (e_2, t_2), ..., (e_n, t_n)\rangle$ contenant les états de p ordonnés par date croissante d'occurrence. Une *STIS symbolique* est alors un ensemble de séquences d'évolution pixellaire dont la cardinalité est égale au nombre de pixels formant chaque image/acquisition.

Une STIS symbolique typique est un ensemble contenant des millions de séquences d'évolution pixellaire. Chaque séquence contient la description symbolique des valeurs du pixel qu'elle décrit. Une STIS symbolique jouet contenant les états de quatre pixels est donnée par l'exemple 1.

Exemple 1

$$\begin{array}{c}((0,0), \langle(1,A),(2,B),(3,C),(4,B),(5,D)\rangle),\\ ((0,1), \langle(1,B),(2,A),(3,C),(4,B),(5,B)\rangle),\\ ((1,0), \langle(1,D),(2,B),(3,C),(4,B),(5,C)\rangle),\\ ((1,1), \langle(1,C),(2,A),(3,C),(4,B),(5,A)\rangle)\end{array}$$

Ce jeu de données décrit l'évolution de quatre pixels localisés aux positions $(0, 0)$, $(0, 1), (1, 0)$ et $(1, 1)$. Les dates $1, 2, 3, 4$ and 5 sont considérées : cinq images successives sont prises en compte. Les états pixellaires sont décrits en utilisant les symboles A, B, C et D. Par exemple, les états successifs du pixel localisé aux coordonnées $(1, 0)$ sont D, B, C, B et C.

Une *base de séquences* typiques, telle que définie dans [AS95], est un ensemble de séquences d'événements dans lequel chaque séquence dispose d'un identifiant unique. En ce qui concerne les STIS symboliques, si nous considérons les paires (x, y) donnant les coordonnées des pixels comme des identifiants des séquences d'évolution pixellaire correspondantes, alors une STIS symbolique est une base de séquences.

9.2.2 Motifs séquentiels fréquents

Dans une base de séquences, Agrawal et Srikant [AS95] proposent d'extraire des *motifs séquentiels*. Ceux-ci peuvent être définis comme suit [2] :

Définition 3 ***(motif séquentiel)*** Un *motif séquentiel* α est un tuple $\langle\alpha_1, \alpha_2, \ldots, \alpha_m\rangle$ où $\alpha_1, \ldots, \alpha_m$ sont des labels appartenant à L et m est la *longueur* d'α. Un tel motif est aussi noté $\alpha_1 \to \alpha_2 \to \ldots \to \alpha_m$.

L'occurrence d'un motif séquentiel et le support de ce dernier sont alors définis ainsi :

Définition 4 ***(occurrence et support)*** Soient $\mathcal{S}$ une STIS symbolique et $\alpha = \alpha_1 \to \alpha_2 \to \ldots \to \alpha_m$ un motif séquentiel. Alors $((x, y), \langle(\alpha_1, t_1), (\alpha_2, t_2), \ldots, (\alpha_m, t_m)\rangle)$, où $t_1 < t_2 < \ldots < t_m$, est une *occurrence* de α dans $\mathcal{S}$ s'il existe $((x, y), seq) \in \mathcal{S}$ telle que (α_i, t_i)

2. Alors que plusieurs éléments peuvent apparaître à une même date selon les définitions originelles, nous considérons qu'une date est ici associée à un unique élément.

apparaît dans *seq* pour tout i appartenant à $\{1, \ldots, m\}$. Une telle séquence d'évolution pixellaire $((x, y), seq)$ est dite *supporter* α. Le *support* d'α dans $\mathcal{S}$, noté $support(\alpha)$, est le nombre de séquences dans $\mathcal{S}$ qui supportent α.

En ce qui concerne l'exemple 1, les occurrences du motif séquentiel $A \rightarrow C \rightarrow B$ sont [3] :

$$((0,0), \langle (1,A), (3,C), (4,B) \rangle),$$
$$((0,1), \langle (2,A), (3,C), (4,B) \rangle),$$
$$((0,1), \langle (2,A), (3,C), (5,B) \rangle),$$
$$((1,1), \langle (2,A), (3,C), (4,B) \rangle)$$

En effet, le motif $A \rightarrow C \rightarrow B$ apparaît dans la séquence d'évolution pixellaire du pixel localisé à la position $(0, 1)$. Le label A apparaît à la date 2, le label C apparaît à la date 3 et le label B apparaît aux dates 4 et 5 : deux occurrences différentes peuvent ainsi être considérées pour le pixel localisé à la position $(0, 1)$. Ça n'est pas le cas pour tous les autres pixels. Plus précisément, le motif $A \rightarrow C \rightarrow B$ apparaît seulement une fois pour les pixels localisés aux positions $(0, 0)$ et $(1, 1)$, et aucune occurrence n'est observée pour le pixel localisé à la position $(1, 0)$. Bien que quatre occurrences puissent être trouvées dans la STIS symbolique définie par l'exemple 1, le motif $A \rightarrow C \rightarrow B$ apparaît seulement à trois positions différentes, dans trois séquences d'évolution différentes. Autrement dit, le motif $A \rightarrow C \rightarrow B$ affecte seulement trois pixels différents. Son support est ainsi $support(A \rightarrow C \rightarrow B) = 3$. Pour finir, le lecteur remarquera qu'un label peut être répété à l'intérieur d'un même motif. Par exemple, le motif $C \rightarrow C$ a deux occurrences, une dans la troisième séquence d'évolution (position $(1, 0)$) et une dans la quatrième séquence d'évolution (position $(1, 1)$).

Définition 5 ***(motif séquentiel fréquent)*** Soit σ un entier strictement positif appelé *support minimum*. Soit α un motif séquentiel, alors α est un *motif séquentiel fréquent* si $support(\alpha) \geq \sigma$. Le support minimum peut aussi être spécifié comme un seuil relatif tel que $\sigma_{rel} \in [0, 1]$. Un motif α est alors fréquent si $support(\alpha)/|\mathcal{S}| \geq \sigma_{rel}$, où $\mathcal{S}$ est le jeu de données (c'est-à-dire une STIS symbolique dans notre cas) et $|\mathcal{S}|$ est le nombre de séquences contenues dans $\mathcal{S}$.

Au final, le support d'un motif est simplement une aire, une surface exprimée en nombre de pixels. De même, le support minimum peut être compris comme une surface minimum. Les pixels affectés par un motif sont également dits *couverts* par un motif. Réciproquement, un motif est dit *couvrir* des pixels. De façon plus formelle, un *pixel couvert* est défini comme suit :

Définition 6 ***(pixel couvert)*** Un pixel associé à une séquence d'évolution $((x, y), seq)$ est *couvert* par un motif séquentiel α si α a au moins une occurrence dans *seq*. L'ensemble des coordonnées des pixels couverts par α est noté $cover(\alpha)$. Par définition, $|cover(\alpha)| = support(\alpha)$.

3. Le lecteur remarquera que les éléments d'une occurrence ne sont pas forcément contigus en temps.

À la lumière de ces définitions, le choix de l'extraction de motifs séquentiels fréquents apparaît comme étant à même de satisfaire une grande partie des objectifs mentionnés à la fin de la section 9.1 En effet :

- ces motifs sont d'une lecture aisée par l'utilisateur final,
- la définition des occurrences des motifs séquentiels n'impose ni de considérer des dates immédiatement successives ni de considérer des occurrences synchrones d'un pixel à l'autre ; ce qui permet, pour certains des motifs, de ne pas considérer les défauts liés aux perturbations atmosphériques et/ou aux capteurs,
- aucune contrainte temporelle n'est posée,
- une seule contrainte spatiale est posée et est directement traduisible en nombre de pixels minimum ou en surface minimum,
- aucun à priori sur la forme des phénomènes observés n'est posé,
- un large éventail d'échelles d'analyse est accessible : tous les phénomènes dont la surface est supérieure au seuil de support minimum sont pris en compte.

La réutilisation des définitions des motifs séquentiels et de leurs occurrences, proposée dans [JMT06b] et [JMT06a], nous permet de bénéficier de l'effort de recherche effectué dans ce domaine pour développer des techniques d'extraction efficaces (cf. [AS95, GRK99, MCP98, PHMAP01, PHW07, SA96, Zak00, Zak01]).

9.2.3 Expériences

Afin d'évaluer le potentiel et la généricité de l'approche retenue, nous avons mené des expériences à la fois sur des données optiques et radar. Pour ce faire, nous avons utilisé le prototype public de Mohammed J. Zaki (http ://www.cs.rpi.edu/ zaki/www-new/pmwiki.php/Software/Software) qui implémente l'algorithme cSPADE [Zak00]. Toutes les expériences ont été réalisées sur un PC standard équipé d'un processeur AMD Athlon(tm) 64 3000+ (1800MHz) avec 512 Mo de RAM fonctionnant sous Linux (kernel 2.6).

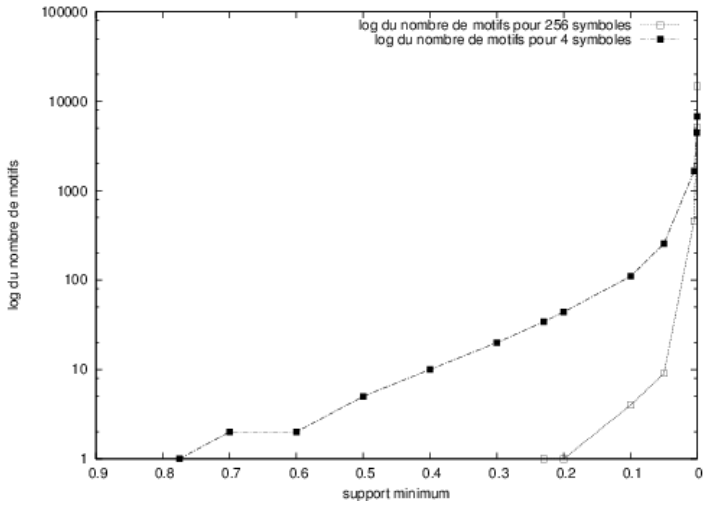
Données et prétraitement des données Nous présentons ici les résultats les plus détaillés à savoir ceux obtenus sur des données optiques. Nous avons utilisé des images panchromatiques (0.45 μm - 1 μm) du satellite géostationnaire Meteosat-7 dont un des objets est la surveillance de l'évolution de la couverture nuageuse. Celles-ci ont été obtenues via le site de la station de réception NERC des images satellitaires de l'Université de Dundee au Royaume-Uni (`http://www.sat.dundee.ac.uk`). Bien que gratuites, ces images sont fournies dans un format dégradé par la compression jpeg. La résolution varie entre 1 $km \times 1$ km et 2.5 $km \times 2.5$ km selon la courbure de la terre et donc l'éloignement et l'orientation par rapport au satellite. Afin de faciliter l'interprétation, nous avons sélectionné une zone géographique qui nous est familière et qui s'étend du nord de la France au sud de l'Algérie. En ce qui concerne la dimension horizontale de l'image, nous avons gardé toute l'étendue possible pour prendre en compte et analyser les influences de l'Atlantique Nord et pour traiter la plus grande image possible. Ainsi, nous avons obtenu des images de 905 lignes et de 2500 colonnes, soit 2 262 500 pixels. Cette sélection a été appliquée sur une collection de huit images acquises les 7, 8, 9, 10, 11, 13, 14 et 15 avril 2006, à 12.00 UTC. Cette heure a été choisie afin d'obtenir une illumination maximale et comparable d'un jour à l'autre pour la zone géographique considérée. Les acquisitions originales (avant sélection) sont présentées dans la figure 9.1. Nous avons cherché à regrouper les valeurs des pixels

en un minimum d'intervalles afin de diminuer l'impact des défauts dus à l'acquisition des images et à la compression jpeg. Après analyse des différentes zones géographiques présentes dans les images, nous avons décidé de considérer quatre intervalles disjoints et contigus pour lesquels nous avons utilisé les symboles « 0 », « 1 », « 2 », et « 3 ». Le symbole « 0 » regroupe les valeurs appartenant à l'intervalle [0, 50] et peut être associé à une présence d'eau (mer, océan) ou de végétation. Le symbole « 1 » rassemble les valeurs appartenant à l'intervalle]50, 100] et indique des zones couvertes de terre ou de nuages foncés. Le symbole « 2 » qui correspond aux valeurs appartenant à l'intervalle]100, 200] représente les zones de sable ou les zones couvertes par des nuages assez clairs. Enfin, le symbole « 3 » correspond aux valeurs appartenant à l'intervalle]200, 255], valeurs qui tracent la présence de nuages très clairs ou de neige. Cette interprétation est conforme à celle donnée sur le site de Meteosat (`http://www.eumetsat.int`).

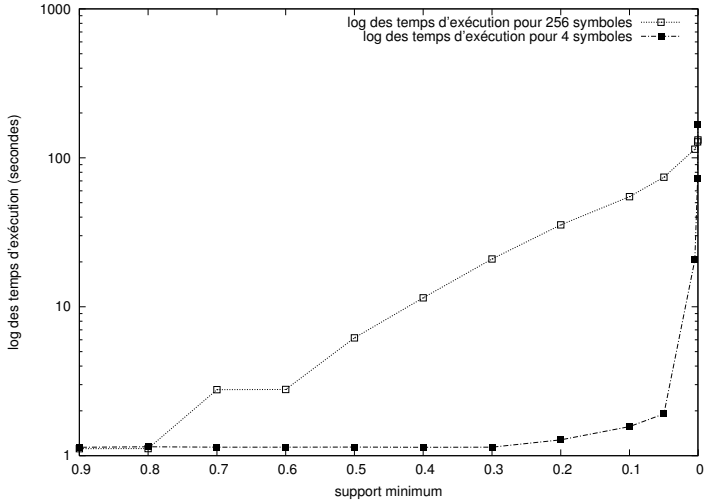
Analyse quantitative Afin de tester les temps d'exécution et le nombre de motifs trouvés en fonction de différents supports possibles, nous avons effectué une série d'expériences dont les résultats sont représentés par les courbes de la figure 9.2 (échelles logarithmiques). Pour différentes valeurs de σ, nous avons pris le temps moyen observé sur 50 éxécutions du prototype. De plus, et afin de tester les performances du prototype, nous avons également mené les mêmes expériences en associant un symbole à chaque niveau de gris de l'intervalle $[0 - 255]$ [4]. Une première remarque s'impose : le nombre de motifs et les temps d'éxécution croissent de manière exponentielle avec la diminution du support minimum σ (cf. figure 9.2a). Ceci est un comportement classique, car plus σ diminue, plus le nombre de symboles fréquents est élévé, et plus il est possible de tester des motifs composés à partir de ces symboles fréquents. Les premiers motifs aparaissent à $\sigma = 77.5\%$ pour une discrétisation à 4 symboles, et à $\sigma = 23\%$ pour une discrétisation à 256 symboles. Pour $\sigma \geq 0.5\%$ le nombre de motifs dans le cas d'une discrétisation à 4 symboles est plus élevé que dans le cas d'une discrétisation à 256 symboles. Ceci est logique car plus on utilise de symboles, plus les supports de ces symboles diminuent, et moins il est possible de trouver des motifs fréquents à des supports élevés. C'est également ce qu'indiquent les temps d'exécution de la figure 9.2b, pour laquelle il est possible de constater que les valeurs de ces temps sont inférieures dans le cas d'une discrétisation à 256 symboles aux temps obtenus pour une discrétisation à 4 symboles. En revanche, avec un très bas support c'est-à-dire $\sigma < 0.5\%$, la majorité des 256 symboles est fréquente. Dans ce cas, de très nombreux motifs peuvent être composés et on obtient plus de motifs qu'avec une discrétisation impliquant moins de symboles. Par exemple, pour $\sigma = 0.05\%$ le nombre de motifs pour la discrétisation à 256 symboles est 14717 et celui pour la discrétisation à 4 symboles est 6787. Enfin, notons que le volume des motifs extraits reste, à l'exception des très bas supports, relativement faible, ce qui facilite d'autant plus l'interprétation des résultats.

Analyse qualitative Les premiers motifs qui apparaissent en baissant le support minimum sont 2 (un motif composé d'un seul symbole, le symbole « 2 ») pour un support minimum de 77.5% et 3 pour un support minimum de 72.5%. Cela signifie que pour plus de 77.5% des pixels, dans au moins une image de la séquence de huit, la valeur du pixel est 2

4. Nous ne commenterons pas au niveau qualitatif les résultats des expériences faites pour une discrétisation à 256 symboles car l'information considérée est bien trop fragmentée pour pouvoir espérer proposer une interprétation des motifs extraits. Par ailleurs, cette information est également très exposée aux défauts introduits par la compression jpeg.



(a) Nombre de motifs en fonction du support minimum (de 0 à 90%).



(b) Temps d'éxécutions en fonction du support minimum (de 0 à 90%).

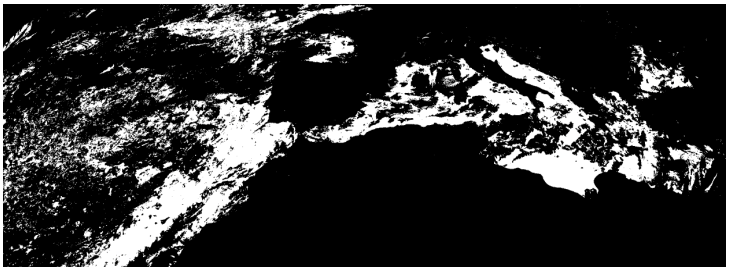
FIGURE 9.2 – Le nombre de motifs SFG (a) et les temps d'exécution (b) en fonction du seuil de support minimum σ.

(sable/nuage clair). Le même type d'analyse peut être conduit pour le motif 3 (nuages très clair/neige). Le motif 1, qui peut représenter le même type d'information au niveau des zones continentales ou des nuages fins et translucides sur de l'eau apparaît pour $\sigma = 57.5\%$. Le dernier 1-motif (motif de longeur 1) est le 0 (mer, océan) avec $\sigma = 52.5\%$. Ceci nous donne donc le nombre minimum de pixels qui, dans une des 8 images, indiquent des zones couvertes par la mer ou la végétation. Le premier 2-motif, $2 \rightarrow 2$, sort pour $\sigma = 52.5\%$. Il est suivi par les motifs $3 \rightarrow 3$ et $0 \rightarrow 0$ pour $\sigma = 47.5\%$. Ainsi, dans au moins deux images, au minimum 52.5% des pixels indiquent des zones de nuages claires/sables, et au minimum 47.5% des pixels indiquent des zones de nuages épais/neige ($3 \rightarrow 3$) et des zones de mer/végétation ($0 \rightarrow 0$). Avec un support minimum assez élevé, 45%, apparaissent les passages de nuages, $2 \rightarrow 3$ et $3 \rightarrow 2$ (des nuages assez clairs laissent place à des nuages très clairs ou l'inverse). Nous avons également caractérisé le passage de nuages sur les océans et les mers avec les motifs suivants :

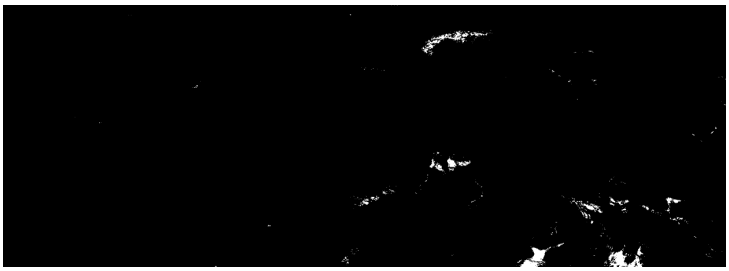
- $0 \rightarrow 0 \rightarrow 3$ ($\sigma = 25\%$)
- $3 \rightarrow 0 \rightarrow 0$, $0 \rightarrow 3 \rightarrow 0$ ($\sigma = 22.5\%$)
- $0 \rightarrow 0 \rightarrow 2$, $2 \rightarrow 0 \rightarrow 0$, $0 \rightarrow 2 \rightarrow 0$ ($\sigma = 20\%$)
- $0 \rightarrow 0 \rightarrow 1$, $1 \rightarrow 0 \rightarrow 0$, $0 \rightarrow 1 \rightarrow 0$ ($\sigma = 20\%$)
- $0 \rightarrow 0 \rightarrow 0 \rightarrow 3$, $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ ($\sigma = 17.5\%$)
- $0 \rightarrow 3 \rightarrow 0 \rightarrow 0$ ($\sigma = 15\%$)
- $0 \rightarrow 0 \rightarrow 1 \rightarrow 0$ ($\sigma = 14\%$)
- $0 \rightarrow 0 \rightarrow 0 \rightarrow 1$, $1 \rightarrow 0 \rightarrow 0 \rightarrow 0$ ($\sigma = 13\%$)
- $0 \rightarrow 0 \rightarrow 2 \rightarrow 0$, $0 \rightarrow 2 \rightarrow 0 \rightarrow 0$ ($\sigma = 13\%$)
- $3 \rightarrow 0 \rightarrow 0 \rightarrow 0$ ($\sigma = 13\%$)
- $0 \rightarrow 0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ ($\sigma = 12\%$)
- $2 \rightarrow 0 \rightarrow 0 \rightarrow 0$ ($\sigma = 12\%$)
- $0 \rightarrow 0 \rightarrow 3 \rightarrow 0 \rightarrow 0$ ($\sigma = 11\%$)
- $0 \rightarrow 0 \rightarrow 0 \rightarrow 2$ ($\sigma = 11\%$)

Comme montré par la figure 9.3a, le motif $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ couvre principalement des zones maritimes. Une large partie de ces zones étant affectée par ce motif, il est même possible de distinguer l'Afrique du Nord et l'Europe (pixels noirs). Le dernier type de phénomène à être découvert concerne des pixels dont l'état ne change pas ou peu. Ainsi nous avons trouvé le motif $0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0$ à $\sigma = 1.4\%$, qui indique que certaines zones maritimes ne sont pas recouvertes par les nuages. Un autre motif intéressant, $3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$, découvert pour $\sigma = 0.7\%$, fait ressortir, comme le montre la figure 9.3b, les zones enneigées des Alpes et les zones désertiques très claires de l'Afrique du Nord.

Il est possible de raffiner les localisations spatiales en introduisant des couleurs représentant les informations temporelles et en considérant les occurrences au plus tôt. Nous avons ainsi créé un image couleur 8 bits en mettant à 0 tous les pixels qui ne sont pas affectés par le motif considéré et en attribuant une valeur différente de 0 aux autres pixels. Cette dernière valeur est obtenue en mettant le $i^{ème}$ bit à 1 si un des labels composant le motif est présent dans la $i^{ème}$ image. La figure 9.4 donne un exemple d'un tel codage pour le motif $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$. Par exemple, les zones en rouge (correspondant au code 99, 1100011 en binaire) mettent en avant les zones où le motif apparaît dès le premier jour, se déroule sur le second et le sixième jour pour s'achever le septième jour. Bien que limité au nombre de bits disponibles, ce type de localisation spatio-temporelle est intéressant pour les utilisateurs finaux car il est possible d'observer simultanément où une évolution



(a) $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$.



(b) $3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$.

FIGURE 9.3 – Localisation (pixel blancs) des motifs $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ (a) et $3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$ (b).

se produit et quelles sont les zones qui sont affectées de façon synchrone.

Synthèse Au final, les motifs fréquents extraits lors de ces expériences ne sont pas trop nombreux (à l'exception des très faibles supports). Ils sont faciles à interpréter conjointement avec la mesure de support et ils permettent de considérer les deux grandes catégories de phénomènes présentes dans cette STIS : les changements (le passage de nuages épais sur des zones de la mer) et les non changements (présence récurrente de mer/végétation, de neige/nuages épais). Les phénomènes retrouvés sont connus et on peut conclure qu'une description pertinente des différents changements présents dans cette série d'images peut être obtenue à l'aide de motifs séquentiels. Ces expériences sont également présentées et détaillées dans [JMT06a], pour ce qui est de la communauté télédétection, et dans [JMT06b] pour ce qui est de la communauté data mining. Des expériences préliminaires sur des images radar confirment le potentiel de l'approche et sont décrites dans [JMT06a]. Dans [JMB08] et [JMTB08], d'autres expériences sur des données optiques de résolution plus élevée ($20\ m \times 20\ m$) confirment également le caractère générique de la méthode proposée. Notre proposition est d'ailleurs reprise dans [PVB$^+$11] et [PMGF11]. Dans [PVB$^+$11] une application à la détection de zones cultivées est présentée. Dans [PMGF11] est proposée une extraction guidée vers les seuls changements (interdiction de composer des motifs comprenant deux symboles fréquents identiques à la suite) et vers des phénomènes qui en plus de couvrir une surface minimum ne doivent pas dépasser une surface maximum. Nous pensons que cela peut être restrictif dans le cadre d'un apprentissage non supervisé : il serait par exemple impossible de décrire les zones enneigées ou certaines zones désertiques. De plus, il est également proposé de considérer, à chaque date d'acquisition, plusieurs informations (plusieurs bandes/canaux)

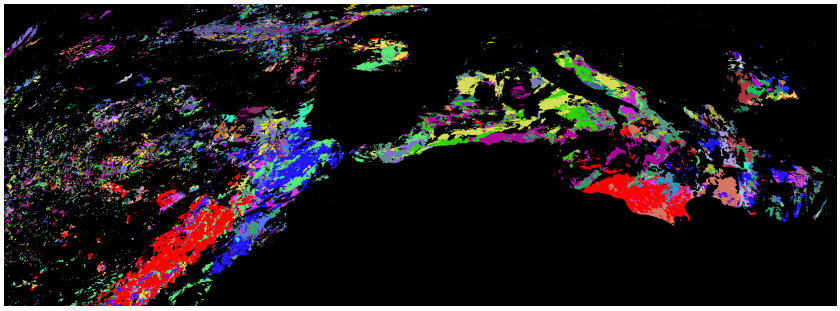


FIGURE 9.4 – Localisation spatio-temporelle du motif séquentiel $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$.

pour chaque pixel. L'objectif poursuivi est d'extraire des motifs plus détaillés où chaque état pixellaire peut être décrit par plusieurs labels, ce qui rejoint les définitions originelles proposées dans [AS95] et correspond également à des propositions que nous avions formulées dans [JMT06a] et [JMT06b]. Néanmoins, et à moins d'imposer les limitations proposées dans [PMGF11], ce type d'approche conduit à une explosion du nombre de motifs fournis à l'utilisateur. Enfin, l'ensemble des expériences menées, et pour partie présentées dans ce mémoire, montrent qu'il est possible de ne considérer qu'une seule bande (souvent synthétique), qu'un seul symbole, pour décrire les états pixellaires et extraire des motifs pertinents pour l'utilisateur final.

9.3 Propositions

Une partie des propositions présentées et retenues dans cette section a fait l'objet d'une étude menée dans le cadre la thèse d'Andreea Maria Julea, de mars 2007 à septembre 2011 (cf. chapitre 4). Cette thèse s'est déroulée en co-tutelle, entre l'Université de Savoie et l'Université Politehnica din Bucaresti. Les financements utilisés proviennent d'un projet BQR dédié à la fusion d'information pour l'évaluation du risque lié aux glaciers ou aux volcans, du projet ANR EFIDIR porté par Emmanuel Trouvé et du projet ANR FOSTER pour lequel je suis responsable scientifique du LISTIC (cf. chapitre 5).

9.3.1 Connexité spatiale : les motifs SFG

Une simple mesure de surface peut vite révéler ses limites. En effet rien n'interdit d'extraire un motif séquentiel couvrant des pixels épars, isolés les uns des autres. Dans ce cas, le phénomène observé n'a aucune cohérence spatiale et est généralement dû simplement à du bruit, des perturbations atmosphériques ou des défauts de capteur. Afin de concentrer l'extraction sur des motifs couvrant des pixels qui forment des régions dans l'espace, nous proposons une mesure additionnelle, la *connexité moyenne*, basé sur un 8-voisinage [FDHF+05]. Elle est définie comme suit :

Définition 7 ***(connexité locale)*** Soit une STIS symbolique $\mathcal{S}$ et soit $occ((x, y), \alpha)$ une

fonction qui, en fonction de coordonnées spatiales (x, y) d'un motif α, indique si α apparaît dans $\mathcal{S}$ à la position (x, y). Plus précisément, $occ((x, y), \alpha)$ renvoie 1 si et seulement si il y a une séquence seq dans $\mathcal{S}$ à la position (x, y) et si α apparaît dans $((x, y), seq)$. La fonction $occ((x, y), \alpha)$ renvoie 0 dans tous les autres cas. Si α apparaît dans $((x, y), seq)$, alors sa *connexité locale* à la position (x, y) est
$LC((x, y), \alpha) = [\sum_{i=-1}^{i=1} \sum_{j=-1}^{j=1} occ((x + i, y + j), \alpha)] - 1$.

La valeur de $LC((x, y), \alpha)$ est le nombre de pixels dans le 8-voisinage de (x, y) qui ont une évolution (ou séquence d'évolution pixellaire) supportant α. La somme est décrémentée de 1 pour ne pas compter l'occurrence d'α à la position (x, y). Dans l'exemple 1, et en se limitant à 4 pixels, pour les motifs séquentiels $A \to C \to B$ et $C \to C$, nous avons :

$$\begin{gathered} LC((0,0), A \to C \to B) = 2 \\ LC((0,1), A \to C \to B) = 2 \\ LC((1,1), A \to C \to B) = 2 \\ LC((0,1), C \to C) = 1 \\ LC((1,1), C \to C) = 1 \end{gathered}$$

Définition 8 ***(connexité moyenne)*** La *connexité moyenne* d'α est définie comme :

$$AC(\alpha) = \frac{\sum_{(x,y) \in cover(\alpha)} LC((x,y),\alpha)}{|cover(\alpha)|}$$

Cette mesure donne, pour les pixels couverts par α, le nombre moyen de voisins dans leur 8-voisinage qui sont également couverts par α. Dans l'exemple 1, on a $AC(A \to C \to B) = 6/3 = 2$ et $AC(C \to C) = 2/2 = 1$. Finalement, nous définissons les *motifs séquentiels fréquents groupés* comme suit :

Définition 9 ***(motif SFG)*** Soient $\mathcal{S}$ une STIS symbolique, α un motif séquentiel fréquent dans $\mathcal{S}$, et κ un nombre réel positif appelé *connexité moyenne minimum*. Le motif α est un *motif Séquentiel Fréquent Groupé* (*motif SFG*) si $AC(\alpha) \geq \kappa$ dans $\mathcal{S}$.

Ainsi, dans l'exemple 1, si $\sigma = 2$ et si $\kappa = 2$, alors $A \to C \to B$ est un motif séquentiel fréquent groupé tandis que le motif $C \to C$ n'en n'est pas un. La notion des motifs SFG a été introduite auprès de la communauté télédétection dans [JMR$^+$10] et [JMB$^+$11].

Comme mentionné dans la section 9.2, différentes techniques efficaces d'extraction de motifs séquentiels sont disponibles et utilisables dans notre contexte. Une solution naïve consiste à extraire les motifs séquentiels fréquents, et à retenir, dans une phase de post-processing, ceux satisfaisant à la contrainte de connexité moyenne $AC(\alpha) \geq \kappa$. Ainsi que cela est rapporté dans [JMB$^+$11], il est possible puisque le surcoût de traitement engendré est linéaire par rapport au nombre de motifs post-traités. Une autre solution consiste à pousser partiellement cette contrainte dans le processus d'extraction afin d'élaguer l'espace de recherche. Celle-ci a été proposée auprès de la commnunauté data mining dans [JMR$^+$11] et [JMR$^+$12]. En effet, la contrainte de connexité moyenne ne correspond à aucune classe de contraintes pouvant être utilisées à des fins d'élagage. Les deux classes principales sont les contraintes anti-monotones (si un motif ne satisfait pas une contrainte aucun des sur-motifs ne la satisfait) et monotones (si un motif satisfait la contrainte tous ses sur-motifs la satisfont également). Dans le cas des motifs séquentiels tels que définis dans ce mémoire, la notion de *sur-motif* peut être définie ainsi :

Définition 10 ***(sur-motif)*** Un motif séquentiel $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \ldots \rightarrow \beta_m$ est un *sur-motif* d'un motif séquentiel $\alpha = \alpha_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_n$ s'il existe des entiers $1 \leq i_1 < i_2 < \ldots < i_n \leq m$ tels que $\alpha_1 = \beta_{i_1}$, $\alpha_2 = \beta_{i_2}$, ..., $\alpha_n = \beta_{i_n}$.

Il est évident que la contrainte de connexité moyenne n'est ni anti-monotone, ni monotone. Il est donc facile de montrer qu'elle n'est ni anti-monotone ou monotone par rapport au préfixe (cf. [PHW07]). En effet, dans ce cas, seule une partie des relations de spécialisation est exploitée, c'est-à-dire celles permettant la construction des préfixes par ajout d'un symbole en fin de ces derniers. De plus elle n'appartient à aucune classe de contraintes utilisées pour l'extraction de motifs fréquents telles que les contraintes succinctes (cf. [NLHP98]), convertibles (cf. [PHL01]) ou faiblement anti-monotones (cf. [BL05]).

Le push partiel de la contrainte de connexité moyenne s'appuie sur l'observation suivante : pour tous les motifs α, puisque $|cover(\alpha)| \geq \sigma$, alors

$$AC(\alpha) = \frac{\sum_{(x,y)\in cover(\alpha)} LC((x,y),\alpha)}{|cover(\alpha)|} \leq \frac{\sum_{(x,y)\in cover(\alpha)} LC((x,y),\alpha)}{\sigma}.$$

Un motif séquentiel fréquent α qui ne satisfait pas $\frac{\sum_{(x,y)\in cover(\alpha)} LC((x,y),\alpha)}{\sigma} \geq \kappa$ ne peut pas être un motif SFG. Et, si nous considérons la conjonction de contraintes

$$\mathcal{C} = support(\alpha) \geq \sigma \wedge \frac{\sum_{(x,y)\in cover(\alpha)} LC((x,y),\alpha)}{\sigma} \geq \kappa,$$

celle-ci apparaît comme anti-monotone puisque la valeur $\sum_{(x,y)\in cover(\alpha)} LC((x,y),\alpha)$ ne peut augmenter pour les sur-motifs d'α. Cette conjonction de contraintes peut donc être utilisée de façon active pour élaguer l'espace de recherche.

L'ensemble des algorithmes d'extraction de motifs séquentiels gèrent et poussent les contraintes anti-monotones au sein du processus d'extraction. Nous avons décidé d'intégrer la conjonction anti-monotone $\mathcal{C}$ dans l'algorithme *PrefixGrowth* proposé dans [PHW07]. Cet algorithme récent est efficace et permet de gérer facilement les contraintes anti-monotones. Mis à part le fait de vérifier $\mathcal{C}$ pour élaguer l'espace de recherche, la seule modification est de vérifier, avant de valider un motif α que $AC(\alpha) \geq \kappa$, puisque $\mathcal{C}$ n'implique pas la satisfaction de la contrainte de connexité moyenne. L'implémentation de l'algorithme a été fait en C en utilisant nos propres structures de données.

9.3.2 Expériences sur les motifs SFG

Dans cette section, des résultats obtenus à la fois sur des STIS optiques et radar sont présentés et rendent compte de l'efficacité et de la généricité de la méthode proposée. Toutes les expériences ont été menées avec un PC standard équipé d'un Intel Core 2 @3GHz avec 4 GB RAM fonctionnant sous Linux (kernel 2.6). Comme évoqué dans la section 9.3.1, nous utilisons cette fois-ci notre propre moteur d'extraction écrit en C. Ce dernier, SPATPAM (SPAtio-TemPorAl Mining), est diffusé gratuitement en tant que délivrable du projet ANR EFIDIR (cf. chapitre 5).

Données optiques

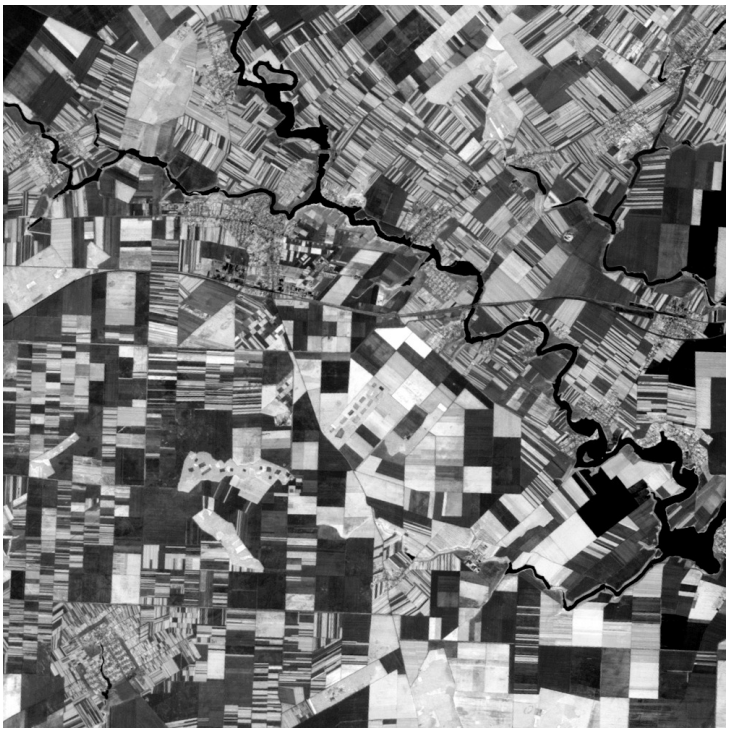
Données et prétraitement des données Les données ici utilisées nous ont été fournies par le CNES (Centre National des Études Spatiales) au travers du projet *Assimilation de Données pour l'Agro-Modélisation (ADAM)* [CNE12]. Nous avons sélectionné

20 images de la STIS ADAM acquises entre octobre 2000 et juillet 2001. De cette façon, suffisamment de données sont prises en compte pour observer des cycles de cultures, des labours et semis d'automne aux récoltes. Ces images ont été acquises par les satellites SPOT. Trois bandes sont disponibles : B1 en vert ($500\ nm - 590\ nm$), B2 en rouge ($R,\ 610\ nm - 680\ nm$ et B3 en poche infra-rouge ($PIR,\ 780\ nm - 890\ nm$). La résolution spatiale est $20\ m \times 20\ m$ et la scène observée est une zone rurale à l'est de Bucarest en Roumanie. Nous avons choisi une sous-scène de 1000×1000 pixels représentant une zone appelée Fundulea. L'intérêt de cette sélection est d'accéder à la vérité terrain, disponible sur la période 2000-2001, concernant les champs appartenant à l'Institut National Roumain de Recherche en Agriculture et de Développement [5]. Ces données d'interprétation représentent 5.9% de la scène et peuvent être utilisées à des fins de validation. Le jeu de données correspondant à cette sous-scène contient du bruit (principalement des perturbations atmosphériques) et a une taille typique dans le domaine de l'analyse pixellaire de STIS (20 images de 1000×1000 pixels). La sous-scène contient principalement des champs dont les dimensions sont supérieures à la résolution spatiale. Différents types de cultures telles que le coton, le maïs, le soja, les petits pois, ou bien encore le millet sont présentes. Les autres objets présents peuvent être classés en « routes », « rivières », « forêts » et « villes ». La topographie de cette région est généralement plate. Seule une petite fraction est affectée par des pentes bordant une rivière ou correspondant à quelques micro-dépressions.

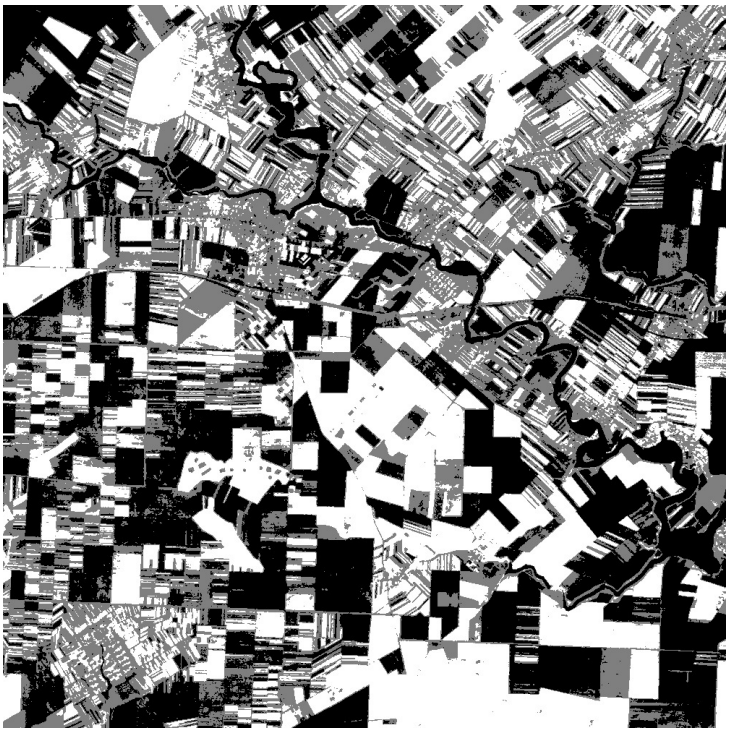
Pour chaque pixel et chaque date, nous avons calculé une bande synthétiques B4 correspondant à l'*Indice de Différence de Végétation Normalisé (IDVN ou NDVI)*[LK00] et qui est défini ainsi : $NDVI = \frac{B3-B2}{B3+B2}$. Le NDVI est très largement utilisé pour détecter la végétation dans les images multispectrales. Un exemple d'une image originale de la STIS ADAM encodé en NDVI est présenté par la figure 9.5a. La quantification des valeurs NDVI a été effectuée en considérant 3 intervalles définis à partir des $33^{\text{ème}}$ et $66^{\text{ème}}$ centiles. De façon à minimiser l'influence des défauts de calibration, la quantification a été envisagée séparément pour chaque image. Pour une date d'acquisition donnée, un pixel est ainsi décrit par un unique label qui indique à quel intervalle sa valeur NDVI appartient. Le label « 1 » donne les valeurs basses, le label « 2 » représente les valeurs intermédiaires et le label « 3 » indique les valeurs hautes. Comme cela est rapporté dans [JMB+11], il s'agit de la pire configuration qui soit par rapport au nombre de motifs extraits par la suite. En effet, si moins ou plus de symboles sont utilisés, alors le nombre de motifs extraits est inférieur à celui obtenu pour 3 symboles. Le résultat de la quantification de l'image de la figure 9.5a est donné par la figure 9.5b. Au final, nous obtenons donc une base de séquence d'un million de séquences de longueur 20 (20 dates soit 20 symboles) décrites avec un alphabet de 3 symboles.

Analyse quantitative Les deux paramètres que peut régler l'utilisateur sont σ, le support minimum et κ, la connexité moyenne minimum. Les valeurs du support minimum ont été prises dans l'intervalle $[0.25\%, 2\%]$ de façon à extraire des aires couvrant au minimum de 2500 pixels ($1\ km^2$) à 20000 pixels ($8\ km^2$). Ces valeurs permettent de considérer tous les champs y compris les plus petits. De façon à caractériser l'impact de κ, les valeurs entre 0 et 7 ont été considérées. En effet, puisque la définition de la mesure de la connexité moyenne s'appuie sur la convention du 8-voisinage, qui ne fait aucune distinction entre les

5. Nous remercions R. Vintila, de l'Institut National Roumain de Recherche en Sciences du Sol et en Agrochimie, et G. Petcu, de l'Institut National Roumain de Recherche en Agriculture et Développement, pour nous avoir fourni la vérité terrain.



(a) Image en NDVI.



(b) Image quantifiée en 3 niveaux basés sur les $33^{\text{ème}}$ et $66^{\text{ème}}$ centiles.

FIGURE 9.5 – Acquisitions SPOT sur la zone de Fundulea, Roumanie, après calcul du NDVI (a) et quantification des valeurs NDVI (b).

pixels des bords de l'image et les autres, la mesure de connexité appartient à l'intervalle $[0, 8[$.

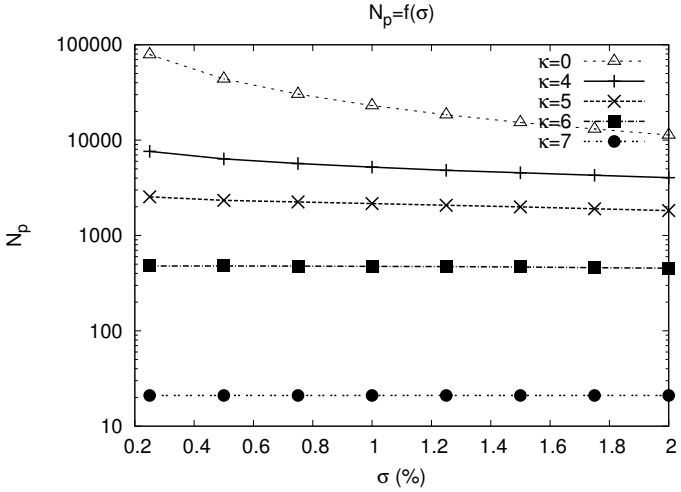
Nos expériences montrent que le nombre de motifs séquentiels fréquents qui sont rejetés grâce à la contrainte de connexité moyenne minimum est important, et que pousser partiellement cette contrainte permet de réduire significativement les temps d'éxécution (de 10 à 20%). Le nombre de motifs extraits N_p peut être réduit de plusieurs ordres de grandeur par rapport au nombre de motifs séquentiels fréquents. Cela est représenté par la figure 9.6a. Si aucune contrainte de connexité moyenne minimum n'est appliquée ($\kappa = 0$), alors tous les motifs séquentiels fréquents sont extraits et N_p atteint jusqu'à 78885 motifs. Comme attendu, N_p est d'autant plus réduit que κ est élevé. Dans le cas le plus défavorable, c'est-à-dire lorsque $\sigma = 0.25\%$, si $\kappa = 4$, alors $N_p = 7623$ tandis que si $\kappa = 7$ alors $N_p = 21$. La contrainte de connexité moyenne est une contrainte très sélective : pour une valeur donnée de κ telle que $\kappa \neq 0$, N_p ne subit que des variations assez limitées en fonction de σ. Par exemple, pour $\kappa = 4$, N_p augmente de 4042 ($\sigma = 2\%$) à 7623 motifs SFG ($\sigma = 0.25\%$) alors que pour $\kappa = 6$, N_p augmente de 454 ($\sigma = 2\%$) à 479 motifs SFG ($\sigma = 0.25\%$). N_p est même stable pour $\kappa = 7$ avec 21 motifs SFG. Comme présenté dans la figure 9.6b, les temps d'extraction sont les mêmes pour toutes les valeurs de κ si la contrainte de connexité n'est pas poussée. Si la contrainte est poussée, alors les temps d'extraction sont réduits, quelle que soit la configuration, de 10% à 20%. Par exemple, pour $\sigma = 0.75\%$ et $\kappa = 7$, l'extraction prend 756 secondes sans push partiel tandis qu'elle ne prend que 599 secondes avec le push partiel.

L'élagage correspondant peut être quantifié via $N_{checked}$, le nombre de motifs séquentiels fréquents qui sont considérés pendant l'extraction et pour lesquels la contrainte de connexité est vérifiée. Les valeurs obtenues pour $N_{checked}$ sont données par la figure 9.7a. Si aucun push partiel n'est mis en œuvre, alors, par exemple, $N_{checked}$ atteint 78885 motifs pour $\sigma = 0.25$ (quelle que soit la valeur κ). Au même seuil de support minimum, si le push partiel est activé, alors, par exemple, avec $\kappa = 7$, $N_{checked}$ est réduit à 50227. Quelle que soit la configuration au niveau de σ et de κ, lorsque la contrainte est poussée, $N_{checked}$ est réduit. Cette réduction (en pourcentage) est mise en évidence par la figure 9.7bb. Elle varie entre 7.7% ($\sigma = 2\%$,$\kappa = 4$) et 36.3% ($\sigma = 0.25\%$,$\kappa = 7$). L'élagage est plus efficace dans les configurations les plus difficiles, c'est-à-dire pour les valeurs faibles de σ.

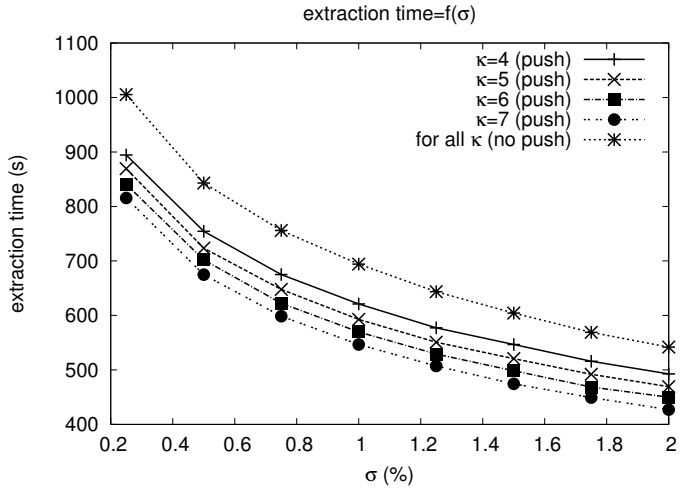
Analyse qualitative Pour ce qui est de l'analyse qualitative, σ a été fixé à 1% de façon à extraire des motifs SFG couvrant au moins 4 km^2 (l'image entière couvre 400 km^2). De cette façon, le focus a été mis sur les cultures les plus présentes dans cette zone, ce qui nous a aidé à caractériser les résultats. La vérité terrain qui nous a été fournie par les experts et qui couvre 5.9% de l'image contient en effet des cultures représentatives de la région.

Comme montré par la suite, en appliquant une contrainte typique de maximalité sur ces motifs il est possible de se concentrer sur un nombre réduit de motifs SFG porteurs de sens pour les experts en agro-modélisation. La contrainte de maximalité est très simple et consiste en la sélection des motifs extraits n'ayant pas de sur-motifs également présents en sortie de l'extraction. Ces motifs, sont en un sens, les plus spécifiques.

Afin de visualiser les résultats, nous avons utilisé le type de localisation spatiale déjà utilisé dans la section 9.2. Puisque nous avons obtenu seulement quelques dizaines d'images de ce type, l'inspection visuelle a pu être effectuée rapidement par l'expert. Il est à noter que si nous n'avions pas appliqué de contrainte de connexité moyenne, c'est-à-dire si nous

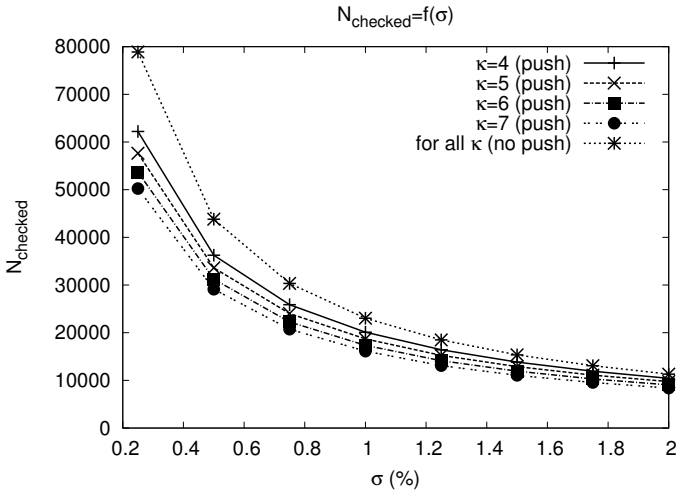


(a) N_p, le nombre de motifs extraits vs. κ, la connexité moyenne minimum et σ, le support minimum.

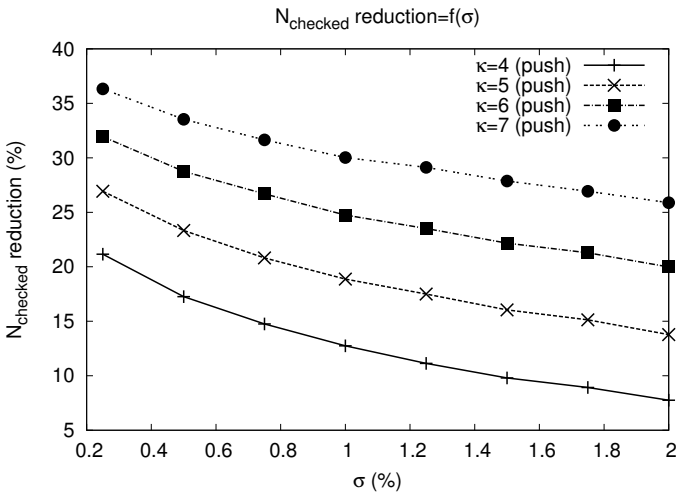


(b) Temps d'extraction (avec et sans push partiel) vs. κ, la connexité moyenne minimum et σ, le support minimum.

FIGURE 9.6 – Impact de la contrainte de connexité moyenne minimum.



(a) $N_{checked}$, le nombre de motifs vérifiés vs. κ, la connexité moyenne minimum et σ, le support minimum.



(b) Réduction du nombre de motifs vérifiés vs. κ, la connexité moyenne minimum et σ, le support minimum.

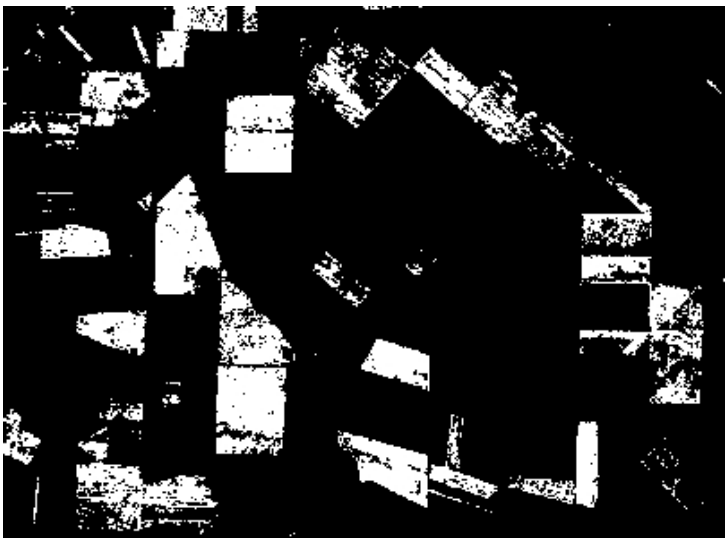
FIGURE 9.7 – Évaluation de l'élagage dû au push partiel.

n'avions considéré que les motifs séquentiels fréquents et non les motifs SFG, alors l'expert n'aurait pas pu analyser nos résultats. En effet, à $\sigma = 1\%$, 23038 motifs sont extraits parmi lesquels 4684 sont des motifs maximaux, ce qui constitue une collection bien trop importante. Un autre fait intéressant à relever est que la quantification et la présence du bruit intrinsèque à une STIS (en particulier les variations atmosphériques et les nuages) ne semblent pas constituer un problème critique En effet, bien que la quantification en 3 niveaux aboutisse à l'extraction de motifs construits sur un alphabet de 3 labels, et bien qu'aucun prétraitement destiné à atténuer le bruit ne soit effectué, l'usage conjoint de l'information spatiale et temporelle permet de trouver des motifs porteurs de sens. Cette technique est donc applicable à des séries d'images de faible qualité obtenues par exemple avec des capteurs aux performances limitées et ne requiert que très peu de prétraitements.

Pour la première expérience, nous avons réglé κ à 7, ce qui est un seuil très sélectif. Dans ce cas, 21 motifs SFG ont été obtenus. Ils expriment des évolutions générales puisque leur longueur n'excède pas 12 symboles. Seulement 7 sont maximaux, et parmi eux, nous avons par exemple le motif $3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$. Les pixels couverts par ce motif sont présentés par la figure 9.8a au niveau de la zone pour laquelle la vérité terrain est disponible. Il couvre 96.2% des pixels de la vérité terrain qui correspondent aux champs cultivés et 98.3% des pixels couverts dans cette zone correspondent à des champs cultivés.

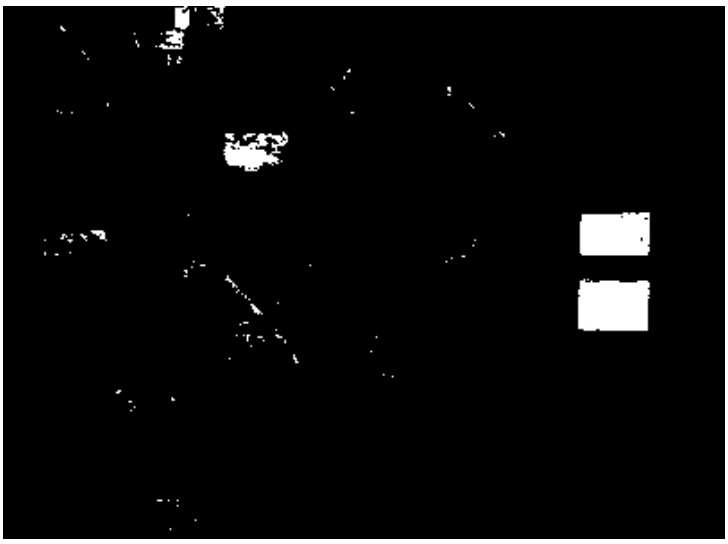
De façon à obtenir des évolutions plus spécifiques, c'est-à-dire des motifs plus longs, κ a été réglé à une valeur moins sélective, $\kappa = 6$. Nous avons obtenu alors 31 motifs SFG maximaux à partir des 474 motifs SFG extraits. L'un d'entre eux est $2 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1$. Les pixels couverts par ce motif sont présentés dans la figure 9.8b. Selon la vérité terrain, il couvre 61.4% des pixels associés à la culture du coton et 96.3% des pixels qu'il couvre correspondent à cette même culture. Des informations intéressantes peuvent être tirées de tels motifs. Par exemple, comme cela peut être observé, des « trous » apparaissent à l'intérieur des champs (polygones presque complètement remplis en blanc) dans la figure 9.8a et la figure 9.8b. Les pixels de ces trous ne sont pas couverts par le motif couvrant les zones blanches. Le comportement temporel est donc différent de celui des pixels qui entourent ces trous bien qu'ils soient tous associés à une même culture. Certains de ces trous correspondent à des différences pédologiques (composition du sol) qui ont été rapportées par les experts tandis que les autres trous sont certainement dus à des différences en termes de fertilisation (engrais) ou d'irrigation. De telles informations sont particulèrement intéressantes et peuvent être utilisées localement pour adapter les pratiques agricoles. De plus, il est possible d'extraire des motifs correspondants à une unique variété d'un type de culture donné. Par exemple, à $\kappa = 5.5$, nous avons 1074 motifs SFG et 66 d'entre sont maximaux. Parmi eux, nous avons le motif SFG $3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1$. La figure 9.9a donne sa localisation. Alors que le motif précédent correspond à un type de culture en général, celui-ci identifie une variété en particulier. En effet, 98.8% des pixels qu'il couvre dans la vérité terrain sont tous associés à une unique variété de coton. Deux champs rectangulaires sont clairement identifiés (partie droite de l'image). Celui d'en haut correspond à une zone partiellement couverte par le motif précédent (cf. figure 9.8b). Ce n'est pas le cas pour l'autre rectangle qui met ainsi en avant un autre champ de coton. Les deux rectangles sont couverts par le motif général correspondant aux champs cultivés (cf. figure 9.8a). Les pixels couverts par les motifs SFG ne correspondent pas toujours à des zones cultivées. Par exemple, pour $\kappa = 6$ nous obtenons aussi le motif SFG $2 \rightarrow 2 \rightarrow 2 \rightarrow 2 \rightarrow 2 \rightarrow 2 \rightarrow 2 \rightarrow 2$ qui correspond aux chemins, aux friches, aux villes et aux bords des champs. Sa localisation est représentée par la figure 9.9b.

(a) $3 \to 3 \to 3 \to 3 \to 3 \to 3$: champs cultivés.



(b) $2 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 1 \to 1 \to 1 \to 1$: coton.

FIGURE 9.8 – Localisation (pixels blancs) de motifs SFG ayant trait aux cultures.



(a) $3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 1 \to 1 \to 1 \to 1$: une espèce particulière de coton.

(b) $2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2$: chemins, friches, villes, bord des champs.

FIGURE 9.9 – Localisation (pixels blancs) de motifs SFG : autres exemples.

Données radar

De façon à investiger la généricité de l'approche, nous avons effectué d'autres expériences sur un type très différent de STIS, une SITS *SAR (Synthetic Aperture Radar ou Radar à Ssynthèse d'Ouverture, RSO)*. Les satellites SAR émettent et enregistrent des ondes radars qui sont réfléchies par la surface de la Terre. Plus précisément, la technique SAR permet de former une antenne synthétique pouvant atteindre plusieurs kilomètres de long alors que l'antenne réelle ne mesure qu'une dizaine de mètres de long. Dans les images radar, les valeurs des pixels sont des complexes. La magnitude (ou module) exprime la quantité d'énergie renvoyée par la surface de la Terre. La phase (ou angle) mesure la propagation des ondes radar, c'est-à-dire la distance entre le satellite et la surface de la terre. Elle inclut également un terme dit de diffusion (scattering) relatif à la nature de la surface réfléchissante. De façon à utiliser l'information géométrique apportée par la phase, il est nécessaire de calculer la différence de phase entre deux images SAR, acquises à des dates différentes, et ce sous l'hypothèse que le terme de diffusion soit stable : le changement dû à l'évolution temporelle et aux angles de visées légèrement différents doit être limité. Dans ce cas, l'image de différence de phase, appelée *interférogramme*, mesure une différence de distance qui est associée à la topographie et aux déplacements/déformations du sol entre deux acquisitions. En utilisant des *Modèles Numériques d'Élévation* (MNE ou DEM en anglais), il est possible de retirer la composante topographique et d'obtenir des mesures de déplacement avec une précision de l'ordre de la fraction de la longeur d'onde utilisée (5.6 cm dans notre cas). La principale limitation de cette technique, appelée *interférométrie SAR différentielle* (D-InSAR dans la communauté) vient des variations des conditions atmosphériques qui peuvent modifier la propagation des ondes et introduire des artefacts difficiles à séparer de l'information de déplacement. La contribution due à l'atmosphère stratifiée peut être estimée avec l'utilisation de DEM et de données météorologiques, mais les effets de l'atmosphère turbulente continuent de dégrader les interferogrammes. Différentes approches ont été développées pour réduire ces difficultés en utilisant les séries temporelles d'interférogrammes : la *technique des permanent scatterers (PS)* [FPR01] qui analyse le signal temporal sur des cibles spécifiques, la *stratégie des Small BAseline Subsets (SBAS)* [BFLS02] qui sélectionne les paires d'images acquises avec les orbites les plus proches possibles (à la fois dans le temps et dans l'espace), ou bien encore la méthode STAMPS [Hoo08] qui incorpore les deux précédentes approches. Elles sont utilisées par les géophysiciens pour surveiller les déformations de surface et leurs évolutions spatiales et temporelles. L'utilité de ce type d'étude concerne la planification de l'utilisation des sols (développement d'infrastructures pour l'agriculture ou le transport).

Données et prétraitement des données Dans cette expérience, une STIS SAR, fournie par le satellite ENVISAT (ENVIronmental SATellite) de l'agence spatiale européenne (ou *European Space Agency, ESA*) [6], a été sélectionnée. Elle contient 25 images acquises entre 2004 et2009. Elle couvre la faille sismique d'Haiyuan, à la frontière nord-est du plateau tibétain. Cette zone a été affectée par des tremblements de terre majeurs au début du XX$^{\text{ème}}$ siècle. Les experts souhaitent localiser et mesurer de possibles déformations de la croûte terrestre au travers des acquisitions qui sont affectées par des perturbations atmosphériques. La STIS sélectionnée a été prétraitée par Cécile Lasserre (laboratoire de Géologie de l'ENS, aujourd'hui au laboratoire ISTerre) et Romain Jolivet (laboratoire ISTerre aujourd'hui au Tectonic Observatory du California Institute of Tech-

6. Nous remercions l'ESA pour la fourniture de ces données au travers du projet Dragon 5305.

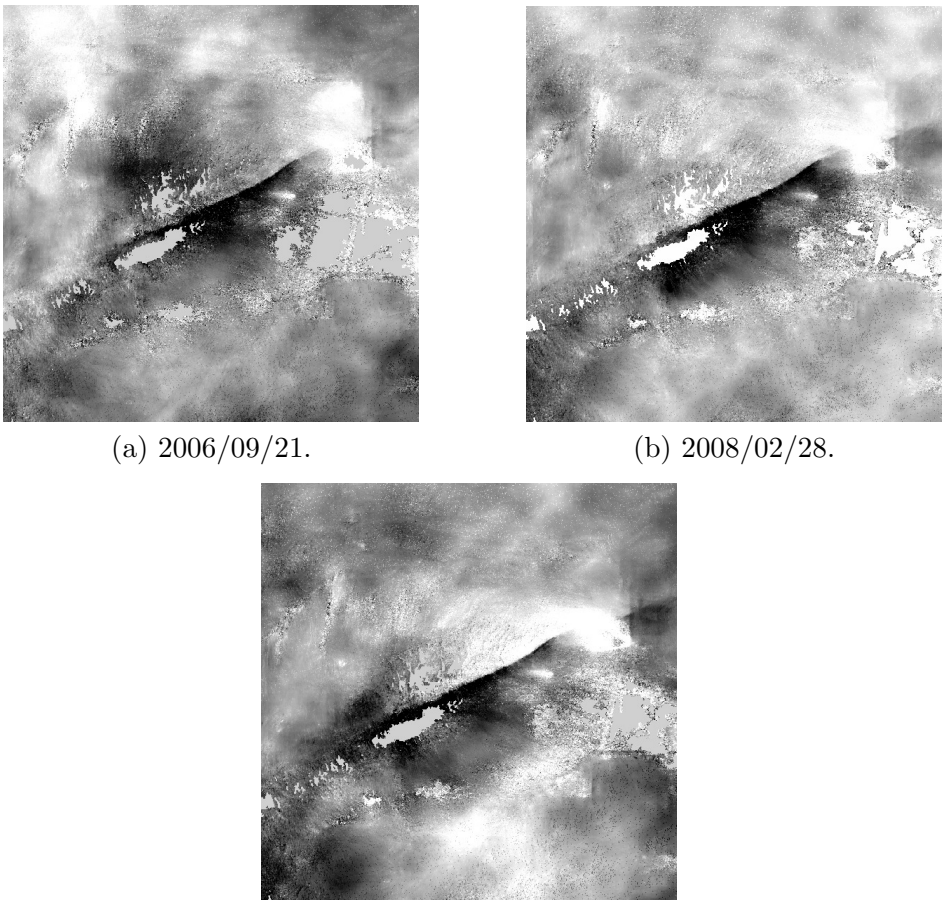
nology) pour calculer des mesures de déplacement sans introduire aucune connaissance spécifique sur la déformation étudiée. Tout d'abord, les interférogrammes ont été générés. Les orbites résiduelles ont été corrigées et les délais atmosphériques ont été retirés en utilisant des DEM et des données météorologiques. Puis, en utilisant une technique de type SBAS, l'évolution de phase entre la première acquisition et les différentes acquisitions successives a été calculée. Au final, ce sont donc 24 images 701X701 [7] contenant les évolutions de phase (déplacements) par rapport à la première acquisition qui ont été produites. Puisque aucun lissage n'est appliqué, à la fois la déformation du sol et l'atmopshère turbulente contribuent à la différence de phase. Le détail des prétraitements permettant la constitution de cette STIS D-InSAR sont disponibles dans [JLD+12]. La résolution spatiale, après traitement, est ici de 80 $m \times 80$ m par pixel et diffère de la précédente STIS. Des images typiques sont présentées par la figure 9.10. Les niveaux de gris correspondent aux déplacements de la surface de la Terre par rapport au satellite selon la direction de visée. Les pixels sombres correspondent aux fortes valeurs négatives exprimant le rapprochement. Les pixels clairs indiquant de fortes valeurs positives représentent l'éloignement. Les autres pixels traduisent de faible déplacements, La STIS a été quantifiée avec trois symboles à partir des $33^{\text{ème}}$ et $66^{\text{ème}}$ centiles. Le symbole $\ll 1 \gg$ représente les fortes valeurs négatives, le symbole $\ll 2 \gg$ correspond aux faibles valeurs négatives et le symbole $\ll 3 \gg$ correspond aux valeurs positives. Le résultat est une base de séquences contenant 491401 séquences formées de 24 symboles chacune.

Analyse qualitative Cette série a été fouillée de façon à découvrir des structures spatio-temporelles pertinentes. L'objectif est d'extraire des motifs SFG et les présenter aux utilisateurs pour attirer leur attention sur des déplacements ou des évolutions de déplacements inconnus. Les interprétations suivantes ont été validées en travaillant avec Marie-Pierre Doin (laboratoire de Géologie de l'ENS, aujourd'hui au laboratoire ISTerre), Cécile Lasserre (laboratoire de Géologie de l'ENS, aujourd'hui au laboratoire ISTerre), Romain Jolivet (laboratoire ISTerre aujourd'hui au Tectonic Observatory du California Institute of Technology), et Catherine Pothier (laboratoire LGCIE de l'INSA Lyon). L'extraction a été effectuée en réglant la surface minimum σ à 4.07% (c'est-à-dire 20000 pixels) et la connexité moyenne minimum κ à 6. Au final 3414 motifs SFG ont été extraits. De façon à se concentrer sur les motifs les plus spéficiques, les motifs maximaux ayant au moins 10 symboles ont été sélectionnnés soit 19 motifs au total. Pour chacun de ces motifs, une image a été construite de façon à observer quelles sont les zones affectées par l'évolution décrite par le motif. Celles-ci sont présentées dans la figure 9.11, pour les trois motifs SFG suivants :

- $\#1 : 2 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1$;
- $\#2 : 2 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 3$;
- $\#3 : 3 \to 3 \to 3 \to 3 \to 3 \to 1 \to 1 \to 3 \to 2 \to 3 \to 3 \to 2 \to 1$.

Le motif SFG #1 indique que certaines zones s'approchent du satellite alors que le motif SFG #2 exhibe des zones s'éloignant du satellite. Comme cela peut être observé dans la figure 9.11, ces motifs sont spatialement complémentaires. Un phénomène de fluage est ainsi révélé par les deux premiers motifs. Il est cohérent avec le mouvement de la partie supérieure de l'image qui a été rapporté par les experts. De plus, la localisation de la faille sismique responsable de ce phénomène de fluage peut être inférée en regardant la frontière nette entre les pixels affectés et les pixels non affectés, particulièrement sur la

7. La résolution de ces images est inférieure aux acquisitions SAR, une réduction d'échelle ayant été appliquée pour réduire le bruit.



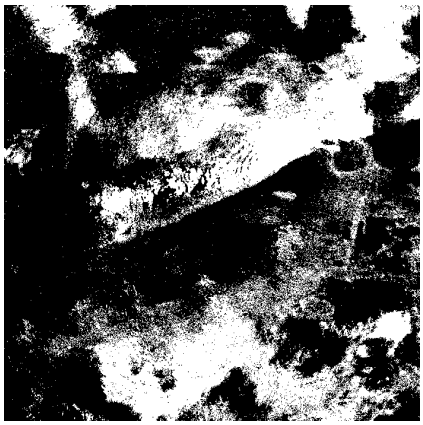
(a) 2006/09/21. (b) 2008/02/28.

(c) 2009/08/06.

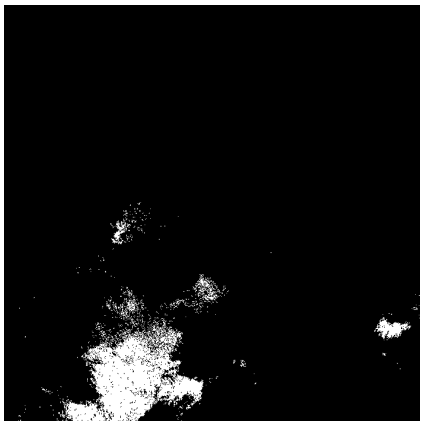
FIGURE 9.10 – Images de déplacement (selon la ligne de visée du satellite radar). Pixels tirant vers le blanc : déplacements positifs importants (éloignement). Pixels tirant vers le noir : déplacement négatifs importants (rapprochement). Autres pixels : faibles déplacements.

diagonale partant du coin inférieur gauche et allant au coin supérieur droit de l'image. Le troisième motif, le motif SFG #3, indique que la surface de la Terre s'éloigne, se rapproche, s'éloigne et puis se rapproche du satellite. Ceci est un comportement atypique. De plus, la zone concernée est très isolée et très compacte comme le montre la figure 9.11 (c). Par conséquent, les ouvrages de génie civil construits dans cette zone devraient être conçus pour résister à des mouvements antagonistes. La localisation spatiale de ces trois motifs permet de discriminer des zones qui ne sont pas facilement dinstinguables dans les images de la SITS D-InSAR originelle (cf. figure 9.10).

(a) Motif SFG #1 : rapprochement.



(b) Motif SFG #2 : éloignement.



(c) Motif SFG #3 : éloignement, rapprochement, éloignement et finalement rapprochement.

FIGURE 9.11 – Localisation (pixels blancs) de 3 motifs SFG exprimant des déplacements par rapport au satellite.

Synthèse

Quel que soit le type ou la résolution des STIS, les expériences ici présentées et rapportées dans [JMR+10, JMB+11, JMR+11, JMR+12] montrent qu'il est possible d'extraire des motifs SFG ayant un sens pour les utilisateurs finaux, qu'ils expriment des phénomènes de changement ou de non-changement. D'autres expériences ont été menées sur des données radar et des données polarimétriques. Les données polarimétriques sont des données radar multivariées formées de 3 canaux différents de polarisation en émission et réception qui apportent une information supplémentaire sur le mécanisme de rétrodiffusion. Ces expériences confirment le caractère générique de l'approche et sont publiées dans [JMT+], [JLM+11] et [JMB+11]. Par ailleurs, quelle que soit l'application considérée, une fois choisie la bande d'intérêt (par exemple NDVI pour les cultures ou les différences de phase pour les déplacements), un simple prétraitement à base de centiles suffit pour construire les STIS symboliques desquelles seront extraits ces motifs. L'ensemble de ces

travaux sont repris et détaillés dans [Jul11].

La contrainte de connexité moyenne minimum sur laquelle s'appuie ces motifs permet de réduire efficacement l'espace de recherche et vient en appui de la contrainte de support/surface minimum lorsque les valeurs de cette dernière sont faibles. La conjonction des contraintes de connexité et de surface permet ainsi de limiter également le nombre de motifs extraits. Cependant, ce nombre peut, selon les jeux de données, rester élevé et nécessite alors la mise en œuvre d'un critère de maximalité permettant de se concentrer sur les motifs les plus spécifiques. Ce type de contrainte associée aux contraintes de connexité et de surface a l'avantage d'écarter, autant que possible, les phénomènes dus à l'incertitude aléatoire, en particulier les phénomènes liés aux perturbations atmosphériques et aux défauts d'acquisition. En effet, ces phénomènes sont dispersés plus ou moins aléatoirement dans l'espace et dans le temps. S'ils affectent des structures spatio-temporelles nombreuses, aux comportements variés, ceux-ci se retrouvent portés par des motifs raffinant ces structures (par ajout d'un/de symbole/s aux motifs caractérisant les structures en question) et dont l'emprise spatiale est encore plus fragmentée et réduite. Ces motifs ont de fait du mal à satisfaire la contrainte de surface minimum et encore plus la contrainte de connexité minimum. Cela est d'autant plus vrai que les motifs sont spécifiques. Au final, ces phénomènes, de par leur dispersion spatio-temporelle, se voient *dilués* dans des motifs qui sont eux-mêmes souvent rejetés lors de l'extraction et de l'application des différentes contraintes. Ainsi, lors de nos différentes expériences, aucun des motifs interprétés par les utilisateurs finaux ne s'est avéré traduire une phénomène lié à des perturbations atmosphériques ou à des défauts de capteur. Il est à noter que les motifs extraits de la STIS Meteosat-7 présentée dans la section 9.2.3 ne rentrent pas dans ce cadre puisque les seules évolutions perceptibles, à cette échelle temporelle et spatiale, sont les passages de nuages à la surface de la Terre.

Au final, les motifs SFG satisfont une grande partie des objectifs mentionnés à la fin de la section 9.1 En effet :

- ces motifs sont d'une lecture aisée par l'utilisateur final,
- la définition des occurrences des motifs séquentiels n'impose ni de considérer des dates immédiatement successives ni de considérer des occurrences synchrones d'un pixel à l'autre ; ce qui permet, pour certains des motifs, de ne pas considérer les défauts liés aux perturbations atmosphériques et/ou aux capteurs,
- aucune contrainte temporelle n'est posée,
- seules 2 contraintes spatiales sont posées et sont directement traduisibles en surface minimum et connexité moyenne minium,
- aucun à priori sur la forme des phénomènes observés n'est posé,
- un large éventail d'échelles d'analyse est accessible : tous les phénomènes dont la surface est supérieure au seuil de support minimum sont pris en compte,
- l'ensemble des contraintes précitées associé à une contraite de maximalité permettent de rejeter une grande partie des phénomènes dus à l'incertitude aléatoire (phénomènes atmosphériques, défauts de capteurs).

9.3.3 Classement des motifs SFG à l'aide de l'Information Mutuelle Normalisée (IMN)

Comme cela a été rapporté dans la section 9.3, l'association des contraintes de surface minimum, de connexité moyenne minimum et de maximalité permet de produire un

nombre de motifs raisonnable. Cependant, rien ne permet de guider l'utilisateur, dans la collection des motifs extraits, vers des motifs dont les occurrences sont singulières au niveau statistique. Plus précisément, quelle assurance avons-nous que la présence d'un motif n'est pas simplement due à une sur-représentation de certains symboles ? Afin de répondre à cette question, considérons les occurrences d'un motif. Celles-ci peuvent être localisées spatialement, comme cela a été régulièrement fait tout au long de ce chapitre, en attribuant la « couleur » noir à tous les pixels qui ne sont pas couverts par le motif en question. La localisation temporelle est alors faite en affectant différentes couleurs aux pixels couverts par le motif, ces couleurs traduisant :

- soit les dates d'occurrence de chacun des éléments de l'occurrence du motif comme cela est proposé dans la section 9.2 (cf. figure 9.4),
- soit la date de début de l'occurrence,
- soit la durée de l'occurrence,
- soit la date de fin de l'occurrence.

Le résultat de ce type de localisation spatio-temporelle est appelé *carte de Localisation Spatio-Temporelle ou carte LST*. Soit α un motif SFG extrait d'un STIS symbolique S. De façon à évaluer si α est significatif, c'est-à-dire qu'il serait difficilement présent dans un jeu aléatoire ayant les mêmes fréquences de symboles, nous proposons de comparer une carte LST de α sur S à sa carte LST obtenue pour une STIS symbolique aléatoire S^*. La STIS S^* est générée à partir de S en échangeant aléatoirement des états pixellaires selon la procédure suivante : soit $B_1, B_2, \ldots, B_N$ et $B'_1, B'_2, \ldots, B'_N$ deux séquences d'évolution pixellaire choisies aléatoirement dans S. Les indices i et j sont retenus pour un échange d'états pixellaires si $B_i = B'_j$ et $B_j = B'_i$. L'échange est effectué en intervertissant B_i et B_j et en intervertissant B'_i et B'_j. Tous les pixels, et donc toutes les séquences d'évolution pixellaire, ont la même probabilité d'être choisis et peuvent être choisis plus d'une fois. En conséquence un échange peut être défait par l'échange inverse. Dans un tel jeu randomisé, les mesures de surface et de connexité des motifs SFG extraits de la STIS originales ont tendance à être plus basses puisque les différents échanges successifs ont modifié les séquences d'évolution pixellaire. Par construction, les séquences d'évolution pixellaire formées d'un seul et même symbole sont les seules séquences à ne pas être modifiées et à garder intactes leurs mesures de surface et de connexité. Cette randomisation garantit deux propriétés : les fréquences des symboles ne sont pas modifiées (1) à l'échelle d'une séquence d'évolution pixellaire et (2) à l'échelle de la STIS. Cette technique dite de *swap-randomization* est adaptée de la technique proposée dans [GMMT07] pour randomiser les matrices booléennes. Considérons maintenant l'information temporelle x d'une carte LST comme la réalisation d'une variable aléotoire X et dont l'espace des réalisations Ω contient toutes les valeurs observées. Cette variable peut être comparée à la variable alétoire Y dont les réalisations sont représentées par la carte LST de α extraite de S^*. Afin de quantifier dans quelle mesure X est indépendante de Y ou, autrement dit, de quantifier dans quelle mesure X prédit/détermine les valeurs de Y, nous proposons d'utiliser la mesure d'Information Mutuelle Normalisée (IMN) proposée dans [CT91] :

$$IMN(X;Y) = \frac{\sum_{x,y\in\Omega^2} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}}{min(H(X),H(Y))}, \tag{9.1}$$

$$H(X) = -\sum_{x\in\Omega} P(x)\log P(x), \tag{9.2}$$

où $P(x, y)$ représente la probabilité de co-occurrence de valeurs d'information temporelle, x et y, aux mêmes positions spatiales.

L'avantage d'une telle mesure est qu'elle capture tout type de dépendance, contrairement aux mesures de corrélation qui s'attachent à déterminer, par exemple, une dépendance linéaire (coefficient de corrélation de Pearson). En effet, elle considère les valeurs comme des labels et teste leurs co-occurrences. Plus les variables sont indépendantes, plus le ratio du log du numérateur tendra vers 0, et plus la mesure d'IMN tendra également vers 0 (cf. équation 9.1). La normalisation (valeurs entre 0 et 1) est assurée en divisant par l'entropie minimum (cf. équation 9.2). L'IMN est calculée pour chaque motif SFG qui est alors classé par rapport à cette valeur. Les motifs les plus prometteurs sont ceux dont la valeur tend vers 0, c'est-à-dire ceux dont les occurrences sont difficilement retrouvées dans les données aléatoires. Par construction de notre méthode, plus un motif évoquera un phénomène de non changement, moins ses occurrences pourront être modifiées (répétitions du même symbole) et plus sa valeur d'IMN tendra vers 1. De même, plus les occurrences d'un motif seront dispersées temporellement, plus il sera difficile de retrouver une telle configuration dans un jeu aléatoire et plus son IMN tendra vers 0. Au final, ce classement, si l'on part de la valeur 0, mettra en avant les motifs en fonction de leur capacité avec évoquer des changements et progressant au fil du temps dans l'espace. À l'inverse, les motifs de non changement ou les motifs dont les occurrences ne sont pas dispersées temporellement se retrouveront à l'autre bout du classement. Même si ces derniers phénomènes peuvent aussi être intéressants, leur singularité statistique est moins évidente. Cette proposition de classement a été faite dans [MRG+12].

9.3.4 Expériences sur le classement IMN

Données et prétraitement des données Afin de valider notre approche, l'information mutuelle normalisée a été calculée pour chacun des motifs SFG extrait d'une STIS Landsat 7 (©USGS 1999 - 2010, LDPAAC distribution). Les images de cette STIS (513x513 pixels) couvrent sur le sud-est de la Nouvelle-Calédonie SITS et plus précisément la commune rurale de Yaté. Ce territoire présente d'importantes mines de nickel dont les experts souhaitent s'assurer la progression afin de maîtriser les risques environnementaux associés. En effet, les lagons et la barrière de corail qui bordent ces zones sont déclarés patrimoine mondial de l'UNESCO depuis 2008. La SITS que nous avons manipulée contient 16 images multispectracles acquises entre 2000 et 2001, à une résolution spatiale de 30 $m \times 30$ m et avec une présence significative de nuages sur 13 d'entre elles comme le montre la figure 9.12. Des défauts de capteurs sont également présents. Le détail de deux de ces images de cette STIS est donné par les figures 9.13a et 9.13b. La vérité terrain a été fournie par Bluecham SAS grâce à sa plate-forme de surveillance environnementale *Qehnelo*[8]. Pour ces expériences, nous avons utilisé, comme dans la section 9.3.2, l'Indice de Différence de Végétation Normalisé (IDVN ou NDVI) [LK00], calculé à partir de la bande rouge et de la bande proche infra-rouge[9]. La quantification des valeurs NDVI a été effectuée en considérant 3 intervalles définis à partir des 33ème et 66ème centiles. De façon à minimiser l'influence des défauts de calibration, la quantification a été conduite séparément pour chaque image. Pour une date d'acquisition donnée, un pixel est ainsi décrit par un unique label qui indique à quel intervalle sa valeur NDVI appartient. Le

8. Disponible à http ://www.yate.nc/ grâce à la commune de Yaté.

9. Nous remercions la société Bluecham par la fourniture des images originales et pour le calcul du NDVI.

label « 1 » donne les valeurs basses, le label « 2 » représente les valeurs intermédiaires et le label « 3 » indique les valeurs hautes.

FIGURE 9.12 – STIS Landsat 7 en couleurs RGB, de 2000 à 2011, Nouvelle-Calédonie, commune de Yaté.

Éléments quantitatifs Une extraction de motifs SFG en imposant une surface minimum σ de 6000 et une connexité moyenne minimum κ de 5 a été réalisée. De façon à sélectionner les motifs les plus spécifiques, seul les motifs SFG maximaux ont été retenus. Au final, 295 motifs SFG sont ainsi conservés. Les cartes LST, à la fois sur la STIS originale et sur la STIS randomisée ont été calculées en prenant en compte les dates de fin d'occurrence au plus tôt des motifs comme cela a été proposé dans [MJL+11]. De façon à nous assurer que la STIS a suffisamment été randomisée, nous avons calculé les classements pour 3 niveaux de randomisation : 4 millions, 40 millions et 80 millions de tentatives d'échange d'états pixellaires. Selon nos expériences, 29 des 30 meilleurs motifs sont les mêmes pour 40 millions et 80 millions de tentatives d'échange. Les 30 meilleurs motifs pour 4 millions d'échanges sont très différents de ces 29. Cela indique donc que,

pour ce jeu de données, un niveau suffisant de randomisation est atteint à 40 millions de tentatives d'échanges. Les résultats présentés ci-après ont été calculés pour ce niveau de randomisation.

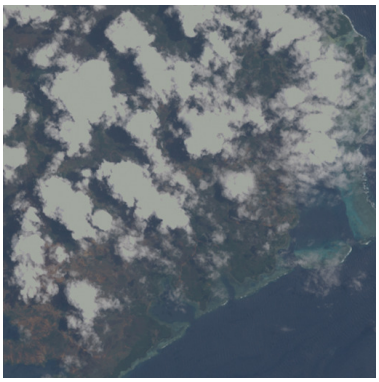
Analyse qualitative Cette analyse a été menée et validée par Catherine Pothier (laboratoire LGCIE de l'INSA Lyon) et Rémi Andréoli (société Bluecham). Nous présentons ici les 4 meilleurs motifs que sont $2, 2, 3, 2, 2, 2, 3$ (IMN=0.037953), $2, 3, 2, 2, 2, 3, 2$ (IMN=0.039520), $2, 3, 2, 2, 3, 2, 2$ (IMN= 0.041644) et $2, 2, 1, 1, 1, 2$ (IMN=0.041737). Les trois premiers motifs alternent des valeurs moyennes ou hautes de NDVI (symboles $\ll 2 \gg$ et $\ll 3 \gg$). Ces trois motifs donnent des cartes LST très similaires. Une d'entre elle est présentée par la figure 9.13c. Les pixels colorés correspondent à du maquis implanté sur un sol pierreux de type ultramafique (roches magmatiques et pré-magmatique pauvres en silice). Les pixels noirs, c'est-à-dire ceux qui ne sont pas affectés par le motif pour lequel a été calculée la carte, correspondent à l'océan (sud-est) et des zones d'activités minières (parties entourées qui seront détaillées plus tard) et à un lac. Une autre partie en noir se situe le long de la côte (indiquée par les croix blanches) et correspond à un type différent de végétation. La carte LST du $4^{\text{ème}}$ motif $2, 2, 1, 1, 1, 2$ est présentée par la figure 9.13d. Elle souligne le contour du lac (cercle en haut en gauche, numéroté 0) et les zones d'activités minières pour lesquelles l'échelle de couleur des cartes permet de retracer l'historique. En effet, pour chaque pixel affecté par ce motif, la couleur indique la date de fin d'occurrence. L'échelle des couleurs est donnée par la figure 9.13f. De façon détaillée, nous avons :

- (1) : une pépinière destinée à la reforestation (bleu clair),
- (2) : la partie la plus récente des mines, une zone d'extraction, avec un gradient allant du bleu au violet représentant l'expansion de cette zone dans le temps,
- (3) : une nouvelle aire d'extraction (3),
- (4) : les bassins de décantation avec un gradient allant du violet sombre au violet, ce qui correspond à une expansion vers le nord ouest dans le temps,
- (5) : les usines de traitement du minerai, avec à nouveau un gradient allant du bleu clair au violet et représentant une utilisation progressive des sols,
- (6) : les logements des équipes de travail.

La cohérence de la méthode est observable avec la figure 9.13e présentant un motif de non changement. Ce motif SFG, $1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$ contient 16 fois le symbole $\ll 1 \gg$ et a la plus haute valeur d'IMN possible ($IMN = 1$) parmi les motifs maximaux extraits. Puisque notre STIS contient 16 images, cela signifie que les pixels couverts par le motif sont décrits par le symbole $\ll 1 \gg$ à chacune des dates. Dans notre méthode de classement, ces pixels ne sont pas considérés comme intéressants car leurs états pixellaires n'évoluent pas sur l'ensemble de la STIS. En effet, les symboles étant les mêmes pour chaque date d'acquisition, les échanges d'états pixellaires ne changent pas les séquences d'évolution pixellaire et l'IMN atteint 1 dans ce cas. Le motif SFG $1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$ correspond à l'océan. Les formes noires dans l'océan correspondent à la présence de nuages dans une ou plusieurs images. Bien que les nuages aient un impact sur des motifs de longueur égale ou proche du nombre d'image de la STIS, ce type de phénomène n'est pas observé pour des motifs plus courts tels que ceux présentés auparavant. En effet, pour les motifs plus courts, l'information cachée par la présence d'un nuage peut être obtenue d'une autre image, à une autre date. Ceci est évidemment plus dur pour les motifs longs, voire impossible pour les motifs dont la taille est égale au nombre total d'images. Ce type de motif n'est pas des plus fréquents. En effet, la valeur moyenne de l'IMN est, en ce qui concerne cette expérience, de 0.1 avec un

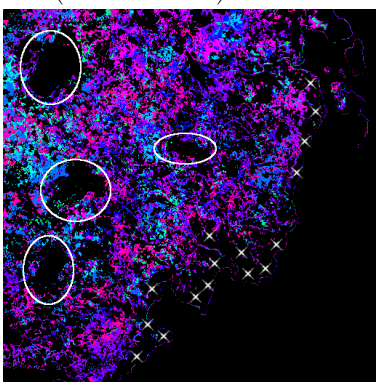
écart-type à 0.07. Il s'agit donc d'un cas plutôt extrême, cas qui par exemple n'est pas retrouvé sur des expériences préliminaires que nous menons actuellement sur des données radar.

Synthèse Classer les motifs SFG par la méthode proposée dans cette section permet de séparer d'un côté les motifs exprimant des changements et progressant au fil du temps dans l'espace et, d'un autre côté, les motifs évoquant le non changement ou des motifs ne progressant pas dans l'espace au fil du temps. Ces derniers sont d'ailleurs considérés comme ayant une plus forte probabilité d'apparaître dans un jeu de données aléatoire où les fréquences des symboles sont conservées. Les deux types de motifs expriment des types de phénomènes différents vers lesquels il est possible de guider automatiquement l'utilisateur. Les expériences menées et publiées dans [MRG+12] montrent le potentiel de l'approche en exhibant des phénomènes spatio-temporels intéressants liés à l'activité minière, et ce tout en rejetant les perturbations atmosphériques. Ces travaux ont été menés en collaboration avec Felicity Lodge, la post-doctorante que j'encadre en tant que responsable scientifique pour le LISTIC de l'ANR FOSTER (cf. chapitre 5).

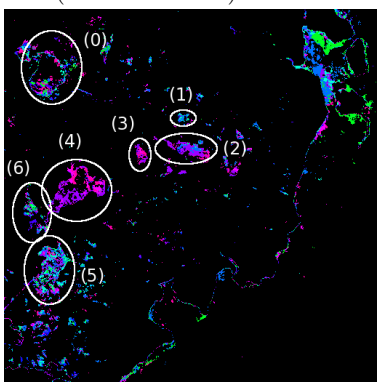


(a) Acquisition Landsat 7 du 2002-10-22 (couleur RGB).

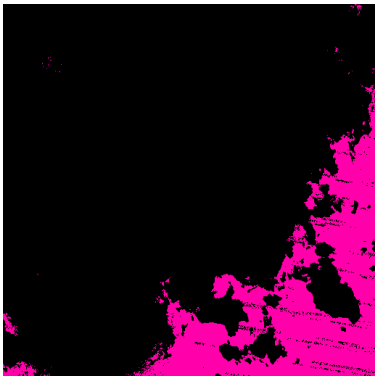
(b) Acquisition Landsat 7 du 2007-04-27 (couleurs RGB).



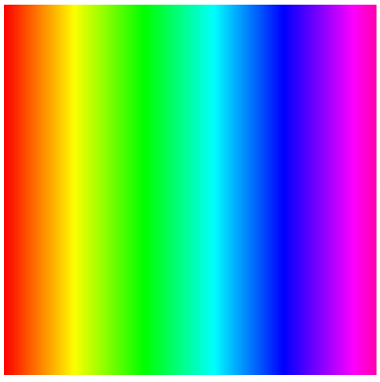
(c) Carte LST du motif SFG $2, 2, 3, 2, 2, 2, 3$.



(d) Carte LST du motif SFG $2, 2, 1, 1, 1, 2$.



(e) Carte LST du motif SFG $1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$.



(f) Échelle de couleurs, du rouge (année 2000) au violet (année 2011).

FIGURE 9.13 – Images Landsat 7 (Nouvelle Calédonie, commune de Yaté), cartes LST et échelle de couleur.

Chapitre 10

Prévision d'événements dans un flot de données

Ce chapitre présente les travaux relatifs à la prévision d'événements dans un flot de données. Après avoir détaillé le contexte thématique, les objectifs associés et les principales contributions existantes dans le section 10.1, présentation est faite de notre étude d'opportunité dans la section 10.2. Cette étude, qui met en avant l'utilisation de *FLM-règles* telles que définies dans [MR04], est reprise et complétée par la section 10.3 avec l'introduction d'une méthode d'apprentissage permettant de sélectionner les FLM-règles les plus génériques générant le moins possible de fausses prévisions et d'une méthode de prévision exploitant au mieux et au plus tôt l'information temporelle des FLM-règles.

10.1 Contexte

Des lignes de production aux composants d'un avion en passant par les réseaux de télécommunications, nombreux sont les systèmes complexes générant un flot de données. Une fois obtenue la description d'un tel flot sur une période considérée comme caractéristique, il est possible d'utiliser cette dernière afin de procéder à de la *prévision d'événements*, c'est-à-dire du *pronostic*. L'objectif poursuivi par la prévision d'événements est l'aide au pilotage de ces systèmes complexes pour lesquels les experts ont du mal à définir des modèles de comportement. Ce pilotage passe le plus souvent par la prévision de pannes, de défaillances. Ces derniers événements, si anticipés, sont en effet source de réduction des coûts [MSI08]. Si l'on s'intéresse aux défaillances d'un système, le pronostic est défini dans [SG07] ou [ISO04] comme la détection des signes précurseurs d'un dysfonctionnement et l'estimation du temps restant avant la défaillance. Bien qu'il s'agisse également d'inférence, cela diffère fondamentalement de la notion prédiction de classe/classification supervisée [TSK05], c'est-à-dire l'inférence d'une variable catégorique d'un objet à partir de l'observation d'autres variables le décrivant sur un laps de temps donné. Il ne s'agit pas non plus de *prédiction au plus tôt* comme définie dans [XPY12]. Dans ce dernier cas, il s'agit de prédire la classe d'un objet dès que possible, au fur et à mesure que les données se présentent. Prévoir un événement, qu'il s'agisse d'une défaillance ou de tout autre type d'événement revient en effet à estimer la date à laquelle ce dernier va apparaître.

La plupart des propositions existantes ont trait à la prévision de défaillances et proviennent de l'industrie aéronautique et aérospatiale (c.f. [LFM99] ou [SG07]) ou de la médecine (cf. [MP01]). Elles s'appuient généralement sur les réseaux neuronaux dont il est dit dans [MP01] qu'ils sont incapables de fournir une explication lisible concernant leurs conclusions. Par exemple, dans [EKGZ08], les auteurs proposent une méthode en deux étapes dédiée à la maintenance d'équipements. D'abord, ils prévoient l'évolution d'un index de dégradation à l'aide de réseaux neuronaux flous adaptatifs. Cette prévision est effectuée pour une date donnée par l'utilisateur et doit être, dans une seconde étape, comparée à un index de dégradation de référence en utilisant un seuil de distance également fixé par l'utilisateur. Outre le fait que la prévision effectuée par un réseau neuronal est difficilement interprétable par l'utilisateur final, cette approche n'a été testée que sur des données synthétiques. Dans [LFM99], Letourneau et al. présentent une approche pour prévoir des défaillances de composants d'avions. À partir de données historiques, ils construisent des classifieurs (des arbres de décision) utilisés par la suite à des fins de prévisions. Les données d'apprentissage, hétérogènes (numériques, textuelles) sont construites en sélectionnant les données précédant les défaillances sur une étendue temporelle fixée par l'utilisateur, généralement 10% de la durée couverte pour l'ensemble des données. Une défaillance est prévue à la volée si les données courantes de production sont classifiées comme « caractérisant une période de défaillance ». La date de défaillance est déduite de l'étendue temporelle fixée par l'utilisateur lors de l'apprentissage.

Une variété de motifs locaux, les *épisodes* tels que définis dans [MTIV97], a montré sa capacité à décrire les flots de données provenant de réseaux de télécommunications [HKM+96], de capteurs sismiques [MR04], de lignes de fabrication [LSU04], de sites web [CG03] ou bien encore de grandes surfaces de vente [AGS04]. Une des classes d'épisodes la plus utilisée est la classe dite des épisodes *sériels*. Par exemple, l'épisode sériel « $A \rightarrow D \rightarrow F$ » indique que l'événement de type « F » suit, immédiatement ou non, l'événement de type « D » qui lui-même survient, immédiatement ou non, après l'événement de type « A ». Un épisode sériel est donc identique, dans sa forme, aux motifs séquentiels tels que définis au chapitre 9. Les *épisodes fréquents* sont sélectionnés grâce à une mesure de fréquence (ou support) et un seuil de fréquence minimum. Contrairement aux motifs séquentiels dont la mesure de fréquence est calculée en nombre de séquences les supportant au sein d'une base de séquence, la fréquence d'un épisode est calculée sur une seule longue séquence d'événements en comptant ses occurrences. Plusieurs définitions d'occurrences ont été avancées : les occurrences basées sur des fenêtres temporelles glissantes [MTV95, HKM+96, MTIV97], les occurrences non imbriquées ou *minimales* [MT96, MT96], les occurrences basées sur des fenêtres glissantes calées sur un événement dont le type est celui du premier événement de l'épisode [IYN04], les occurrences ne se recouvrant pas [LSU05, LSU07], ou bien encore les occurrences non entrelacées [Lax06]. Une vision unifiée de l'extraction de tous ces types d'occurrence est aujourd'hui disponible dans [ALS12]. Sur la base des épisodes fréquents, les *règles d'épisodes* [MTIV97] du type « $A \rightarrow D \rightarrow F \Rightarrow G$ » sont construites. Cette dernière est lue comme « si $A \rightarrow D \rightarrow F$ alors G suit avec une certaine *confiance*, une certaine probabilité ». Cette confiance associée à un seuil de confiance minimum permet de sélectionner les règles dites *confiantes*. Puisque ces règles sont construites sur des épisodes fréquents, celles-ci sont également fréquentes.

Les épisodes et les règles d'épisodes étant des motifs facilement interprétables par

l'utilisateur final, des propositions visant à la prévision d'événements ont été faites. Dans [CZC07], les auteurs proposent une méthode qui recherche dans un flot de données les règles d'épisodes extraites d'un historique. Dès que la prémisse d'une règle d'épisodes est reconnue, elle est utilisée pour prévoir la conclusion. Plus précisément, ils proposent de construire et de maintenir une file d'évenements susceptibles de former les prémisses des règles préalablement extraites. Une fois qu'une prémisse est identifiée dans le flot de données, ils calculent la date à laquelle la conclusion est censée apparaître en ajoutant la taille de la fenêtre temporelle maximale utilisée lors de l'apprentissage (et identique pour toutes les règles) à la date d'occurrence du premier événement de la prémisse. Cette méthode est intéressante car elle évite de balayer tout le flot de données mais une information cruciale, l'information temporelle, c'est-à-dire la taille de fenêtre temporelle maximale d'apprentissage, doit être fournie par les utilisateurs. Par ailleurs, cette méthode n'a été testée que sur des données synthétiques. Une autre technique, à base d'épisodes, publiée dans [LTW08], vise à prévoir, à l'instant n, les types d'événements d'intérêt susceptibles d'apparaître à l'instant $n + 1$. L'ordre est pris en compte mais aucune date de prévision n'est fournie. Pour ce faire, les épisodes fréquents sont à nouveau extraits à l'aide d'une fenêtre temporelle maximale identique pour chacun d'entre eux. Les fenêtres temporelles sont considérées si elles finissent sur un type d'événement d'intérêt. Chaque épisode fréquent découvert est alors associé à un modèle de Markov caché. Ces modèles sont ensuite combinés pour chaque type d'événement d'intérêt. Pour finir, la fenêtre courante est comparée aux différentes combinaisons obtenues, ce qui sert de base à la prévision des types d'événements d'intérêt.

Les méthodes présentées dans cette section imposent à l'utilisateur de fixer une fenêtre temporelle d'apprentissage, fenêtre qui sera utilisée par la suite pour la prévision. Cela paraît peu réaliste dans le cas où les systèmes surveillés et pilotés dépassent de par leur complexité la connaissance de l'expert. Afin de susciter la découverte de connaissances et de ne pas limiter la prévision d'événements à la connaissance de l'expert, qui peut être parcellaire voire fausse, notre travail vise à réduire à sa plus simple expression l'appel à des connaissances concernant les aspects temporels tout en exprimant et en expliquant nos prévisions le plus simplement possible.

10.2 Étude d'opportunité

Cette étude a été menée dans le cadre du stage ingénieur CNAM de Nicolas Le Normand, de septembre 2007 à juillet 2008 (cf. chapitre 4). La spécialité visée a pour intitulé *Management des systèmes d'information*. Les financements utilisés proviennent d'un BQR dédié au pilotage de chaînes logistiques.

10.2.1 Définitions préliminaires

Les flots de données provenant des systèmes complexes peuvent être appréhendés comme des flots d'*événements*, c'est-à-dire des symboles ou *des types d'événements* associés à des dates d'occurrences. Afin de fournir des descriptions explicites à même d'expliquer nos prévisions, nous avons d'emblée choisi d'utiliser des motifs locaux. Deux contextes d'extraction de motifs locaux dans de telles séquences d'événements peuvent

être distingués : les bases de séquences comme définies dans [AS95] et vues au chapitre 9 ou la *longue séquence d'événements* comme définie dans [MTIV97]. C'est dans ce dernier contexte que nous plaçons les travaux présentés dans ce chapitre. En effet, une base de séquence contient de très nombreuses séquences de taille réduite (500 événements au maximum) tandis qu'une longue séquence d'événements peut contenir des centaines de milliers d'événements, ce qui correspond plus à la réalité des flots de données provenant de systèmes complexes pour lesquels la fréquence d'acquisition est élevée en regard de leur durée de vie. Comme évoqué dans la section précédente, plusieurs types d'occurrence sont disponibles. Néanmoins, que ce soit explicite au niveau des définitions ou implicite car utilisée en pratique lors des extractions, une contrainte temporelle de fenêtre maximum doit être fournie par l'utilisateur et est appliquée indifféremment à tous les motifs, quelle que soit leur taille. Afin de ne pas avoir à fixer une telle fenêtre et de considérer différentes tailles de fenêtres en fonction des motifs eux-mêmes, nous avons décidé de nous baser sur la proposition faite dans [MR04]. Celle-ci revient à extraire des règles d'épisodes ayant une *fenêtre temporelle optimale* et satisfaisant à des contraintes de fréquence, de confiance et de gap maximum. La contrainte de gap maximum correspond à une contrainte sur le temps maximum écoulé entre deux événements consécutifs formant l'occurrence d'un épisode. Elle permet de contraindre linéairement la durée des occurrences des règles par rapport au nombre de symboles composant ces dernières. Il n'est ainsi plus nécessaire de fixer une seule fenêtre temporelle maximum pour toutes les règles. La fenêtre temporelle optimale, si elle existe, est la plus petite fenêtre temporelle correspondant à une maximum local de confiance : la confiance est plus basse pour de plus petites ou de plus grandes fenêtres temporelles. Ces règles sont appelées des règles *First Local Maximum* ou *FLM-règles*. Afin de définir plus en détail ce type de règle, commençons par définir les données elles-mêmes :

Définition 11 (événement, séquence d'événements) Soit E un ensemble de symboles appelés *types d'événements*. Un *événement* est défini par le couple (e, t) où $e \in E$ et t est un entier donnant la date d'occurrence de e. Une séquence d'événement est un triplet $S = (s, T_s, T_e)$ où s est une *séquence ordonnées d'événements* $\langle (e_1, t_1), (e_2, t_2), ..., (e_n, t_n) \rangle$ telle que $\forall i \in \{1, ..., n\}$, $e_i \in E \wedge \forall i \in \{1, ..., n-1\}$, $t_i \leq t_{i+1}$. T_s, T_e sont des entiers exprimant la date de début (*starting date*) et la date de fin (*ending date*) de la séquence d'événements.

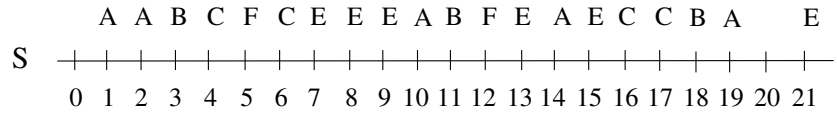


FIGURE 10.1 – La séquence d'événements S.

La figure 10.1 représente un exemple jouet d'une telle séquence. Les règles d'épisodes sont construites à partir d'une type d'épisodes que sont les *épisodes sériels* :

Définition 12 (épisode sériel, préfixe, suffixe) Un épisode sériel est un n-uplet $\alpha = \langle e_1, e_2, ...e_k \rangle$ tel que $\forall i \in \{1, ..., k\}, e_i \in E$. Le préfixe d'$\alpha$, noté $prefix(\alpha)$ est le n-uplet $\langle e_1, e_2, ...e_{k-1} \rangle$. Le suffixe d'$\alpha$, noté $suffix(\alpha)$ est le singleton $\{e_k\}$.

Afin de faciliter la lecture, un épisode sériel $\alpha = \langle e_1, e_2, ...e_k \rangle$ est aussi noté $e_1 \rightarrow e_2 \rightarrow ... \rightarrow e_k$. Étant donné que nous ne considérons que des épisodes sériels [1], nous utiliserons le terme *épisodes* dans la suite de ce mémoire. Par exemple, $A \rightarrow B \rightarrow C$ est un épisode indiquant que B apparaît après A et est suivi de C. Le préfixe de $A \rightarrow B \rightarrow C$ est $A \rightarrow B$ et son suffixe est C. Définissons maintenant la façon dont nous considérons qu'un épisode apparaît dans une séquence :

Définition 13 (occurrence) Un épisode $\alpha = \langle e_1, e_2, ..., e_k \rangle$ apparaît dans une séquence $S = (s, T_s, T_e)$ s'il existe au moins une séquence ordonnée d'événements $s' = \langle (e_1, t_1), (e_2, t_2), ..., (e_k, t_k) \rangle$ telle que s' puisse être obtenue en enlevant des éléments de s ou $s' = s$ (ce qui sera noté $s' \sqsubseteq s$ dans ce mémoire) et en s'assurant que $\forall i \in \{1, ..., k-1\}$, $0 < t_{i+1} - t_i \leq maxgap$ avec $maxgap$ une contrainte définie par l'utilisateur représentant le temps maximum autorisé entre deux événements consécutifs. L'intervalle $[t_1, t_k]$ est une occurrence de α. L'ensemble des occurrences de α dans S est noté $occ(\alpha, S)$.

La contrainte $maxgap$ est utilisée à la fois pour réduire l'espace de recherche et pour satisfaire à des besoins applicatifs récurrents. Elle a été introduite dans [MR04]. Elle contraint linéairement la durée des occurrences des épisodes en fonction du nombre de symboles formant ces derniers et évite d'imposer une même fenêtre temporelle maximale à tous les épisodes. Selon cette définition, en considérant l'exemple de la figure 10.1 et en réglant $maxgap$ à 4 tout au long de cette section, $occ(A \rightarrow B, S) = \{[1, 3], [2, 3], [10, 11], [14, 18]\}$. Les intervalles $[1, 11], [2, 11], [1, 18], [2, 18], [10, 18]$ ne satisfont pas la contrainte de $maxgap$. De façon à réduire le nombre d'occurrences et à ne considérer que celles ne contenant pas d'autres occurrences, nous nous intéressons aux *occurrences minimales* telles que définies dans [MT96, MTIV97] et utilisées dans [MR04] :

Définition 14 (occurrence minimale) Une occurrence minimale d'un épisode α dans une séquence S est un intervalle de temps $[t_s, t_e]$ contenant α et tel qu'il n'y ait pas d'autre occurrence $[t'_s, t'_e]$ vérifiant $[t'_s, t'_e] \subset [t_s, t_e]$. L'ensemble des occurrences minimales de α dans S est noté $mo(\alpha, S)$.

Si l'on reprend notre exemple, les occurrences minimales de $A \rightarrow B$ dans S sont $mo(A \rightarrow B, S) = \{[2, 3], [10, 11], [14, 18]\}$. L'occurrence $[1, 3]$ n'appartient pas à $mo(A \rightarrow B, S)$ car elle inclut l'occurrence $[2, 3]$. Rappelons à toute fin utile que, dans notre cas, et contrairement à la définition originelle proposée dans [MTIV97], notre définition intègre la contrainte de gap maximum définie au niveau des occurrences elles-mêmes. L'algorithme *Minepi* permettant d'extraire les occurrences minimales proposé dans [MTIV97] n'a donc pas été conçu pour gérer cette contrainte. Or, cette dernière est source d'incomplétude. Plus précisément, si les occurrences minimales d'un épisode de taille k (c'est-à-dire ayant k types d'événements/symboles) sont considérés, alors il est possible de calculer les occurrences des épisodes de $k + 1$ dont il est le préfixe. Considérons une séquence S, l'épisode $A \rightarrow B \rightarrow C$ et l'épisode A. Avec $maxgap = 4$ unités temporelles, $mo(A \rightarrow B \rightarrow C, S) = \{[2, 4]\}$ et $mo(A, S) = \{[1, 1], [2, 2], [10, 10], [14, 14], [19, 19]\}$ ne peuvent pas être utilisées pour générer l'occurrence minimale $\{[2, 10]\}$ de l'épisode $A \rightarrow B \rightarrow C \rightarrow A$. En effet, l'occurrence minimale de $(A \rightarrow B \rightarrow C)$ dans S apparaît trop tôt par rapport à la date de fin de $A \rightarrow B \rightarrow C \rightarrow A$. C'est pourquoi l'algorithme *WinMiner* a été proposé dans [MR04]. Il extrait toutes les occurrences minimales des épisodes

1. Une définition plus générale des épisodes est proposée dans [MTIV97].

satisfaisant une contrainte de *maxgap*. Il s'appuie sur la notion d'*occurrence minimale préfixée* :

Définition 15 (occurrence minimale préfixée) Soit $o = [t_s, t_e]$ l'occurrence de l'épisode α dans une séquence S. o est une *occurrence minimale préfixée* de α si est seulement si : $\forall [t_1, t_2] \in mo(prefix(\alpha), S)$, si $t_s < t_1$ alors $t_e < t_2$. L'ensemble des occurrences minimales préfixées de α dans S est noté $mpo(\alpha, S)$.

En utilisant cette définition, $mpo(A \rightarrow B \rightarrow C, S) = \{[2,4],[2,6]\}$. Il est maintenant possible de joindre $[2,6]$ avec $[10,10]$ pour construire $mpo(A \rightarrow B \rightarrow C \rightarrow A, S) = mo(A \rightarrow B \rightarrow C \rightarrow A, S) = \{[2,10]\}$. La notion d'occurrence minimale préfixée sera considérée plus tard dans ce chapitre pour détecter les prémisses des FLM-règles dans les flots de données. Plus de détails sont disponibles dans [MR04].

Les règles d'épisodes sont construites à partir des épisodes. Soit α un épisode. Une *règle d'épisodes* est l'expression $prefix(\alpha) \Rightarrow suffix(\alpha)$. Par exemple si $\alpha = A \rightarrow B \rightarrow C$, la règle d'épisode construite sur α est $A \rightarrow B \Rightarrow C$ [2]. Les règles d'épisodes sont caractérisées à l'aide de deux mesures :

- *le support* : le nombre d'occurrences d'une règles d'épisodes dans la séquence considérée, c'est-à-dire le nombre d'occurrence de l'épisode sur laquelle elle est construite. Par exemple, le support de $A \rightarrow B \Rightarrow C$, noté $support(A \rightarrow B \Rightarrow C)$, est égal au support de l'épisode $A \rightarrow B \rightarrow C$, noté $support(A \rightarrow B \rightarrow C)$.
- *la confiance* : la *probabilité conditionnelle observée* de voir apparaître la conclusion de la règle d'épisodes sachant que la prémisse est déjà apparue. La confiance de $A \rightarrow B \Rightarrow C$ est ainsi définie comme suit :
  $confiance(A \rightarrow B \Rightarrow C) = \frac{support(A \rightarrow B \rightarrow C)}{support(A \rightarrow B)}$.

Ces mesures sont utilisées pour sélectionner les règles d'épisodes fréquentes, c'est-à-dire dont le support est supérieur ou égal à un support minimum σ , et confiantes, c'est-à-dire dont la confiance est supérieure ou égale à une confiance minimum γ. Comme proposé dans [MR04], le support et la confiance peuvent être définis sur la base des occurrences minimales et pour chaque *largeur de fenêtre*, c'est-à-dire la durée maximale des occurrences d'un épisode. Dans l'exemple de la figure 10.1, $mo(A \rightarrow B, S) = \{[2,3],[10,11],[14,18]\}$ et $mo(A \rightarrow B \rightarrow F, S) = \{[2,5],[10,12]\}$. Si l'on considère une largeur de fenêtre de deux unités temporelles, alors nous avons $support(A \rightarrow B \Rightarrow F, S, 2) = 1$ and $confiance(A \rightarrow B \Rightarrow F, S, 2) = \frac{1}{2}$. Cela signifie que $A \rightarrow B \Rightarrow F$ apparaît une fois et a une confiance de 50% pour une largeur de fenêtre de deux unités temporelles. Pour une règle d'épisode λ, et pour la plus petite largeur de fenêtre possible w, si

- le support de λ est supérieur ou égal à σ,
- la confiance c_w de λ est supérieure ou égale à γ,
- il existe une largeur de fenêtre $w'|w' > w$ telle que la confiance de λ pour w' est $decreaseRate$% plus petite que c_w,
- il n'y pas de largeur de fenêtre entre w et w' pour laquelle la confiance est supérieure à c_w,

alors, la règle d'épisodes λ est une règle présentant un premier maximum local de confiance nommé *First Local Maximum (FLM)*. La règle λ est, dans ce cas, une *FLM-règle*. Le paramètre *decreaseRate* est défini par l'utilisateur et permet de sélectionner des

2. Une définition plus générique est disponible dans [MTIV97].

maxima locaux de confiance plus ou moins prononcés. La largeur de fenêtre w correspondant au premier FLM est appelée *fenêtre temporelle optimale* de la FLM-règle λ. Si nous réglons $decreaseRate$ à 30%, σ à 2, γ à 100% et $maxgap$ à 4, alors la règle $r = A \rightarrow B \Rightarrow F$ (figure 10.2) est une FLM-règle qui a un FLM pour une largeur de fenêtre de 3 unités temporelles. Cela peut être interprété comme : « si la prémisse de r apparaît à $t_p s$ et finit à $t_p e$, alors sa conclusion doit apparaître après $t_p e$ et jusqu'à $t_p s + w$, avec $w = 3$ ». Le paramètre γ peut bien sûr être réglé à une valeur inférieure à 100%. Dans ce cas, la probabilité d'observer la conclusion entre $t_p e$ et $t_p s + w$ est supérieure ou égale à γ. Des définitions plus formelles sont disponibles dans [MR04]. Les FLM-règles permettent de réduire les collections de règles d'épisodes aux seules règles présentant une fenêtre temporelle optimale, ce qui améliore l'échange avec l'utilisateur final : les collections de motifs extraits sont en effet moins importantes. Par ailleurs, utiliser de telles règles pour prévoir des événements fait sens : chaque règle dispose de sa propre information temporelle, sa fenêtre temporelle optimale.

window width	1	2	3	4	5
$mo(A \rightarrow B \rightarrow F, S)$	$\varnothing$	$\{[10, 12]\}$	$\{[2, 5], [10, 12]\}$	$\{[2, 5], [10, 12]\}$	$\{[2, 5], [10, 12]\}$
$mo(A \rightarrow B, S)$	$\varnothing$	$\{[2, 3], [10, 11]\}$	$\{[2, 3], [10, 11]\}$	$\{[2, 3], [10, 11], [14, 18]\}$	$\{[2, 3], [10, 11], [14, 18]\}$
$support(A \rightarrow B \rightarrow F, S)$	0	1	2	2	2
$support(A \rightarrow B, S)$	0	2	2	3	3
$confidence(A \rightarrow B \Rightarrow F, S)$	0	$1/2 = 50\%$	$2/2 = 100\%$	$2/3 = 66\%$	$2/3 = 66\%$

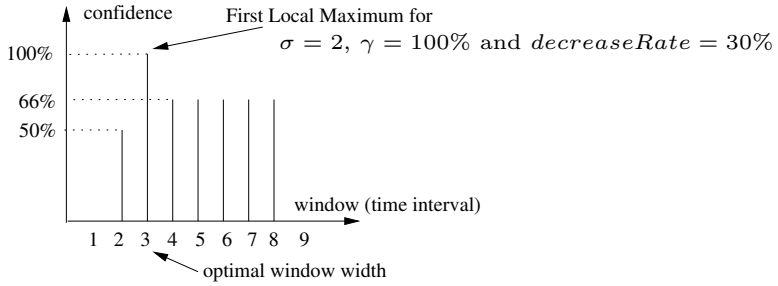
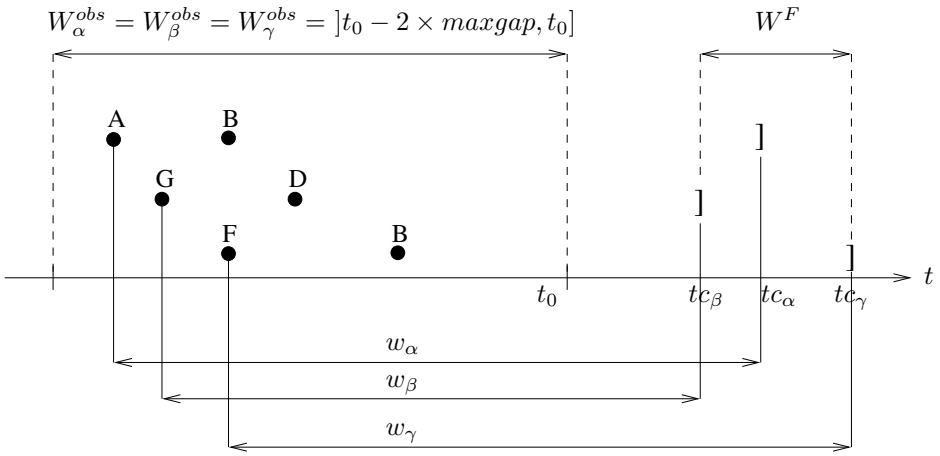


FIGURE 10.2 – Confiance et support de la règle $A \rightarrow B \Rightarrow F$ dans la séquence S (figure 10.1), pour $maxgap = 4$.

10.2.2 Prévisions : une approche *au plus tard*

Comme évoqué dans la section 10.2.1, nous proposons de décrire le comportement d'un système complexe à l'aide de FLM-règles puis d'utiliser celles-ci pour prévoir des événements dont les types sont au préalable indiqués par l'utilisateur. Généralement, comme expliqué dans la section 10.1, il s'agira de défaillances, de pannes affectant le système.



α : $A \rightarrow B \Rightarrow C$, w_α
β : $G \rightarrow D \Rightarrow C$, w_β
γ : $F \rightarrow B \Rightarrow C$, w_γ

FIGURE 10.3 – Prévision d'un problème de classe C à partir des règles α, β et γ.

Au niveau de la méthode de prévision, pour chaque FLM-règle r, nous proposons d'établir une *fenêtre d'observation* $W_r^{obs} =]t_0 - k \times maxgap, t_0]$ où t_0 est la date à laquelle l'utilisateur souhaite effectuer une prévision, *maxgap* est la contrainte de gap maximum utilisée lors de l'apprentissage et k est le nombre de types d'événements de la prémisse. Soit $[t_ps, t_pe]$ l'occurrence la plus tardive de la prémisse de r observée dans W_r^{obs}. Cette observation est faite sur la base d'une modification de l'algorithme *WinMiner* permettant de gérer la contrainte de gap maximum, de considérer des occurrences minimales, et de remonter le flot à partir de t_0 afin d'éviter tout traitement inutile, comme cela est proposé dans [CZC07]. La conclusion de la règle r est alors réputée apparaître *au plus tard* à $tc_r = t_ps + w_r$ avec $tc_r > t_0$ et w_r la fenêtre optimale de la règle r. La probabilité d'apparition est égale à la confiance associée à la fenêtre temporelle de la règle r. Soit T^C l'ensemble des dates de prévisions associées à un type d'événement C à prévoir. Soit $W^F = [t_s^f, t_e^f]$ la fenêtre de prévision associée à ce même type d'événement et définie comme suit : $t_s^f = min(T^C)$ et $t_e^f = max(T^C)$. Au final, pour chaque type d'événement à prévoir, nous générons, s'il y a lieu, un avertissement associé à une fenêtre de prévision qui est une fenêtre de prévision au plus tard. Par définition des FLM-règles, cela garantit le maximum de chances, sur cette période, d'avoir observé l'événement prévu, y compris avant cette même période. La figure 10.3 résume cette méthode pour la prévision d'un type d'événement C à partir des règles α, β et γ.

10.2.3 Expériences

Afin d'évaluer les performances de notre approche, nous avons mené des expériences sur des données synthétiques en utilisant, pour l'extraction des règles, le prototype *WinMiner*, issu des travaux présentés dans [MR04], et dont j'assure la distribution au sein de

la communauté. La plate-forme d'exécution est un PC standard équipé d'un processeur AMD Athlon(tm) 64 3000+ (1800 MHz) avec 512 Mo de RAM fonctionnant sous Linux (kernel 2.6).

Données et prétraitement des données Notre approche a été testée sur un jeu de données issu de la simulation d'une chaîne logistique. Cette simulation a été effectuée à l'aide du logiciel spécialisé ARENA par Julien Boissière, maître de conférence au LISTIC. Une chaîne logistique est un système complexe dont le pilotage est source de compétitivité [SKM03]. Ce type de système produit un flot de données, généralement enregistré au sein d'un ERP (Enterprise Resource Planning), caractérisant les stocks, les commandes et les livraisons pour l'ensemble des dépôts au cours du temps. Si un modèle est induit à partir de ces données, il devient envisageable de prévoir les problèmes pouvant affecter la chaîne logistique. Symeonidis *et al* [SKM03] ou Chen *et al* [CHCW05] utilisent ainsi une approche basée sur du clustering et des règles d'association pour adapter et piloter une chaîne logistique en fonction des fournisseurs et des clients. Pour en revenir à notre simulation, la chaîne logistique simulée est une chaîne *divergente* : elle est composée d'un entrepôt D_0 qui fournit 3 dépôts D_1, D_2, D_3 eux-mêmes connectés à 9 clients dits *agrégés*. Chaque client représente en effet une famille de clients achetant le même type de bien. Les politiques de réapprovisionnement sont basées sur des seuils : lorsque les stocks passent en-dessous d'un certain seuil, appelé *seuil de réapprovisionnement*, une quantité fixe est commandée au fournisseur. Notre chaîne distribue 10 produits différents notés p_i avec $i \in [1 \dots 10]$. La demande des clients finaux est simulée à l'aide d'un intervalle entre les demandes suivant une loi normale et de la quantité demandée pour chaque produit suivant une distribution de Poisson. Aucun délai de production n'est introduit dans le système, seuls des délais dus au transport sont considérés. Différents indicateurs sont utilisés pour qualifier les états de la chaîne. Tout d'abord, nous avons les niveaux de stocks pour chaque produit sur chaque dépôt. Il sont notés $inv_{i,j,k}$, c'est-à-dire le niveau de stock (inventory level) k du produit i au dépôt j avec $j \in [0 \dots 3]$ et $k \in \{low, medium, large\}$. Les stocks sont *low* si le niveau est en-dessous du seuil de sécurité, *medium* s'il est en-dessous du seuil de réapprovisionnement et $large$ s'il est au-dessus. Chaque commande génère un facteur de satisfaction noté $sat_{b,l}$ qui reflète le niveau de satisfaction l du client b en fonction des dates de livraion estimées et effectives tel que $l \in \{ontime, late\}$ et $b \in [1 \dots 9]$. Si le dépôt dispose de suffisamment de stock, alors la demande peut être satisfaite dans les 2 jours : la demande est satisfaite (*ontime*). Si le client doit attendre plus longtemps, il est insatisfait (*late*). Dans tous les cas, les commandes sont distribuées. De façon à gérer efficacement notre chaîne logistique, l'attention sera mise sur la satisfaction des clients finaux. En effet, il est intéressant pour les décideurs de connaître à l'avance quels sont les clients qui ne seront pas satisfaits de façon à en limiter le nombre. Le type d'événement $sat_{b,l}$ est donc celui dont on veut prévoir les occurrences en particulier pour $l = late$. Deux jeux de données ont été simulés : le jeu d'apprentissage et le jeu de test. Chaque jeu couvre une année de fonctionnement. La demande globale est plutôt régulière bien que 3 clients aient une demande saisonnière, c'est-à-dire qu'ils demandent deux fois plus de produits durant l'été. Cette saisonnalité génère d'intéressantes fluctuations dans la chaîne logistique. L'unité temporelle est la minute et 300000 événements sont générés à partir d'un alphabet de 250 symboles. Les trois clients qui commandent le plus sont les clients 1 à 3 et les paramètres sont réglés de façon à ce qu'il soient le plus souvent insatisfaits pendant la simulation.

Apprentissage Le choix des différente paramètres a été fait après plusieurs expériences. Tout d'abord, la confiance minimum γ a été fixée à 0.9 pour obtenir des règles fortes. Le gap maximum $maxgap$ a été fixé à 10000 minutes. De façon à obtenir assez de règles pour tous les clients insatisfaits, le support minimum a été fixé 100, ce qui est un support très bas pour ce jeu de données. De façon à obtenir des règles génériques et éviter le sur-apprentissage, nous avons limité le nombre de symboles par règle à 3, ce qui par ailleurs permet d'envisager un support aussi faible que celui évoqué précédemment. Le dernier paramètre à régler est le taux de décroissance, $decreaseRate$. Dans ce cas particulier, il nous a fallu le mettre à 0%. En effet, dès 2%, trop peu de FLM-règles sont fournies : 12 exactement. Autrement dit, la confiance augmente généralement avec les tailles de fenêtres jusqu'à atteindre un niveau, supérieur ou égal à γ qui par la suite reste stable puis, dans certains cas, augmente à nouveau. La plus petite fenêtre pour laquelle ce niveau est atteint est la fenêtre optimale. En moyenne, pas plus de 4% d'augmentation pour la confiance et de 8% d'augmentation pour le support sont observés pour les fenêtres plus larges que les fenêtres optimales. En 9 heures, environ 3000000 de FLM-règles sont extraites. Ce nombre est important car (1) le support est très bas et (2) un taux de décroissance à 0% revient à extraire autant de FLM-règles que de règles d'épisodes. Une fois filtrées sur les conclusions évoquant des clients insatisfaits, le nombre de FLM-règles est réduit à 37282. La distribution de ces règles correspond à la fréquence d'insatisfaction de simulation des 3 clients affectés par des problèmes de livraison. Ainsi, nous avons 36877 règles pour le client 1, 226 règles pour le client 2 et 179 pour le client 3. À propos des fenêtres optimales, il convient à nouveau de faire le distingo au niveau des clients : pour le client 1, la fenêtre optimale moyenne est de 1070 minutes, pour le client 2 elle passe à 6391 minutes et à 2434 minutes pour le client 3. Cela correspond également à notre simulation : les clients 1 et 3 commandent plus souvent que le client 2.

Prévisions Nous avons ensuite testé notre méthode de prévision pour différentes dates t_0 en utilisant notre jeu de test. Nous avons fait une prévision par semaine, c'est-à-dire 52 prévisions. Nous avons ensuite évalué nos prévisions en s'appuyant sur les cas suivants :

- cas 1 : notre méthode prévoit un problème de classe C et fournit l'intervalle $[t_s^f, t_e^f]$. Dans ce cas, nous vérifierons qu'au moins un problème de classe C apparaît bien dans l'intervalle de prévision fourni.
- cas 2 : aucune alerte n'est émise. Aucun intervalle n'est donc fourni. Dans ce cas, nous vérifierons qu'aucun problème n'apparaît dans l'intervalle $]t_0, t_0 + maxgap]$. En effet, le dernier événement de la prémisse peut être détecté à t_0 et la conclusion qui suit peut apparaître jusqu'à $t_0 + maxgap$. Au-delà, du fait de l'utilisation de la contrainte de gap maximum, notre méthode d'apprentissage n'aurait en aucun cas pu fournir de FLM-règle à même de prévoir une conclusion aussi tardive.

Seulement 3 clients sont insatisfaits dans notre jeu de test. La question principale est donc de prévoir correctement le moment où ils seront insatisfaits. Les résultats sont données par la figure 10.4. Pour chaque client b, nous notons $\widehat{late}$, le nombre de prévisions indiquant que le client b va être insatisfait, c'est-à-dire des FLM-règles finissant sur la conclusion $sat_{b,late}$ sont en voie d'apparaître. Quant aux prévisions ne prévoyant aucun problème à venir, celle sont notées $\widehat{ontime}$.

Comme on peut le constater, pour le client 1, nous sommes capables de prévoir 90% des problèmes sans fausse alarme et ce bien que ce client soit toujours insatisfait. Pour le client 2, 88% des problèmes sont prévus mais 4 fausses alarmes sont émises. Pour le client

customer 1	$\widehat{late}$	$\widehat{ontime}$
late	47	5
ontime	0	0

customer 2	$\widehat{late}$	$\widehat{ontime}$
late	32	2
ontime	4	14

customer 3	$\widehat{late}$	$\widehat{ontime}$
late	16	0
ontime	2	34

FIGURE 10.4 – Apprentissage au plus tard : performances.

3, 96% des problèmes sont prévus et seulement 2 fausses alarmes sont à mentionner. En ce qui concerne les fenêtres de prévision, les dates moyennes de début et de fin d'intervalle sont : $t_o + 6$ heures et $t_0 + 54$ heures pour le client 1, $t_0 + 16$ heures et $t_0 + 208$ heures pour le client 2, $t_0 + 15$ heures et $t_0 + 87$ heures pour le client 3. Ceci laisse suffisamment de temps aux utilisateurs finaux pour mettre en oeuvre des actions correctrices, c'est-à-dire piloter la chaîne logistique. Au final, en moyenne, nous sommes capables de prévoir 93% des problèmes.

Synthèse Nous proposons donc de s'appuyer sur les FLM-règles pour construire un modèle de comportement d'un système complexe et d'utiliser ce modèle pour prévoir, au plus tard, les occurrences d'événements d'intérêt en fournissant les fenêtres temporelles de prévision associées. Cette méthode a été testée sur un jeu de données issu de la simulation d'une chaîne logistique. Bien qu'utilisées dans une configuration singulière avec un taux de décroissance à 0%, les FLM-règles extraites ainsi que les fenêtres optimales associées ont permis de prévoir, dans le temps, 93% des problèmes sur un nouveau jeu de données, le tout en n'émettant que 4 fausses alarmes sur 52 prévisions. De plus, bien qu'il s'agisse de prévisions au plus tard, suffisament de temps est laissé aux utilisateurs finaux pour décider des actions correctrices à déclencher dans le cas où des alarmes sont émises. Ces expériences montrent également que les motifs locaux peuvent être directement utilisés pour de la prévision et ce sans que l'utilisateur final ait à les consulter. Néanmoins, ces expériences ne portent que sur des données synthétiques et n'ont pas permis de valider la méthode pour des FLM-règles typiques, c'est-à-dire dont le taux de décroissance est supérieur 0, comme indiqué dans [MR04] ou [MLLR04]. De plus, le mode d'apprentissage des règles pourrait être raffiné, voire durci, en ne sélectionnant que les règles les plus génériques ne générant pas de fausse alarme. Par ailleurs, la méthode de prévision proposée dans cette section s'appuie sur une prévision au plus tard des événements, ce qui intrinsèquement peut disqualifier la méthode pour certaines applications, de maintenance prédictive en particulier [SG07]. Or, les FLM-règles permettent de prévoir, avec une confiance/probabilité connue, l'apparition d'un événement entre la fin de la prémisse et la fenêtre optimale : l'utilisation proposée ici n'utilise donc pas toutes les informations disponibles, informations qui permettent d'adopter une stratégie de prévision au plus tôt. De plus, les prévisions sont recalculées pour chaque instant d'observation et de prévision t_0. Une même occurrence

d'une prémisse de FLM-règle peut donc être retrouvée plusieurs fois ce qui serait inutile si un effet mémoire était envisagé. Enfin, tout le flot de données est conservé alors qu'un phénomène *d'oubli* est mis en place au travers des fenêtres d'observations. À propos de ces dernières, et afin de réduire leurs étendues temporelles, celles-ci pourraient être restreintes non pas en fonction de la contrainte de gap maximum mais en fonction des fenêtres optimales.

10.3 Propositions

Afin de répondre aux réserves émises en fin de section 10.2, de nouveaux travaux ont été engagés dans le cadre de la thèse Florent Martin. Celle-ci a été menée d'octobre 2007 à juin 2011 sur le support d'un contrat CIFRE avec ADIXEN (Alcatel-Lucent, aujourd'hui Pfeiffer Vacuum). L'entreprise ADIXEN produit des pompes à vides. Ce sont des systèmes mécaniques complexes fonctionnant dans des conditions sévères et variables pour lesquels il nous a été demandé de prévoir un certain type de panne. Un BQR dédié à l'apprentissage sur des systèmes dynamiques a également servi de support à ces travaux.

10.3.1 Apprentissage : une approche *leave-one-out*

L'approche présentée ici est à décliner pour chaque type d'événement à prévoir. Afin de sélectionner les règles les plus fiables, c'est-à-dire les plus génériques et générant le moins possible de fausses alarmes, nous proposons d'extraire et sélectionner les FLM-règles ayant une confiance de 100% tout en s'appuyant sur une approche de validation croisée de type *leave-one-out* [TSK05]. Soit une séquence d'événements d'apprentissage D découpée en sous-séquences Seq_i. Ce découpage peut par exemple être fait en fonction des occurrences des événements à prévoir, chaque sous-séquence finissant sur un tel événement. Nous proposons que chaque sous-séquence Seq_i serve, à tour de rôle, et une fois, de jeu de validation. Les séquences restantes constituent alors le jeu d'apprentissage. À chaque fois, les paramètres d'extraction des FLM-règles sont maintenus identiques, et les FLM-règles extraites du jeu d'apprentissage sont sélectionnées en fonction de leur capacité à prévoir un type d'événement et à n'émettre aucune fausse alarme sur le jeu de validation.

Pour ce faire, les règles extraites du jeu d'apprentissage et dont les prémisses ne sont pas retrouvées dans le jeu de validation sont rejetées : elles ne sont pas assez génériques. Afin de vérifier si une FLM-règle r ne génère aucune fausse alarme, on contrôle que sa conclusion apparaisse après chaque occurrence de sa prémisse. Si tel n'est pas le cas, la règle r pourrait générer des fausses alarmes : elle doit être écartée. Des contraintes temporelles sont bien sûr prises en compte. Soit w_r la fenêtre optimale de la règle r. Par construction des FLM-règles, chaque occurence d'une prémisse doit s'étendre sur moins de w_r unités temporelles. Si tel n'est pas le cas, nous sommes face à une configuration qu'il ne nous a pas été possible d'apprendre sur le jeu d'apprentissage et qu'il nous sera donc impossible de prévoir. Dans cette situation, la règle r est rejetée. Si l'occurrence de la prémisse, notée $[t_ps, t_pe]$, respecte la contrainte précédente, nous vérifions que la conclusion de r apparaît dans $]t_pe, t_pe + maxgap]$. Par définition, $t_ps + w_r \leq t_pe + maxgap$. Nous permettons donc à la conclusion d'apparaître après $t_ps + w_r$ ce qui est permissif par rapport à la définition des FLM-règles. Si tel est le cas, en utilisant r et w_r, nous serions capables de prévoir l'événement et ce trop tôt. Cela est toujours mieux que de ne rien prévoir : la règle r peut

être retenue. De plus, cela est cohérent avec l'approche *au plus tôt* développée dans le section 10.3.2. Si la conclusion apparaît après $t_pe + maxgap$, alors la règle r a été apprise avec un $maxgap$ trop faible. Une règle ayant plus de types d'événements pourrait peut-être gérer ce cas : la règle r est écartée.

Une fois que le processus d'apprentissage et de validation s'arrête, une collection de FLM-règles génériques ne générant pas de fausse alarme est constituée. Une même FLM-règle pouvant être extraite plusieurs fois avec des fenêtres temporelles différentes, la plus petite est sélectionnée comme fenêtre temporelle optimale afin de prévoir au plus tôt. L'ensemble des FLM-règles sélectionnées ainsi que les fenêtres temporelles optimales associées forment une *FLM-base*. Cette dernière sera utilisée pour la prévision du type d'événement pour lequel elle a été construite. Une version plus souple de cet apprentissage peut être envisagée en extrayant les FLM-règles avec une confiance minimum inférieure à 100% et en fixant un taux de fausses alarmes tolérées par l'utilisateur final.

10.3.2 Prévisions : une approche *au plus tôt*

Notre méthode de prévision est à dérouler pour chaque type d'événement à prévoir et se présente en 3 étapes : (1) les occurrences des prémisses des FLM-règles sont recherchées dans le flot de données surveillé, (2) les dates de prévisions au plus tard sont calculées à partir de chacune des occurrences et (3) un intervalle unique de prévision au plus tôt est construit sur la base des informations calculées précédemment.

De façon à retrouver les occurrences des prémisses des FLM-règles servant à prédire un type d'événement dans un flot de données, une file d'événements est construite. Cette file permet de ne pas conserver tout le flot de données, contrairement à ce qui était décrit dans la section 10.2. Elle est maintenue en ajoutant les nouveaux événements dont la date d'occurrence est la date système t_0. Les événements dont la date d'occurrence est égale ou antérieure à $t_0 - W$, avec W la plus grande fenêtre optimale de la FLM-base, sont supprimés. De cette façon, nous nous assurons qu'assez de données sont conservées pour idendifier toutes les prémisses des règles de la FLM-base. Cette file d'événement sera parcourue de façon à trouver les dernières occurrences en date des prémisses des FLM-règles. Ainsi, nous n'avons pas forcément besoin de lire toute la file, il suffit de remonter celle-ci depuis t_0, date à laquelle le dernier type d'événement d'au moins une des prémisses des règles de la FLM-base vient d'apparaître. Si un événement apparaît mais ne satisfait pas à cette dernière condition, aucun recherche n'est déclenchée, les occurrences antérieures à t_0 ayant déjà été identifiées par construction. Ce type de recherche est également proposé dans [CZC07] et permet de plus rapidement identifier l'occurrence d'une prémisse tout en conservant le moins possible de données. Lors de la rercherche des occurences des prémisses, il doit être tenu compte de la contraite $maxgap$. Comme expliqué dans la section 10.2, le risque d'incomplétude est avéré et l'utilisation des occurrences minimales préfixées doit être envisagée. L'algorithme WinMiner proposé dans [MR04] a donc été simplement modifié pour rechercher la prémisse d'une règle r, à partir de t_0 et tout s'assurant que la contrainte de $maxgap$ soit respectée (cf. [MMGB12]). Par ailleurs, cette prémisse est recherchée dans la fenêtre d'observation $W_r^{obs} =]t_0 - w_r, t_0]$ avec w_r la fenêtre temporelle optimale de r.

Une fois une occurrence $[t_ps, t_pe]$ de la prémisse d'une règle r retrouvée, la date d'apparition au plus tard de la conclusion tc_r est valorisée à $t_ps + w_r$. Pour chaque prémisse, et selon la définition des FLM-règles 10.2.1, sa conclusion doit en effet apparaître dans $]t_0, tc_r]$ avec 100% de confiance (si la confiance minimum a été réglée 100%). Un exemple

est donné par la figure 10.5 pour la règle $\alpha = A \rightarrow B \rightarrow C \Rightarrow P$.

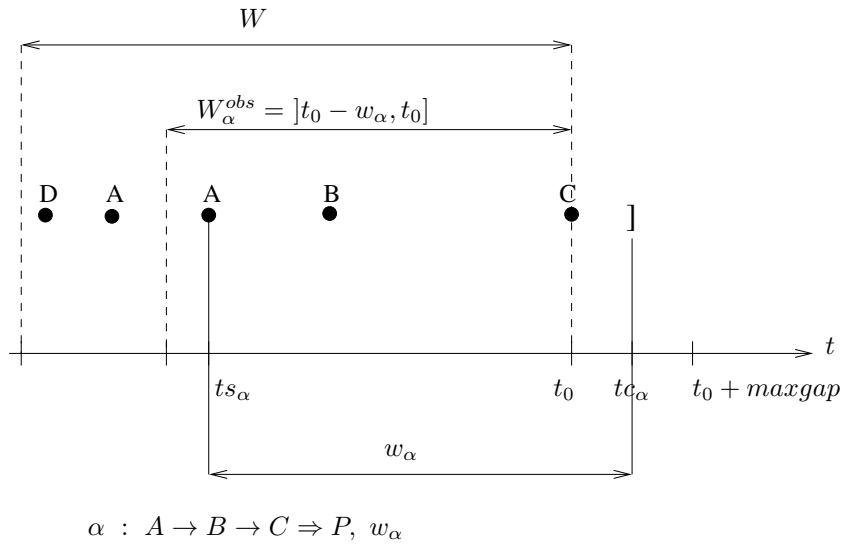


FIGURE 10.5 – Projection de la fenêtre temporelle optimale de la règle $A \rightarrow B \rightarrow C \Rightarrow P$.

Les prévisions effectuées à des dates antérieures à t_0 sont mémorisées. Comme cela vient d'être évoqué, dans cette configuration, les occurrences précédentes, c'est-à-dire finissant à des dates antérieures à t_0, ont par construction déjà été trouvées et les dates de prévision associées calculées. Si ces dernières sont postérieures à t_0, celles-ci sont dites *actives* et utilisées pour la prévision. Un effet mémoire est donc introduit et permet de ne pas avoir à retrouver une même occurrence d'une prémisse plusieurs fois comme cela était le cas dans le proposition précédente (cf. section 10.2). Soit T^C l'ensemble des dates de prévisions au plus tard tc_r qui sont actives pour un type d'événement C à t_0 ; ces dates pouvant avoir été calculées avant et à t_0. L'intervalle de prévision associé $]t_s^f, t_e^f]$, également appelé *fenêtre de prévision* W^F, est tel que $t_s^f = t_0 \wedge t_e^f = min(Tc)$. En choisissant l'opérateur *min* comme opérateur d'agrégation, nous fournissons un intervalle de prévision au plus tôt. La figure 10.6 fournit une intervalle de prévision calculé à partir des règles α, β, δ dont les prémisses ont été reconnues à t_0^{n-1} et à t_0^n.

10.3.3 Expériences

Afin d'évaluer les performances de cette nouvelle approche brevetée [BMP+10] et proposée dans [MMGB10b, MMGB10a, MMGB12], nous avons mené des expériences sur des données réelles, issues de pompes à vides et fournies par ADIXEN. L'extraction des règles a à nouveau été faite à l'aide du prototype *WinMiner*. La plate-forme d'exécution est un PC standard équipé d'un processeur Intexl Xeon 5160 (3 GHz) avec 4 GB de RAM fonctionnant sous Linux (kernel 2.2).

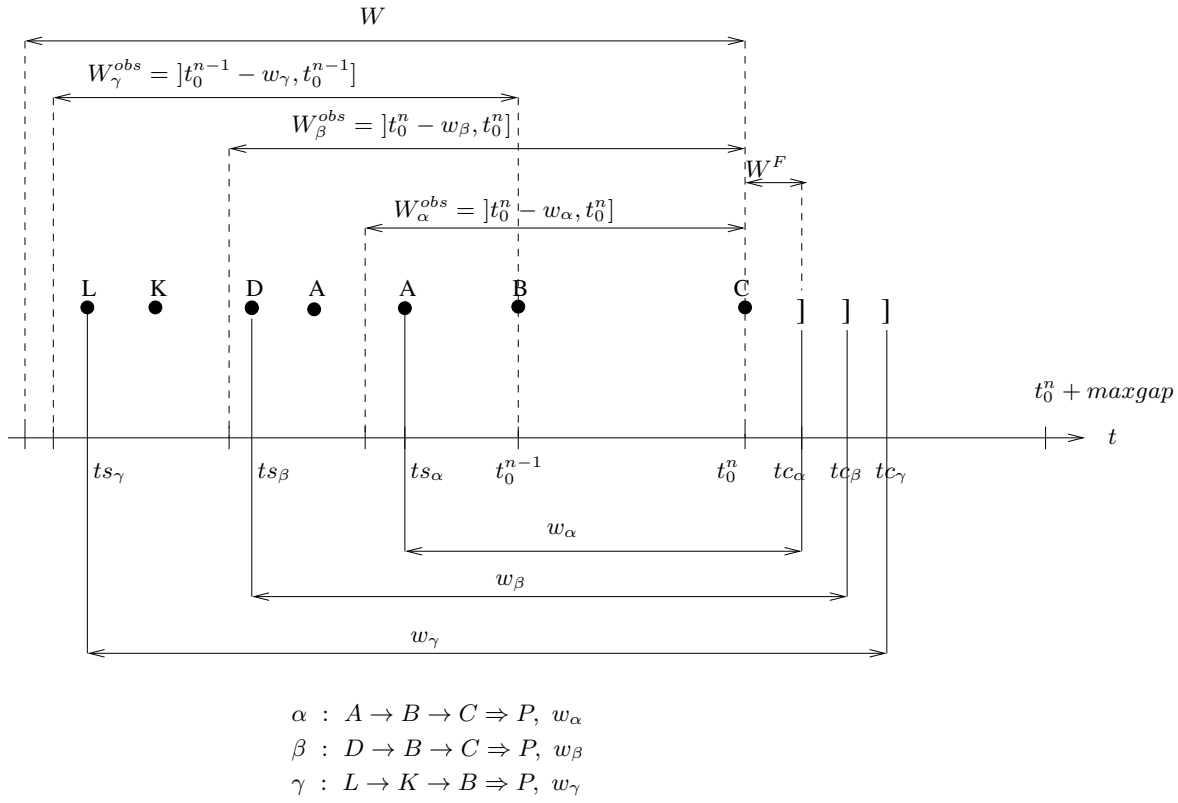


FIGURE 10.6 – Agrégation des informations temporelles de prévision.

Données et prétraitement des données Les pompes à vide, dont la fonction principale est de créer du vide en extrayant l'air d'un volume étanche, sont des systèmes dont la cinétique est complexe. Ces systèmes fonctionnent sous des conditions sévères et imprévisibles. Une défaillance majeure est le *grippage* des axes de la pompe. Ce blocage peut être provoqué par différentes causes comme la chaleur (augmentation du volume des axes) et/ou la condensation de gaz. Dans ces conditions, il est difficile d'établir un planning de maintenance préventive et seule une maintenance prédictive peut être envisagée. Autrement dit, il s'agit pour nous de prévoir de tels grippages. Comme expliqué dans [DFP98], l'analyse de signaux vibratoires constituent une piste prometteuse pour ce genre d'équipements, l'intensité de ceux-ci traduisant la présence de défauts affectant les systèmes surveillés. Les signaux vibratoires de 64 pompes en fonctionnement ont donc été collectés chez un client du marché des semi-conducteurs sur une durée de 2 ans (en cumulé). Le postulat de départ est que ces signaux sont représentatifs des motifs de dégradation des pompes qui peuvent émerger des différentes conditions d'utilisation possibles. Si notre méthode de prévision fonctionne, nous ne surveillerons pas (ou peu) de systèmes endommagés. Les signaux enregistrés avant les premiers grippages sont donc sélectionnés. Apprendre à partir de systèmes dégradés fausserait en effet notre FLM-base.

La valeur efficace de signaux vibratoires, qui peut être interprétée en terme de puis-

sance, est ensuite calculée pour 20 bandes de fréquence (plus de détails sont disponibles dans [MMGB12]). Décrire ces signaux par bande de fréquence permet en effet de remonter aux causes des grippages, aux défauts affectant les pompes. La période d'acquisition de ces valeurs est de 80 secondes. La valeur efficace, notée $Vrms$ (Valeur root mean square), est ensuite associée à un niveau de sévérité traduisant un défaut apparaissant dans la bande de fréquence concernée. Pour ce faire, et comme expliqué dans [Lal99, Bro73], la sévérité est établie grâce au ratio des puissances :

$$R_n = \frac{Vrms_{(n,T)}}{Vrms_0} \tag{10.1}$$

avec $Vrms_{(n,T)}$ la n$^{\text{ème}}$ mesure de valeur efficace du signal vibratoire sur une période $T = 80s$ et $Vrms_0$ la valeur efficace de référence de la pompe lorsque celle-ci fonctionne dans de bonnes conditions. $Vrms_0$ est définie comme la plus petite valeur efficace calculée sur une période P, avec $P \gg T$ et telle que P corresponde à une période complète d'utilisation de la pompe. Déterminer P n'est pas chose aisée. Selon les experts d'ADIXEN, elle doit commencer au moins 24 heures après le démarrage de la pompe pour ne pas considérer la période de mise en route et doit se finir lorsque la valeur de $Vrms$ est stabilisée, celle-ci baissant graduellement par effet de rodage. Dans notre cas, cela se produit généralement après 2 semaines de fonctionnement. Afin de prendre en compte tout type de choc, les valeurs efficaces sont prises sur l'enveloppe maximale et le ratio des puissances est calculé à partir de ces valeurs. Ce dernier est alors discrétisé en trois niveaux. Le premier niveau correspond à $R_n < \alpha$, ce qui signifie que $Vrms_{(n,T)}$ est inférieur à $\alpha \times Vrms_0$. Le deuxième niveau correspond à $\alpha \leq R_n < \beta$ et le troisième à $R_n \geq \beta$.

Chaque fois que $Vrms_{(n,T)}$ change de niveau, un événement est généré. Son type est précisé par un symbole donnant 3 informations : la bande fréquence (20 bandes disponibles dans notre cas), la sévérité du défaut observé avant le changement (3 niveaux) et la durée écoulée à ce niveau de sévérité (4 types de période). Les 4 types de période associés à cette durée vont de quelques minutes à plus de 10 jours. Ces périodes sont issues de différentes expériences. Au final, un dictionnaire de 240 symboles est défini et 13 séquences (seuls 13 grippages sont présents dans les données) contenant environ chacune 2000 événements sont construites. La figure 10.7 illustre notre prétraitement. La première courbe représente un échantillon de $Vrms_{(n,T)}$. La deuxième montre son enveloppe maximale. Le troisième graphique donne un exemple d'encodage de l'enveloppe maximale : les 2 premiers chiffres donnent la bande fréquence, le troisième et le quatrième donnent respectivement le niveau de sévérité et la durée à ce niveau. Ainsi, le symbole 1314 signifie que « le signal correspondant à la bande de fréquence 13 était bas (1) pendant quelques semaines (4) ». Le symbole 1322 signifie quant à lui que « le signal correspondant à la bande de fréquence 13 était élevé (2) pendant quelques heures (2) » et le symbole 1333 signifie que « le signal correspondant à la bande de fréquence 13 était très haut (3) pendant quelques jours (3) ».

Apprentissage Lors de l'apprentissage, à chaque itération de notre approche leave-one-out (cf 10.3.1), les séquences servant de jeu d'apprentissage sont concaténées en une seule séquence et espacées de $maxgap + 1$ unités temporelles de façon à ne pas trouver de règle dont les occurrences s'étendraient sur 2 grippages différents. Sachant que les grippages peuvent provenir de différentes causes et que seuls 13 d'entre eux sont à notre disposition, le support minimum σ a été réglé à 2, ce qui est très bas. De façon à extraire les règles les plus confiantes et à limiter le nombre de fausses alarmes, la confiance minimum

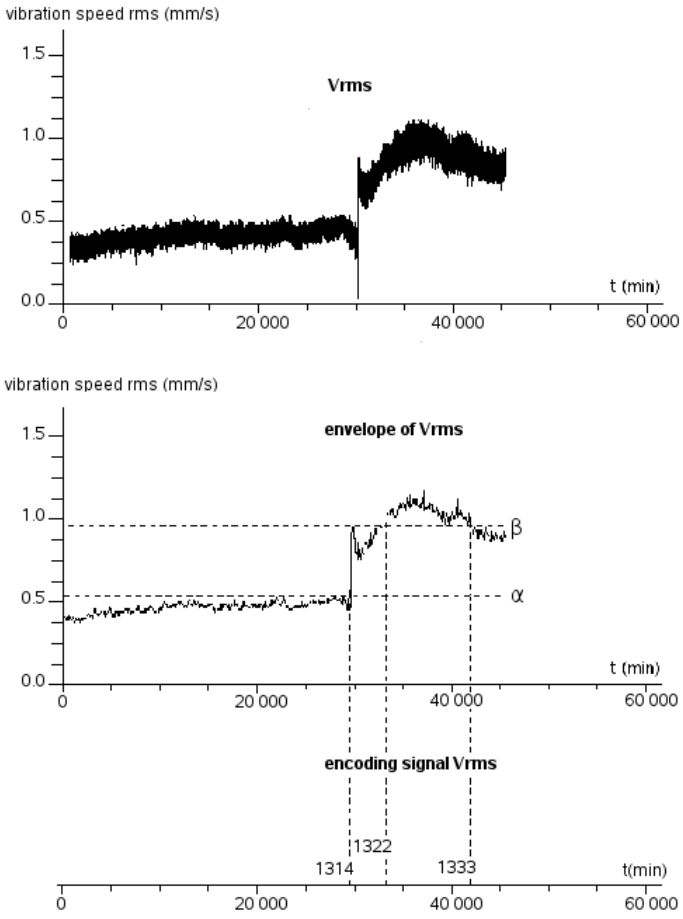


FIGURE 10.7 – Prétraitement des signaux vibratoires.

γ a été réglée à 100%. Le taux de décroissance *decreaseRate* a cette fois pu être réglé à 30% pour sélectionner des fenêtres temporelles optimales marquées et le gap maximum *maxgap* a été réglé à une semaine de façon à couvrir de très larges fenêtres optimales. Enfin, seules les règles contenant 4 types d'événements au maximum ont été considérées de façon à obtenir des règles génériques et à procéder à des extractions standard en termes de temps et d'espace. Ainsi, le processus de construction de la FLM-base ne prend pas plus de 3 heures.

Pour chaque itération de notre approche, le nombre de règles extraites varie de 7509368 à 8713574 alors que le nombre de FLM-règles varie de 2760307 à 3554548. Parmi elles, de 431 à 486 se concluent par le symbole *grippage*. À la fin du processus d'extraction et de sélection, nous obtenons une FLM-base contenant 29 FLM-règles et leurs fenêtres temporelles optimales respectives. Les bandes de fréquence impliquées dans ces règles étaient toutes attendues à l'exception d'une seule. Après simulation et validation de nos experts, il

s'est avéré que nous venions de découvrir une nouvelle bande de fréquence qui n'avait pas été mentionnée jusqu'à présent dans la littérature ou dans des rapports internes. La distribution des fenêtres temporelles optimales présente 2 modes particulièrement marqués : 3 jours (15 FLM-règles) et 10 jours (14 FLM-règles). Cela montre clairement qu'utiliser une seule fenêtre temporelle d'apprentissage et de prévision est trop restrictif. Les règles extraites sont du type $(1931- > 1933- > 1933 => 4, 100, 2448, 3)$, ce qui signifie « si le signal de la bande de fréquence 19 est très élevé (3) pendant quelques minutes (1) puis très élevé (3) pendant quelques jours (3) et ce 2 fois ($1933- > 1933$), alors la pompe a une probabilité de 100% de gripper (événement 4) entre la date du dernier événement (1933) et la date du premier événement (1931) augmentée de 2448 minutes. Cela a été observé 3 fois. ».

Prévisions

Lorsque des grippages sont prévus, deux informations sont fournies aux utilisateurs finaux. La première, la *prévision de défaillance* indique que le grippage va survenir ou non. La deuxième donne la *fenêtre de prévision*, c'est-à-dire l'intervalle temporel dans lequel le grippage est censé apparaître. La contrainte industrielle portant essentiellement sur la première information et sur la capacité de la méthode à ne pas générer de fausses alarmes, nous avons évalué de façon séparée ces deux informations, contrairement à ce qui a été fait dans la section 10.2 où la prévision d'un événement était jugée correcte si et seulement l'intervalle de prévision associé était également correct.

En ce qui concerne la prévision de défaillance, chaque fois qu'un événement appartenant au dernier type d'événement d'une prémisse d'une des règles de la FLM-base apparaît, une prévision de défaillance est déclenchée. Son résultat est soit « un grippage est prévu » (classe minoritaire notée « + ») soit « aucun grippage n'est prévu » (classe majoritaire notée « - »). Ces deux classes caractérisent les pompes dont l'état vient de changer. Bien entendu, il faut prendre en compte les aspects temporels : les objets sont inhabituels puisque définis pour une pompe et une date t_0 et la classe de ces objets ne peut être vérifiée qu'en observant les données après t_0. Cette vérification est effectuée en prenant en compte l'intervalle $]t_0, t_0 + maxgap]$. En effet, aucun événement ne peut être prévu après $]t_0, t_0 + maxgap]$ par construction. Ayant à disposition des objets et leurs classes, il est possible de produire une matrice de confusion rassemblant les quatre cas suivants :

- cas 1 : vrais positifs (VP) - une défaillance est prévue et apparaît dans $]t_0, t_0 + maxgap]$.
- cas 2 : faux négatifs (FN) - aucune défaillance n'est prévue mais une défaillance apparaît dans $]t_0, t_0 + maxgap]$.
- cas 3 : faux positifs (FP) - une défaillance est prévue mais aucune défaillance n'apparaît dans $]t_0, t_0 + maxgap]$.
- cas 4 : vrai négatifs (VN) - aucune défaillance n'est prévue et aucune défaillance n'apparaît dans $]t_0, t_0 + maxgap]$.

Les nombres rapportés pour ces cas servent de base au calcul des mesures standard de classification comme l'exactitude, le rappel, la précision ou la spécificité. Une présentation complète de ces mesures est disponible dans [TSK05].

Nous avons appliqué cette évaluation sur 2 jeux de données : les 13 séquences utilisées pour construire notre FLM-base (dataset 1) et 21 nouvelles séquences de production (dataset 2). Pour rappel, la FLM-base a été construite à partir du dataset 1. Elle a été utilisée ici en simulant 2 flots de données, un par jeu de données. Au final, nous avons fait 24125 prévisions pour le dataset 1 et 32525 prévisions pour le dataset 2. Les résultats des évaluations sont données dans le tableau 10.1 et dans le tableau 10.2 en utilisant des matrices de confusion.

TABLE 10.1 – Dataset 1 - Matrice de confusion.

		Classes prédites	
		+	−
Classes observées	+	492	262
	−	20	23351

TABLE 10.2 – Dataset 2 - Matrice de confusion.

		Classes prédites	
		+	−
Classes observées	+	300	0
	−	404	31821

Les mesures d'exactitude, de rappel, de précision et de spécificité qui en découlent sont présentées dans le tableau 10.3).

TABLE 10.3 – Mesures de classification pour le dataset 1 et le dataset 2.

dataset	exactitude	rappel	précision	spécificité
dataset 1	0.98	0.64	0.96	0.99
dataset 2	0.98	1	0.43	0.99

Bien que nous n'ayons appris que très peu de grippages, les résultats sont encourageants puisque nous prévoyons 10 grippages sur 13 avec 98% d'exactitude sur le dataset 1 et que nous annonçons 2 grippages sur 2 avec 98% d'exactitude sur le dataset 2. En regard de la classe majoritaire, il y a très peu de fausses alarmes ce qui est important dans notre contexte : les mesures de spécificité atteignent 99% pour les deux jeux de données. De plus, les 20 et 404 fausses alarmes (tableaux 10.1 et 10.2) sont générées pour des pompes qui ont réellement subi un grippage quelques jours plus tard. La mesure de précision (96% pour le dataset 1, 43% pour le dataset 2) doit donc être considérée avec précaution. Un exemple typique est donné par la figure 10.8. Elle montre en effet l'évolution d'un intervalle de prévision concernant le grippage #11. On remarquera que la première fenêtre de prévision, fournie 7 jours avant le grippage, n'incluait pas la date d'occurrence du grippage. Cependant, toutes les fenêtres fournies à partir du cinquième jour avant la défaillance incluent cette date. Cette figure montre également que plus l'on se rapproche de la défaillance et plus l'intervalle de prévision est précis. Puisque la faible mesure de précision obtenue pour le dataset 2 est due à des prévisions trop précoces, les fenêtres optimales apprises sur le dataset 1 étaient trop étroites. Ceci peut être expliqué par le nombre réduit de grippages (13) qui a imposé un support minimum très bas lors de l'apprentissage. D'un point de vue applicatif, cela ne représente pas un problème majeur puisque les pompes concernées ont toutes grippé quelques jours plus tard. Finalement, les 262 faux négatifs sur le dataset 1 sont dus à trois 3 grippages qui n'ont pu être prédits.

La mesure de rappel associée est ainsi de 64% tandis qu'elle monte à 100% sur le dataset 2. Une si faible mesure de rappel indique que notre processus d'apprentissage tend à ne conserver que les règles très génériques. C'est en effet le cas puisque les règles spécifiques sont écartées par l'approche leave-one-out. Ceci nous empêche de prédire 3 grippages. Cela doit néanmoins être balancé par le fait que les règles sont suffisamment génériques pour prévoir les grippages du dataset 2 qui a été construit en surveillant des pompes assignées à des processus de fabrication très différents de ceux concernant le dataset 1 (les gaz utilisés sont différents).

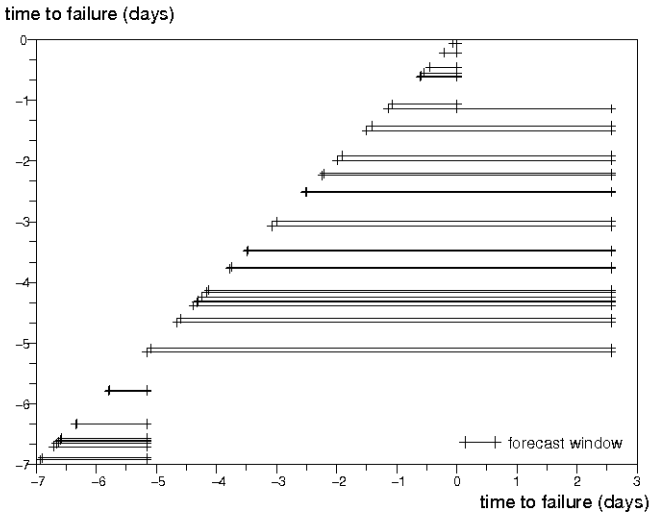


FIGURE 10.8 – Évolution de l'intervalle de prévision associé au grippage #11.

En ce qui concerne l'évaluation des intervalles de prévision $]t_0, t_e^f]$, nous allons nous attacher à vérifier que ces derniers sont plus pertinents qu'un simple intervalle $]t_0, t_0 + maxgap]$ qui pourrait être fourni en première approche. Dans [ZGP10], les auteurs proposent différentes mesures pour évaluer les techniques de pronostic dont l'exactitude et la précision. Celles-ci diffèrent de celles utilisées au paragraphe précédent et sont définies comme suit :

- exactitude : la distance moyenne entre les dates des défaillances et les dates des défaillances prévues.
- précision : l'écart-type de la distance entre les dates des défaillances et les dates des défaillances prévues.

Puisque nous ne fournissons pas une unique date mais un intervalle $]t_0, t_e^f]$, nous proposons de calculer ces mesures de moyenne μ_{error} (cf. équation 10.2) et d'écart-type σ_{error} (cf. équation 10.3) sur les distances/erreurs entre t_r, date à laquelle la défaillance est observée, et les dates t_e^f, t_0 et $t_0 + maxgap$. Les dates t_e^f, t_0 et $t_0 + maxgap$ sont notées t_i dans les équations 10.3 et 10.2 avec N le nombre d'intervalles de prévision.

$$\mu_{error} = \frac{1}{N}\sum_{i=1}^{N}(t_r - t_i) \tag{10.2}$$

$$\sigma_{error} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_r - t_i)^2} \tag{10.3}$$

De façon à vérifier si les dates de défaillance prévues sont, en moyenne, avant ou après les dates de défaillance observées, $(t_r - t_i)$ est préféré à $|t_r - t_i|$ dans l'équation 10.2. Ces mesures sont rapportés dans les tableaux 10.4 et 10.5 pour chacun des jeux de données. Pour le dataset 1, la plus petite distance moyenne est obtenue pour $t_e^f = -0.85$ jours. Le même comportement est observé par l'écart-type avec 1.48 jours. La date t_e^f est donc la plus exacte et la plus précise concernant l'estimation de t_r. Pour le dataset 2, la date t_e^f reste la plus précise (2.15 jours) mais elle n'est pas la plus exacte. Dans ce cas, la date t_0 est une meilleure estimation de t_r avec une exactitude de 2.27 jours. Quel que ce soit le cas et le jeu de données, la date t_e^f et la date t_0 sont plus précises et exactes que $t_0 + maxgap$. La réduction de l'intervalle de prévision de $]t_0, t_0 + maxgap]$ à $]t_0, t_e^f]$ fait donc sens et valide le choix d'une approche au plus tôt.

TABLE 10.4 – Exactitude et précision pour le dataset 1.

	Date à laquelle t_r est comparée.		
	$t_s^f = t_0$	t_e^f	$t_0 + maxgap$
μ_{error} (jours)	2.06	-0.85	-4.88
σ_{error} (jours)	2.26	1.48	2.26

TABLE 10.5 – Exactitude et précision pour le dataset 2.

	Date à laquelle t_r est comparée.		
	t_s^f	t_e^f	$t_0 + maxgap$
μ_{error} (jours)	2.27	-3.14	-4.68
σ_{error} (jours)	2.69	2.15	2.69

De façon plus détaillée, pour le dataset 1, la distance par rapport t_0 est toujours positive et s'étale de 0 à 4 jours. Les grippages apparaissent après t_0 dans 100% des cas tandis que, dans 90% des cas, ils apparaissent juste avant t_e^f. De plus, comme déjà vu, dans la plus part des cas, la distance par rapport à t_e^f est inférieure à 1 jour. Par ailleurs, dans 97% des cas, $t_0 + maxgap > t_r$. Généralement, la durée qui sépare ces 2 dates se situe entre 4 et 6 jours. L'intervalle de prévision fourni est donc une estimation précise et exacte des dates d'occurrences des grippages. Cette estimation reste meilleure que l'intervalle $]t_0, t_0 + maxgap]$. Pour le dataset 2, la distribution reste similaire pour $t_0 + maxgap$. Dans la plus part des cas, $t_0 + maxgap$ se situe entre 5 et 7 jours après le grippage alors que t_e^f est postérieure de 2 à 4 jours (6 jours parfois) à la date du grippage.

À nouveau l'intervalle $]t_s^f = t_0, t_e^f]$ se révèle être un meilleur choix que $]t_0, t_0 + maxgap]$. Au final, les dernières prévisions arrivent au moins 3 heures avant défaillance, et, la plupart du temps, 2 jours avant. Ceci laisse suffisamment de temps pour planifier des opérations de maintenance.

Synthèse Bien que la méthode proposée dans la section 10.2 ait donné satisfaction, des réserves avaient été émises. La première concernait le fait que les expériences n'avaient pas été faites sur des données réelles, dans lesquelles des FLM-règles typiques, c'est-à-dire dont le taux de décroissance est supérieur 0%, peuvent être extraites. Dans cette section nous avons pu effectuer nos expériences sur des données réelles caractérisant le fonctionnement de pompes à vide et nous appuyer sur des FLM-règles typiques. La deuxième réserve concernait le mode d'apprentissage, considéré comme permissif en terme de généricité et de taux de fausses alarmes. Nous avons donc proposé une approche de type leave-one-out permettant de tester la généricité des règles extraites et d'écarter au maximum celles générant de fausses alarmes. Par ailleurs, la méthode de prévision proposée dans la section 10.2 s'appuyait sur une prévision au plus tard des événements, ce qui peut être une politique risquée selon le contexte. En s'appuyant sur la définition des FLM-règles, nous avons proposé un intervalle de prévision au plus tôt dont les expériences ont montré qu'il était précis et exact et qu'il permettait d'anticiper plus de 75% des grippages de pompes à vides. Il est à noter que les performances de notre méthode ont été comparées dans [MMGB12] à la méthode proposée par [CZC07] qui impose une même fenêtre pour l'apprentissage et les prévisions. Les résultats sont disponibles dans [MMGB12] : nous prédisons environ deux fois plus de grippages. Nos intervalles de prévision sont par ailleurs bien plus précis. En effet, la plus petite distance moyenne rapportée par rapport à la date de défaillance concerne t_e^f et dépasse 14 jours. Enfin, en ne recherchant que les dernières occurrences des prémisses dans une file restreinte à la taille de la plus grande fenêtre optimale extraite et en mémorisant les dates de prévision associées, nous évitons de conserver l'intégralité du flot de données et d'avoir à retrouver plusieurs fois une même prémisse contrairement à ce qui était proposé dans la section 10.2. Notre approche est aujourd'hui brevetée [BMP+10].

Chapitre 11

Conclusion et perspectives

Les travaux de fouille de données engagés au LISTIC s'appuient naturellement sur les contextes applicatifs et les compétences disponibles au sein du laboratoire. Bien qu'unique data miner du LISTIC, il nous a été possible de progresser et de dépasser le cadre du laboratoire en impliquant des chercheurs d'autres laboratoires, publics ou privés, qu'ils soient spécialisés en fouille de données ou experts des domaines applicatifs visés. En adoptant une démarche progressive et construite, basée sur une étape d'étude d'opportunité et une étape de proposition, deux axes de recherches ont pu émerger : la description non supervisée de séries temporelles d'images satellitaires et la prévision d'événements dans un flot de données.

En ce qui concerne la description non supervisée de séries temporelles d'images satellitaires, les travaux ont abouti à la proposition d'un nouveau type de motifs locaux, simple à interpréter, les motifs Séquentiels Fréquents Groupés ou motifs SFG. Ces motifs permettent d'extraire des groupes de pixels partageant une même évolution ou sous-évolution, couvrant une surface minimum et étant suffisamment connectés entre eux. La contrainte de connexité moyenne minimum associée permet de réduire efficacement l'espace de recherche et vient en appui de la contrainte de support/surface minimum lorsque les valeurs de cette dernière sont faibles. La conjonction des contraintes de connexité et de surface permet ainsi de limiter également le nombre de motifs extraits. Cependant, ce nombre peut, selon les jeux de données, rester élevé et nécessite alors la mise en œuvre d'un critère de maximalité permettant de se concentrer sur les motifs les plus spécifiques. Ce type de contrainte associée aux contraintes de connexité et de surface a l'avantage d'écarter, autant que possible, les phénomènes dus à l'incertitude aléatoire, en particulier les phénomènes liés aux perturbations atmosphériques et aux défauts d'acquisition. Ainsi, lors de nos différentes expériences, aucun des motifs interprétés par les utilisateurs finaux ne s'est avéré traduire un phénomène lié à des perturbations atmosphériques ou à des défauts de capteur. Ces expériences confirment également le caractère générique de l'approche : quelle que ce soit la résolution des images, qu'il s'agisse de données radar ou optiques, il est possible d'extraire des motifs SFG ayant un sens pour les utilisateurs finaux. Par ailleurs, quelle que soit l'application considérée, une fois choisie la bande d'intérêt (par exemple le NDVI pour les cultures ou les différences de phase pour les déplacements), un simple prétraitement à base de centiles suffit pour construire les STIS symboliques desquelles seront extraits ces motifs. Afin de guider l'utilisateur dans l'étude de la collection des motifs extraits, nous avons proposé de classer ces derniers en fonction de leurs cartes

de localisation spatio-temporelles calculées sur un jeu de données réelles et sur un jeu de données aléatoires obtenu par swap-randomisation, c'est-à-dire un jeu de données dans lequel la fréquence des symboles est conservée. Afin d'établir si ces cartes sont similaires, la mesure d'Information Mutuelle Normalisée (IMN) est utilisée. C'est également selon cette dernière que sont classés les motifs SFG. Cette méthode permet de séparer d'un côté les motifs exprimant des changements progressant au fil du temps dans l'espace et, d'un autre côté, les motifs évoquant le non changement ou ceux ne progressant pas dans l'espace au fil du temps. Ces derniers sont d'ailleurs considérés comme ayant une plus forte probabilité d'apparaître dans un jeu de données aléatoire où les fréquences des symboles sont conservées. Les deux types de motifs expriment des types de phénomènes différents vers lesquels il est possible d'orienter automatiquement l'utilisateur. Les premières expériences menées montrent le potentiel de l'approche en exhibant des phénomènes spatio-temporel intéressants et pertinents. La swap-randomisation utilisée est une proposition originale s'appuyant sur la swap-randomisation de matrices booléennes proposée dans [GMMT07].

Néanmoins, de nombreux éléments théoriques restent à apporter. L'atteignabilité de tous les jeux aléatoires possibles depuis n'importe quel jeu aléatoire doit être montrée comme l'a fait Ryser pour les matrices booléennes [Rys57]. Une fois cela effectué, il serait intéressant de montrer l'uniformité de processus, c'est-à-dire la garantie que chaque jeu aléatoire a la même chance d'être atteint. Cela pourrait peut être se faire en s'appuyant sur les chaînes de Markov comme proposé dans [GMMT07] pour les matrices booléennes. Une propriété de convergence serait également intéressante de façon à pouvoir calculer le nombre minimum d'échanges à faire pour que le jeu de données soit considéré comme suffisamment randomisé. Aujourd'hui, à notre connaissance, seules des méthodes empiriques sont proposées. Une autre piste, prometteuse au niveau des utilisateurs finaux, est la construction d'un seul clustering de séries temporelles d'images satellitaires à partir des motifs SFG en suivant une approche de type *Lego* comme proposé dans [KCFS08]. Le calcul/la caractérisation de champs de déplacements à partir de séries d'images, satellitaires ou non, nous semble également prometteur : de nombreux experts utilisent ces représentations pour observer et simuler des phénomènes en particulier géophysiques et géomécaniques. Dans ce cadre, de telles techniques peuvent être envisagées comme support à l'inversion de données. Aidé de collègues géophysiciens, géomécaniciens, traiteurs de signal et fouilleurs de données, nous rédigeons actuellement un projet ANR en ce sens que je serai amené à diriger si ce projet aboutit. Enfin, en vue de considérer de longues STIS (plus de 50 images), l'extraction de motifs SFG, à partir de représentations compactes construites à l'aide d'auto-encodeurs, constitue une approche intéressante. De telles représentations ont par exemple montré leur intérêt pour la classification de séquences vidéo (cf. [BMW$^+$12]).

En ce qui concerne la prévision d'événements dans un flot de données, nous avons proposé une approche de type leave-one-out permettant d'extraire des FLM-règles génériques, générant peu de fausses alarmes et dont l'information temporelle, c'est-à-dire les fenêtres temporelles, a permis de construire une prévision au plus tôt. Cette méthode, fournit, lorsque des événements sont prévus, des intervalles temporels de prévision. L'objectif poursuivi est le pilotage de systèmes complexes, particulièrement en anticipant les défaillances ou les pannes de ces derniers. Les expériences menées sur des pompes à vides pour le compte de la société ADIXEN ont montré que ces intervalles étaient précis et exacts, et qu'ils permettaient d'anticiper les défaillances tout en générant peu de fausses alarmes. Ces expériences ont également permis de découvrir de nouvelles connaissances relatives à

la cinématique des pompes. La méthode retenue donne de bien meilleures performances que la méthode proposée par [CZC07] qui impose une même fenêtre pour l'apprentissage et les prévisions. Notre approche est aujourd'hui brevetée [BMP+10].

Afin de proposer des alarmes progressives, il serait intéressant d'exploiter plus précisément la courbe de confiance associée à chaque FLM-règle et de déclencher des alarmes sur l'observation des parties des prémisses et non des prémisses entières. De même, considérer chaque événement comme pouvant appartenir à différents types d'événements en utilisant la théorie des ensembles flous permettrait peut être de gagner en généricité. Dans ce dernier cas, le risque à maîtriser est bien entendu lié au taux de fausses alarmes. Enfin, la question de la définition de motifs ou de modèles dédiés à la prévision est également posée. En effet, les FLM-règles ont été conçues dans un but de description et non d'inférence. Ces définitions permettraient par exemple de prendre en compte la nature dynamique des flots de données de façon à faire évoluer les modèles/motifs de prévision en fonction du temps. Cela constituait une des motivations premières du projet ARC *SÉcurité et SURveillance dans les flots de données (SÉSUR)* [INR], qui a été suivi du projet ANR *MIning DAta Streams (MIDAS)* [ENS], et dans lequel les motifs séquentiels et les chroniques [GCRR08] ont été identifiés comme étant prometteurs. D'autres motifs proposés dans [FMLT09], les motifs d'évolution et les tendances graduelles, constituent également des alternatives intéressantes. Ces motifs, basés sur les motifs séquentiels et les motifs graduels [AYLP10], permettent de prendre en compte la temporalité des phénomènes tout en traduisant des enchaînements de tendances.

Enfin, de façon plus globale et à plus long terme, trois axes de recherche se dessinent. Le premier axe concerne la prise en compte des erreurs et des incertitudes liées aux données. Les incertitudes aléatoires et systémiques doivent pouvoir être considérées et des propositions s'appuyant sur la logique floue comme [YMTP12] sont à étudier. Le deuxième axe concerne la définition d'une algèbre permettant d'utiliser les motifs locaux dans le cadre des bases de données inductives comme proposé dans [IM96]. Les langages existants sont effet peu flexibles quant à l'utilisation de motifs et contraintes définis par l'utilisateur final (cf. [BCF+10]). Le troisième et dernier axe, abordé en perspective de la prévision dans un flot de données, consiste en la description de systèmes dynamiques [FMLT09].

Chapitre 12

Acronymes

ADAM : Assimilation de Données pour l'Agro-Modélisation, projet du CNES.

ADIXEN : anciennement filiale d'Alcatel-Lucent, elle est depuis 2011 filiale du groupe Pfeiffer Vacuum Technology AG.

AEGIS : Ability Enlargement for Geophysicists and Information technology Specialists, projet européen.

ANR : Agence Nationale de la Recherche.

CNES : Centre National d'Études Spatiales.

D-InSAR : Differential Interferometric Synthetic Aperture Radar - Interférométrie RSO différentielle.

EFIDIR : Extraction et Fusion d'Informations pour la mesure de Déplacement par Imagerie Radar, projet ANR.

ENS : École Nationale Supérieure.

ENST : École Nationale Supérieure des Télécommunications, aujourd'hui Télécom ParisTech.

ENVISAT : ENVIronement SATellite - satellite de l'ESA lancé en 2002 doté d'un capteur RSO multi-polarisations en bande C.

ERS-1/2 : European Remote Sensing - satellites de l'ESA dotés d'un capteur RSO en bande C mono-polarisation (VV), lancés en 1991 (ERS-1) et 1995 (ERS-2) sur des orbites permettant l'acquisition de couples interférométriques à 1 jour (données tandem).

ESA : European Space Agency - Agence spatiale européenne.

EURINSA : premier cycle EURopéen de l'INSA de Lyon.

FLM : First Local Maximum.

FOSTER : FOuille de données Spatio-Temporelles : application à la compréhension et à la surveillance de l'ERosion, projet ANR.

IDVN - NDVI : Indice de Différence de Végétation Normalisé - Normalized Difference Vegetation Index.

IMN : Information Mutuelle Normalisée.

INFO (département INFO) : département INFOrmatique.

INRIA : Institut National de Recherche en Informatique et en Automatique.

INSA : Institut National des Sciences Appliquées.

InSAR : Interférométrie radar à synthèse d'ouverture.

IPGS : Institut de Physique du Globe.

ISIS (GdR) : Groupement de Recherche Information Signal Image Vision.

ISTerre : Institut des Sciences de la Terre.

IUT : Institut Universitaire de Technologie.

LANDSAT : programme spatial d'observation de la Terre dédié à des fins civiles développé par l'agence spatiale américaine, la NASA. Premier satellite lancé en 1972.

LGCIE : Laboratoire de Génie Civil et d'Ingénierie Environnementale (INSA de Lyon).

LIRIS : Laboratoire d'InfoRmatique en Image et Systèmes d'information (INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / École Centrale de Lyon).

LISTIC : Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (Université de Savoie).

MEGATOR : Mesure de l'Évolution des Glaciers Alpins par Télédétection Optique et Radar - projet soutenu par une ACI masse de données.

METEOSAT : famille de satellites météorologiques réalisés sous maîtrise d'œuvre de l'Agence spatiale européenne, premier satellite lancé en 1977.

MNE - DEM : Modèle Numérique d'Elévation - Digital Elevation Model.

SFG (motif) : motif Séquentiel Fréquent Groupé.

NASA : National Aeronautics and Space Administration, États-Unis d'Amérique.

PolSAR : polarimétrie radar à synthèse d'ouverture.

RADARSAT-2 : satellite canadien doté d'un capteur RSO en bande C, pleinement polarimétrique, lancé en 2007.

RSO - SAR : Radar à Synthèse d'Ouverture - Synthetic Aperture Radar.

SBAS : Small BAselines Subsets, technique D-InSAR.

SPATPAM : SPAtio-TemPorAl Mining, prototype du projet ANR EFIDIR dédié à l'extraction de motifs SFG.

SPOT : Satellites Pour l'Observation de la Terre, premier lancement en 1985.

STAMPS : Stanford Method for Persistent Scatterers, technique D-InSAR.

STIS : Série Temporelle d'Images Satellitaires.

TerraSAR-X : satellite allemand doté d'un capteur RSO en bande X, multi-polarisations (pleinement polarimétrique en mode expérimental), lancé en 2007.

UNESCO : United Nations Educational, Scientific and Cultural Organization.

Bibliographie

[AGS04] Atallah M., Gwadera R., Szpankowski W. *Detection of significant sets of episodes in event sequences*. **In** : *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04. IEEE Computer Society, Washington, DC, USA, 2004, pp. 3–10.

[ALS12] Achar A., Laxman S., Sastry P.S. *A unified view of the apriori-based algorithms for frequent episode discovery*. **In** : Knowledge Information Systems, vol. 31, 2, 2012, pp. 223–250.

[AS95] Agrawal R., Srikant R. *Mining sequential patterns*. **In** : Yu P.S., Chen A.S.P., Eds., *Proc. of the 11th International Conference on Data Engineering (ICDE'95)*. IEEE Computer Society Press, Taipei, Taiwan, 1995, pp. 3–14.

[ASBF+12] Alatrista Salas H., Bringay S., Flouvat F., et al. *The pattern next door : Towards spatio-sequential pattern discovery*. **In** : Tan P.N., Chawla S., Ho C., et al., Eds., *Advances in Knowledge Discovery and Data Mining*, *Lecture Notes in Computer Science*, vol. 7302. Springer Berlin Heidelberg, 2012, pp. 157–168.

[AYLP10] Ayouni S., Yahia S.B., Laurent A., et al. *Fuzzy gradual patterns : What fuzzy modality for what result ?*. **In** : Martin T.P., Muda A.K., Abraham A., et al., Eds., *SoCPaR*. IEEE, 2010, pp. 224–230.

[BBTD08] Bontemps S., Bogaert P., Titeux N., et al. *An object-based change detection method accounting for temporal dependences in time series with medium to coarse spatial resolution*. **In** : Remote Sensing of Environment, vol. 112, 2008, pp. 3181–3191.

[BCF+10] Blockeel H., Calders T., Fromont E., et al. *A practical comparative study of data mining query languages*. **In** : *Inductive Databases and Constraint-Based Data Mining*. Springer New York, 2010, pp. 59–77.

[BFLS02] Berardino P., Fornaro G., Lanari R., et al. *A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms*. **In** : IEEE Transactions on Geoscience and Remote Sensing, vol. 40, 2002, pp. 2375–2382.

[BL05] Bonchi F., Lucchese C. *Pushing tougher constraints in frequent pattern mining*. **In** : *Proc. of the Pacific-Asia Knowledge Discovery and Data Mining conference (PAKDD'05)*. 2005, pp. 114–124.

[BMP+10] Bécourt N., Martin F., Pariset C., et al. *Method for predicting a rotation fault in the rotor of a vacuum pump and associated pumping devices, Alcatel-Lucent patent*, 2010.

[BMW+12] Baccouche M., Mamalet F., Wolf C., et al. *Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification*. **In** : R. Bowden J.C., Mikolajczyk K., Eds., *British Machine Vision Conference (BMVC)*. BMVA Press, sept. 2012, pp. 124.1–124.1.

[Bro73] Broch J.T. *Application of B and K Equipment to Mechanical Vibraton and Shock Measurement*. Bruel & Kjaer, 1973.

[CG03] Casas-Garriga G. *Discovering unbounded episodes in sequential data*. **In** : *Proceedings of 7th Eur. COnf. on Principles and Practice of Knowledge Discovery in Databases (PKDD '03)*. Springer, 2003, pp. 83–94.

[CHCW05] Chen M.C., Huang C.L., Chen K.Y., et al. *Aggregation of orders in distribution centers using data mining*. **In** : Expert Systems with Applications, 2005, pp. 453–460.

[CJN+04] Coppin P., Jonckheere I., Nackaerts K., et al. *Digital change detection methods in ecosystem monitoring : a review*. **In** : International Journal of Remote Sensing, vol. 25, 9, May 2004, pp. 1565–1596.

[CMC05] Cao H., Mamoulis N., Cheung D.W. *Mining frequent spatio-temporal sequential patterns*. **In** : *ICDM '05 : Proceedings of the Fifth IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, 2005, pp. 82–89.

[CMC07] Cao H., Mamoulis N., Cheung D.W. *Discovery of periodic patterns in spatiotemporal sequences*. **In** : IEEE Transaction on Knowledge and Data Engineering, vol. 19, 4, 2007, pp. 453–467.

[CNE12] *Centre National d'Etudes Spatiales (CNES) : Database for the data assimilation for agro-modeling (adam) project*. Online, 2012. Http ://kalideos.cnes.fr/spip.php ?article21.

[CT91] Cover T., Thomas J. *Elements of information theory*. John Wiley & Sons, New York, 1991.

[CZC07] Cho C.W., Zheng Y., Chen A.L.P. *Continuously matching episode rules for predicting future events over event streams*. **In** : *Proceedings of the joint 9th Asia-Pacific web and 8th international conference on web-age information management conference on Advances in data and web management*, APWeb/WAIM'07. Springer-Verlag, Berlin, Heidelberg, 2007, pp. 884–891.

[DFP98] Doebling S., Farrar C., Prime M. *A summary review of vibration-based damage identification methods*. **In** : Shock and Vibration Digest, vol. 30, 2, 1998, pp. 91–105.

[EKGZ08] El Koujok M., Gouriveau R., Zerhouni N. *Towards a Neuro-Fuzzy System for time series forecasting in Maintenance Applications.*. **In** : *17th Triennal World Congress of the International Federation of Automatic Control, (IFAC'08)*. Elsevier, Seoul, Korea, 2008.

[ENS] *ANR project : MIning DAta Streams (MIDAS) (2008-2010). Partners : ENST, LIRMM, CEREGMIA, INRIA, EDF, and France Telecom. http ://www2.lirmm.fr/tatoo/spip.php ?article100*.

[FDHF+05] Fisher R., Dawson-Howe K., Fitzgibbon A., et al. *Dictionary of Computer Vision and Image Processing*. John Wiley and Sons, New York, 2005, 364 pp.

[FMLT09] Fiot C., Masseglia F., Laurent A., et al. *Evolution patterns and gradual trends*. **In** : Int. J. Intell. Syst., vol. 24, 10, October 2009, pp. 1013–1038.

[FPR01] Ferretti A., Prati C., Rocca F. *Permanent scatterers in SAR interferometry*. **In** : IEEE Transactions on Geoscience and Remote Sensing, vol. 39, 1, 2001, pp. 8–20.

[FPSM91] Frawley W.J., Piatetsky-Shapiro G., Matheus C.J. *Knowledge discovery in databases : An overview*. **In** : *Knowledge Discovery in Databases*. AAAI/MIT Press, Cambridge, MA, 1991, pp. 1–30.

[FPSS96] Fayyad U.M., Piatetsky-Shapiro G., Smyth P. *Knowledge discovery and data mining : Towards a unifying framework*. **In** : *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, Menlo Park, CA, 1996, pp. 82–88.

[GCRR08] Guillou X.L., Cordier M.O., Robin S., et al. *Chronicles for on-line diagnosis of distributed systems*. **In** : *ECAI*. 2008, pp. 194–198.

[GKS07] Gudmundsson J., Kreveld M., Speckmann B. *Efficient detection of patterns in 2D trajectories of moving points*. **In** : Geoinformatica, vol. 11, 2, 2007, pp. 195–215.

[GMC08] Gallucio L., Michel O., Comon P. *Unsupervised clustering on multi-components datasets : Applications on images and astrophysics data*. **In** : *16th European Signal Processing Conference EUSIPCO-2008*. Lausanne, Switzerland, august 2008, pp. 25–29.

[GMMT07] Gionis A., Mannila H., Mielikäinen T., et al. *Assessing data mining results via swap randomization*. **In** : ACM Transaction on Knowledge Discovery from Data, vol. 1, 3, déc. 2007.

[GRK99] Garofalakis M., Rastogi R., K. S. *Spirit : Sequential pattern mining with regular expression constraints*. **In** : *Proc. of the 25th International Conference on Very Large Databases (VLDB'99)*. Edinburgh, United Kingdom, September 1999, pp. 223–234.

[HD05] Heas P., Datcu M. *Modelling trajectory of dynamic cluster in image-time-series for spatio-temporal reasoning*. **In** : IEEE Transactions on Geoscience and Remote Sensing, vol. 43, 7, 2005, pp. 1635–1647.

[HK01] Honda R., Konishi O. *Temporal rule discovery for time-series satellite images and integration with RDB*. **In** : *PKDD '01 : Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*. Springer-Verlag, London, UK, 2001, pp. 204–215.

[HKM+96] Hatonen K., Klemettinen M., Mannila H., et al. *TASA : Telecomunications Alarm Sequence Analyzer or : How to enjoy faults in your network*. **In** : *1996 IEEE Network Operations and Management Symposium (NOMS'96)*. Kyoto, Japan, April 1996, pp. 520–529.

[HMS01] Hand D.J., Mannila H., Smyth P. *Principles of Data Mining*. Bradford, MIT Press, Cambridge, MA, 2001.

[Hoo08] Hooper A. *A multi-temporal InSAR method incorporating both persistent scatterer and small baseline approaches*. **In** : Geophysical Research Letters, vol. 35, 5 pages, 2008.

[HZZ08] Huang Y., Zhang L., Zhang P. *A framework for mining sequential patterns from spatio-temporal event data sets*. **In** : IEEE Transactions on Knowledge and Data Engineering, vol. 20, 4, 2008, pp. 433–448.

[IFY+03] Inglada J., Favard J.C., Yesou H., et al. *Lava flow mapping during the Nyiragongo January, 2002 eruption over the city of Goma (D.R. Congo) in the frame of the international charter space and major disasters*. **In** : *In proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS03)*, vol. 3. july 2003, pp. 1540–1542.

[IM96] Imielinski T., Mannila H. *A database perspective on knowledge discovery*. **In** : Commun. ACM, vol. 39, 11, nov. 1996, pp. 58–64.

[INR] *INRIA ARC project : SÉcurité et SURveillance dans les flots de données (SÉSUR) (2007-2008). Teams : INRIA AxIS, IRISA Dream, LGI2P/EMA KDD, and LIRMM TATOO. http ://www-sop.inria.fr/axis/sesur/0708/index.htm*.

[ISO04] *Condition monitoring and diagnostics of machines prognostics - Part1 : General guidelines, ISO 13381-1*, 2004.

[IYN04] Iwanuma K., Y. T., Nabeshima H. *On anti-monotone frequency measures for extracting sequential patterns from a single very-long sequence*. **In** : *In Proceedings of IEEE conference cybernetics and intelligent systems*. December 2004, pp. 213–217.

[JLD+12] Jolivet R., Lasserre R., Doin M.P., et al. *Shallow creep on the haiyuan fault (gansu, china) revealed by sar interferometry*. **In** : Journal of Geophysical Research, 2012.

[JLM+11] Julea A., Ledo F., Méger N., et al. *Polsar radarsat-2 satellite image time series mining over the chamonix mont-blanc test site*. **In** : *proc. of IEEE Int. Geoscience And Remote Sensing Symposium (IGARSS' 11)*. Vancouver, Canada, July 2011, pp. 1191–1194.

[JMB08] Julea A., Méger N., Bolon P. *On mining pixel based evolution classes in satellite image time series*. **In** : *Proc. of the 5th Conference on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008)*. ESRIN - Frascati, Italy, 6 pages, March 2008.

[JMB+11] Julea A., Méger N., Bolon P., et al. *Unsupervised spatiotemporal mining of satellite image time series using grouped frequent sequential patterns*. **In** : IEEE Transactions on Geoscience and Remote Sensing, vol. 49, 4, 2011, pp. 1417–1430.

[JMR+10] Julea A., Méger N., Rigotti C., et al. *Extraction of frequent grouped sequential patterns from satellite image time series*. **In** : *Proc. of the IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS'10)*. Honolulu, Hawaii, USA, July 2010, pp. 3434–3437.

[JMR+11] Julea A., Méger N., Rigotti C., et al. *Mining pixel evolutions in satellite image time series for agricultural monitoring*. **In** : *In Advances in Data Mining. Applications and Theoretical Aspects - 11th Industrial Conference on Data Mining ICDM 2011*. New-York, USA, August 2011, pp. 189–203.

[JMR+12] Julea A., Méger N., Rigotti C., et al. *Efficient Spatiotemporal Mining of Satellite Image Time Series for Agricultural Monitoring*. **In** : Transactions on Machine Learning and Data Mining, vol. 5, 1, 2012, pp. 23–45.

[JMT+] Julea A., Méger N., Trouvé E., et al. *Spatio-temporal mining of polsar satellite image time series*. **In** : *ESA Living Planet Symposium - The 2010 European Space Agency Living Planet Symposium*. Bergen, Norway, 6 pages.

[JMT06a] Julea A., Méger N., Trouvé E. *On mining METEOSAT and ERS multitemporal images*. **In** : *Proceedings of the 4th Conference on Image Information Mining for Security and Intelligence (ESA-EUSC 2006)*. Madrid, Spain, 6 pages, November 2006.

[JMT06b] Julea A., Méger N., Trouvé E. *Sequential patterns extraction in multitemporal satellite images*. **In** : *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'06), Practical Data Mining Workshop : Applications, Experiences and Challenges*. Berlin, Germany, September 2006, pp. 94–97.

[JMTB08] Julea A., Méger N., Trouvé E., et al. *On extracting evolutions from satellite image time series*. **In** : *Proc. of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS' 08)*, vol. 5. Boston, MA, USA, 2008, pp. 228–231.

[Jul11] Julea A. Ph.D. thesis, Université de Grenoble - Université Politehnica din Bucaresti, Bucharest, Romania, September 2011.

[KCFS08] Knobbe A.J., Crémilleux B., Fürnkranz J., et al. *From local patterns to global models : The lego approach to data mining*. **In** : *In Proc. of From Local Patterns to Global Models : Proceedings of the ECML/PKDD-08 Workshop (LeGo-08)*. 2008, pp. 1–16.

[Lal99] Lalanne C. *Vibrations aléatoires*. Hermes Science, 1999.

[Lax06] Laxman S. *Discovering frequent episodes : fast algorithms, connections with HMMs and generalizations*. Ph.D. thesis, Bangalore University, 2006.

[LFM99] Letourneau S., Famili F., Matwin S. *Data mining for prediction of aircraft component replacement*. **In** : IEEE Intelligent Systems and their Applications, vol. 14, 6, december 1999, pp. 59–66.

[LK00] Lillesand T., Kiefer R. *Remote Sensing and Image Interpretation*. 4th edn. John Wiley and Sons, New York, 2000.

[LL01] Li L., Leung M. *Robust change detection by fusing intensity and texture differences*. **In** : *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, vol. 1. 2001, pp. 777–784.

[LMBM04] Lu D., Mausel P., Brondizio E., et al. *Change detection techniques*. **In** : International Journal of Remote Sensing, vol. 25, 12, juin 2004, pp. 2365–2407.

[LSU04] Laxoman S., Sastry P., Unnikrishnan K. *Fast algorithms for frequet episode discovery in event sequences*. Tech. rep., GM R&D Center, Warren, 2004. Technical Report CL-2004-04/MSR.

[LSU05] Laxman S., Sastry P.S., Unnikrishnan K.P. *Discovering frequent episodes and learning hidden markov models : A formal connection*. **In** : IEEE Trans. on Knowl. and Data Eng., vol. 17, 11, nov. 2005, pp. 1505–1517.

[LSU07] Laxman S., Sastry P.S., Unnikrishnan K. *A fast algorithm for finding frequent episodes in event streams*. **In** : Berkhin P., Caruana R., Wu X., Eds., *KDD*. 2007, pp. 410–419.

[LTW08] Laxman S., Tankasali V., White R.W. *Stream prediction using a generative model based on frequent episodes in event sequences*. **In** : *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08), Las Vegas, Nevada, USA, August 24-27, 2008*. 2008, pp. 453–461.

[MCP98] Masseglia F., Cathala F., Poncelet P. *The PSP approach for mining sequential patterns*. **In** : *Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery in Databases (PKDD'98)*, vol. 1510. LNAI, Springer Verlag, Nantes, France, September 1998, pp. 176–184.

[MJL+11] Méger N., Jolivet R., Lasserre C., et al. *Spatiotemporal mining of envisat sar interferogram time series over the haiyuan fault in china*. **In** : *Analysis of Multi-temporal Remote Sensing Images (Multi-Temp), 2011 6th International Workshop on the*. july 2011, pp. 137 –140.

[MLLR04] Méger N., Leschi C., Lucas N., et al. *Mining episode rules in stulong dataset*. **In** : *In Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04), Discovery Challenge*. Pise, Italy, September 2004, pp. 33–45.

[MMGB10a] Martin F., Méger N., Galichet S., et al. *Episode rule-based prognosis applied to complex vacuum pumping systems using vibratory data*. **In** : *Proceedings of the 10th industrial conference on Advances in data mining : applications and theoretical aspects*, ICDM'10. Springer-Verlag, Berlin, Heidelberg, 2010, pp. 376–389.

[MMGB10b] Martin F., Méger N., Galichet S., et al. *Data-driven prognosis applied to complex vacuum pumping systems*. **In** : *Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems - Volume Part I*, IEA/AIE'10. Springer-Verlag, Berlin, Heidelberg, 2010, pp. 468–477.

[MMGB12] Martin F., Méger N., Galichet S., et al. *Forecasting failures in a data stream context : Application to vacuum pumping system prognosis*. **In** : Transactions on Machine Learning and Data Mining, vol. 5, 2, October 2012, pp. 87–116. ISSN 1865-6781.

[MP01] Magoulas G., Prentza A. *Machine learning in medical applications*. **In** : Machine Learning and Its Applications, vol. 2049, 2001, pp. 300–307.

[MR04] Méger N., Rigotti C. *Constraint-based mining of episode rules and optimal window sizes*. **In** : *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD '04. Springer-Verlag New York, Inc., New York, NY, USA, 2004, pp. 313–324.

[MRG+12] Méger N., Rigotti C., Gueguen L., et al. *Normalized mutual information-based ranking of spatio-temporal maps*. **In** : *In Proc. of the 8th Conf. on Image Information Mining : Knowledge Discovery from Earth Observation Data (ESA-EUSC 2012)*. German Aerospace Centre (DLR), Oberpfaffenhofen, Germany, 4 pages, October 2012.

[MSI08] Muller A., Suhner M.C., Iung B. *Formalisation of a new prognosis model for supporting proactive maintenance implementation on industrial system*. **In** : Reliability Engineering and System Safety, vol. 93, 2, 2008, pp. 234–253.

[MT96] Mannila H., Toivonen H. *Discovering generalized episodes using minimal occurrences*. **In** : *In Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining (KDD'96)*. AAAI Press, 1996, pp. 146–151.

[MTIV97] Mannila H., Toivonen H., Inkeri Verkamo A. *Discovery of frequent episodes in event sequences*. **In** : Data Mining and Knowledge Discovery, vol. 1, 3, jan. 1997, pp. 259–289.

[MTV95] Mannila H., Toivonen H., Verkamo I. *Discovering frequent episodes in sequences*. **In** : *Proc. of the 1st International Conference on Knowledge Discovery and Data Mining (KDD'95)*. AAAI Press, Montreal, Canada, August 1995, pp. 210–215.

[NGSR96] Nezry E., Genovese G., Solaas G., et al. *ERS - Based early estimation of crop areas in Europe during winter 1994-95*. **In** : T.-D. G., Ed., *ERS Application, Proceedings of the Second International Workshop held 6-8 December 1995 in London*, *ESA Special Publication*, vol. 383. 1996, p. 13.

[NLHP98] Ng R.T., Lakshmanan L.V.S., Han J., et al. *Exploratory mining and pruning optimizations of constrained associations rules*. **In** : *Proc. of the ACM SIGMOD international conference on Management of data (SIGMOD'98)*. Seattle, Washington, USA, 1998, pp. 13–24.

[NP06] Nanni M., Pedreschi D. *Time-focused clustering of trajectories of moving objects*. **In** : Journal of Intelligent Information Systems, vol. 27, 3, 2006, pp. 267–289.

[PHL01] Pei J., Han J., Lakshmanan L.V. *Mining frequent itemsets with convertible constraints*. **In** : *Proc. of the 17th International Conference on Data Engineering (ICDE'01)*. Heidelberg, Germany, 2001, pp. 433–442.

[PHMAP01] Pei J., Han B., Mortazavi-Asl B., et al. *Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth*. **In** : *Proc. of the 17th International Conference on Data Engineering (ICDE'01)*. 2001, pp. 215–226.

[PHW07] Pei J., Han J., Wang W. *Constraint-based sequential pattern mining : the pattern-growth methods*. **In** : Journal of Intelligent Information Systems, vol. 28, 2, 2007, pp. 133–160.

[PIG12] Petitjean F., Inglada J., Gançarski P. *Satellite image time series analysis under time warping*. **In** : IEEE T. Geoscience and Remote Sensing, vol. 50, 8, 2012, pp. 3081–3095.

[PMGF11] Petitjean F., Masseglia F., Gançarski P., et al. *Discovering significant evolution patterns from satellite image time series*. **In** : Int. J. Neural Syst., vol. 21, 6, 2011, pp. 475–489.

[PVB+11] Pitarch Y., Vintrou E., Badra F., et al. *Mining sequential patterns from modis time series for cultivated area mapping*. **In** : Geertman S., Reinhardt W., Toppen F., Eds., *Advancing Geoinformation Science for a Changing World*, Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, 2011, pp. 45–62.

[Rys57] Ryser H.J. *Combinatorial properties of matrices of zeros and ones*. **In** : Canad. J. Math, vol. 9, 1957, pp. 371–377.

[SA96] Srikant R., Agrawal R. *Mining sequential patterns : Generalizations and performance improvements*. **In** : *Proc. of the 5th International Conference on Extending Database Technology (EDBT'96)*. Avignon, France, September 1996, pp. 3–17.

[SG07] Schwabacher M., Goebel K. *A Survey of Artificial Intelligence for Prognostics*. **In** : *Working Notes of 2007 American Institute in Aeronautics and Astronautics Fall Symposium : AI for Prognostics*. 2007.

[SKM03] Symeonidis A.L., Kehagias D.D., Mitkas P.A. *Intelligent policy recommendations on enterprise resource planning by the use of agent technology and data mining techniques*. **In** : Expert Systems with Applications, vol. 25, 2003, pp. 589–602.

[TSK05] Tan P.N., Steinbach M., Kumar V. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[VERR04] Vina A., Echavarria, R. F., et al. *Satellite change detection analysis of deforestation rates and patterns along the colombia-ecuador border*. **In** : AMBIO : A Journal of the Human Environment, vol. 33, 2004, pp. 118–125.

[XPY12] Xing Z., Pei J., Yu P.S. *Early classification on time series*. **In** : Knowledge Information Systems, vol. 31, 1, avr. 2012, pp. 105–127.

[YMTP12] Yan Y., Mauris G., Trouvé E., et al. *Fuzzy uncertainty representations of coseismic displacement measurements issued from sar imagery*. **In** : IEEE T. Instrumentation and Measurement, vol. 61, 5, 2012, pp. 1278–1286.

[Zak00] Zaki M. *Sequence mining in categorical domains : incorporating constraints*. **In** : *Proc. of the 9th International Conference on Information and Knowledge Management (CIKM'00)*. Washington, DC, USA, November 2000, pp. 422–429.

[Zak01] Zaki M. *Spade : an efficient algorithm for mining frequent sequences*. **In** : Machine Learning, Special issue on Unsupervised Learning, vol. 42, 1/2, Jan/Feb 2001, pp. 31–60.

[ZGP10] Zemouri R., Gouriveau R., Patic P. *Combining a recurrent neural network and a PID controller for prognostic purpose : a way to improve the accuracy of predictions*. **In** : WSEAS Transactions on Systems and Control, vol. 5, 5, 2010, pp. 353–371.

Résumé

Les travaux présentés concernent l'extraction de connaissances dans les données à des fins de description et d'inférence. Comment décrire des Séries Temporelles d'Images Satellitaire (STIS) en mode non supervisé ? Comment prévoir des événements tels que des pannes dans des systèmes complexes ? Des réponses originales s'appuyant sur des techniques de fouille de données extrayant des motifs locaux, les motifs séquentiels, sont développées. Ainsi, de nouveaux motifs, les motifs Séquentiels Fréquents Groupés (motifs SFG), sont-ils proposés afin d'extraire d'une STIS des groupes de pixels faisant sens spatialement et temporellement. Une technique originale permettant de pousser les contraintes associées à ces motifs au sein du processus d'extraction est également détaillée. Des expériences sur des données optiques et radar, à des résolutions différentes, confirment leur potentiel. Un classement de ces motifs basé sur l'information mutuelle et la swap-randomization est par ailleurs proposé afin de mettre en avant les motifs ayant peu de chances d'apparaître dans un jeu de données aléatoires où les fréquences sont conservées, exprimant des changements et progressant dans l'espace. Quant à la prévision d'événements, une approche de type leave-one-out est proposée pour sélectionner des motifs séquentiels, les FLM-règles, génériques et déclenchant le moins possible de fausses alarmes. Une méthode de prévision au plus tôt tirant parti de ces motifs est également avancée et validée sur des données réelles provenant de systèmes mécaniques complexes. Les expériences menées montrent qu'il est possible de prévoir des défaillances pour lesquelles l'expertise technique est insuffisante. Cette méthode de prévision est aujourd'hui brevetée.

Mots-Clés

fouille de données, motifs séquentiels, connexité, push partiel, information mutuelle, swap-randomization, séries temporelles d'images satellitaires, règles d'épisodes, FLM-règles, prévision d'événements, pronostic.

Abstract

The work presented concerns knowledge discovery in databases with the aim of improving both the description of the database and inference from the database. How can a Satellite Image Times Series (SITS) be described in an unsupervised way ? How can events, such as failures affecting complex systems, be forecast ? Original answers based on local pattern-based data mining techniques have been developed. A new pattern type, Grouped Frequent Sequential patterns (GFS-patterns), is proposed to extract pixel groups that make sense both spatially and temporally. An original technique for pushing the constraints relating to these patterns is also detailed. Experiments on optical and radar data, at different resolutions, show the potential of these new techniques. Furthermore, a mutual information-based ranking, built using swap-randomization, is also proposed to highlight patterns that show changes while moving through space and that would not occur in a randomized data set in which symbol frequencies are maintained. With regard to forecasting events, a leave-one-out approach is proposed to select sequential patterns, namely FLM-rules, that are general and that trigger as few false alarms as possible. An early warning forecast method using these patterns is also defined and validated on real data originating from complex mechanical systems. Experiments show that it is possible to forecast failures for which technical expertise is missing. This forecasting method is now patented.

Key words

Data mining, sequential patterns, connectivity, partial push, mutual information, swap-randomization, satellite image time series, episode rules, FLM-rules, event forecasting, prognostic.