Prise en compte de la qualité des données lors de l'extraction et de la sélection d'évolutions dans les séries temporelles de champs de déplacements en imagerie satellitaire

Tuan Nguyen

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel :

Présentée par

Hoang Viet Tuan NGUYEN

Thèse dirigée par **Nicolas MÉGER**
codirigée par **Christophe RIGOTTI, Catherine POTHIER, Emmanuel TROUVÉ**

préparée au sein **du Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (LISTIC)**
et de **l'École Doctorale Sciences et Ingénierie des Systèmes, de l'Environnement et des Organisations (SISEO)**
et comme membre **du Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS)**

Prise en compte de la qualité des données lors de l'extraction et de la sélection d'évolutions dans les séries temporelles de champs de déplacements en imagerie satellitaire

Thèse soutenue publiquement le **10 octobre 2018**,
devant le jury composé de :

Mme. Florence TUPIN
Professeure des Universités, Télécom ParisTech, Présidente

Mme. Élisa FROMONT
Professeure des Universités, Université de Rennes 1, Rapportrice

M. Bruno CRÉMILLEUX
Professeur des Universités, Université de Caen Normandie, Rapporteur

M. Dino IENCO
Chargé de recherche, IRSTEA, Examinateur

M. Nicolas MÉGER
Maître de Conférences HDR, Université Savoie Mont Blanc, Directeur de thèse

M. Christophe RIGOTTI
Maître de Conférences HDR, INSA de Lyon, Co-Directeur de thèse

Mme. Catherine POTHIER
Maître de Conférences, INSA de Lyon, Co-Directrice de thèse

M. Emmanuel TROUVÉ
Professeur des Universités, Université Savoie Mont Blanc, Co-Directeur de thèse

Le travail de Hoang Viet Tuan NGUYEN est soutenu financièrement par la Région Auvergne-Rhône-Alpes.

Remerciements

Cette thèse a été menée au sein du Laboratoire d'Informatique, Systèmes et Traitement de l'Information et de la Connaissance (LISTIC) et du Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS). Mes premiers remerciements vont donc tout naturellement à ces deux institutions.

Je voudrais remercier grandement M. Nicolas MÉGER, mon directeur de thèse qui s'est investi sans compter pour que ce projet de recherche soit de qualité. Je suis ravi d'avoir travaillé en sa compagnie car outre son appui scientifique, il a toujours été là pour me soutenir et me conseiller au cours de l'élaboration de cette thèse.

Je remercie mes co-directeurs de thèse, M. Christophe RIGOTTI, Mme. Catherine POTHIER et M. Emmanuel TROUVÉ pour leurs encadrements continus et pour leurs précieux conseils durant toute la période de ma thèse.

J'adresse tous mes remerciements à Mme. Élisa FROMONT et M. Bruno CRÉMILLEUX de l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de cette thèse.

J'exprime ma gratitude à Mme. Florence TUPIN et à M. Dino IENCO, qui ont bien voulu être examinateurs.

Je remercie également M. Noël GOURMELEN, M. Jean-Louis MUGNIER pour des collaborations dans divers expériences menées pendant la thèse.

Je tiens à adresser mes remerciements aux membres du LISTIC que j'ai pu côtoyer tout au long de cette thèse. Une agréable équipe de chercheurs, d'enseignants, de personnels et de doctorants avec laquelle j'ai passé des moments inoubliables.

Enfin, ma gratitude et mes très sincères remerciements s'adressent à ma famille, qui m'a toujours encouragé, et soutenu dans les moments difficiles.

Résumé

Ce travail de thèse traite de la découverte de connaissances à partir de Séries Temporelles de Champs de Déplacements (STCD) obtenues par imagerie satellitaire. De telles séries occupent aujourd'hui une place centrale dans l'étude et la surveillance de phénomènes naturels tels que les tremblements de terre, les éruptions volcaniques ou bien encore le déplacement des glaciers. En effet, ces séries sont riches d'informations à la fois spatiales et temporelles et peuvent aujourd'hui être produites régulièrement à moindre coût grâce à des programmes spatiaux tels que le programme européen Copernicus et ses satellites phares Sentinel.

Nos propositions s'appuient sur l'extraction de motifs Séquentiels Fréquents Groupés (SFG). Ces motifs, à l'origine définis pour l'extraction de connaissances à partir des Séries Temporelles d'Images Satellitaires (STIS), ont montré leur potentiel dans de premiers travaux visant à dépouiller une STCD. Néanmoins, ils ne permettent pas d'utiliser les indices de confiance intrinsèques aux STCD et la méthode de *swap* randomisation employée pour sélectionner les motifs les plus prometteurs ne tient pas compte de leurs complémentarités spatiotemporelles, chaque motif étant évalué individuellement.

Notre contribution est ainsi double. Une première proposition vise tout d'abord à associer une mesure de fiabilité à chaque motif en utilisant les indices de confiance. Cette mesure permet de sélectionner les motifs portés par des données qui sont en moyenne suffisamment fiables. Nous proposons un algorithme correspondant pour réaliser les extractions sous contrainte de fiabilité. Celui-ci s'appuie notamment sur une recherche efficace des occurrences les plus fiables par programmation dynamique et sur un élagage de l'espace de recherche grâce à une stratégie de *push* partiel, ce qui permet de considérer des STCD conséquentes. Cette nouvelle méthode a été implémentée sur la base du prototype existant SITS-P2miner, développé au sein du LISTIC et du LIRIS pour extraire et classer des motifs SFG.

Une deuxième contribution visant à sélectionner les motifs les plus prometteurs est également présentée. Celle-ci, basée sur un critère informationnel, permet de prendre en compte à la fois les indices de confiance et la façon dont les motifs se complètent spatialement et temporellement. Pour ce faire, les indices de confiance sont interprétés comme des probabilités, et les STCD comme des bases de données probabilistes dont les distributions ne sont que partielles. Le gain informationnel associé à un motif est alors défini en fonction de la capacité de ses occurrences à compléter/affiner les distributions caractérisant les données. Sur cette base, une heuristique est proposée afin de sélectionner des motifs informatifs et complémentaires. Cette méthode permet de fournir un ensemble de motifs faiblement redondants et donc plus faciles à interpréter que ceux fournis par *swap* randomisation. Elle a été implémentée au sein d'un prototype dédié.

Les deux propositions sont évaluées à la fois quantitativement et qualitativement en utilisant une STCD de référence couvrant des glaciers du Groenland construite à partir de données optiques Landsat. Une autre STCD que nous avons construite à partir de données radar TerraSAR-X couvrant le massif du Mont-Blanc est également utilisée. Outre le fait d'être construites à partir de données et de techniques de télédétection différentes, ces séries se différencient drastiquement en termes d'indices de confiance, la série couvrant le massif du Mont-Blanc se situant à des niveaux de confiance très faibles. Pour les deux STCD, les méthodes proposées ont été mises en œuvre dans des conditions standards au niveau consommation de ressources (temps, espace), et les connaissances des experts sur les zones étudiées ont été confirmées et complétées.

Mots-clés — Fouille de données, Motif séquentiel, Série Temporelle de Champs de Déplacements (STCD), Qualité des données, Surveillance de glaciers

Abstract

This PhD thesis deals with knowledge discovery from Displacement Field Time Series (DFTS) obtained by satellite imagery. Such series now occupy a central place in the study and monitoring of natural phenomena such as earthquakes, volcanic eruptions and glacier displacements. These series are indeed rich in both spatial and temporal information and can now be produced regularly at a lower cost thanks to spatial programs such as the European Copernicus program and its famous Sentinel satellites.

Our proposals are based on the extraction of grouped frequent sequential patterns. These patterns, originally defined for the extraction of knowledge from Satellite Image Time Series (SITS), have shown their potential in early work to analyze a DFTS. Nevertheless, they cannot use the confidence indices coming along with DFTS and the swap method used to select the most promising patterns does not take into account their spatiotemporal complementarities, each pattern being evaluated individually.

Our contribution is thus double. A first proposal aims to associate a measure of reliability with each pattern by using the confidence indices. This measure allows to select patterns having occurrences in the data that are on average sufficiently reliable. We propose a corresponding constraint-based extraction algorithm. It relies on an efficient search of the most reliable occurrences by dynamic programming and on a pruning of the search space provided by a partial push strategy. This new method has been implemented on the basis of the existing prototype SITS-P2miner, developed by the LISTIC and LIRIS laboratories to extract and rank grouped frequent sequential patterns.

A second contribution for the selection of the most promising patterns is also made. This one, based on an informational criterion, makes it possible to take into account at the same time the confidence indices and the way the patterns complement each other spatially and temporally. For this aim, the confidence indices are interpreted as probabilities, and the DFTS are seen as probabilistic databases whose distributions are only partial. The informational gain associated with a pattern is then defined according to the ability of its occurrences to complete/refine the distributions characterizing the data. On this basis, a heuristic is proposed to select informative and complementary patterns. This method provides a set of weakly redundant patterns and therefore easier to interpret than those provided by swap randomization. It has been implemented in a dedicated prototype.

Both proposals are evaluated quantitatively and qualitatively using a reference DFTS covering Greenland glaciers constructed from Landsat optical data. Another DFTS that we built from TerraSAR-X radar data covering the Mont-Blanc massif is also used. In addition to being constructed from different data and remote sensing techniques, these series differ drastically in terms of confidence indices, the series covering the Mont-Blanc massif being at very low levels of confidence. In both cases, the proposed methods operate under standard conditions of resource consumption (time, space), and experts' knowledge of the studied areas is confirmed and completed.

Index terms — Data mining, Sequential pattern, Displacement Field Time Series (DFTS), Data quality, Glacier monitoring

Table des matières

Chapitre 1

Introduction

1.1 Les séries temporelles d'images satellitaires et de champs de déplacements

Les données d'origine satellitaire ne cessent de croître en termes de qualité et de volume grâce au développement continu depuis les années 70 des Satellites d'Observation de la Terre (SOT). De nombreux pays disposent aujourd'hui de SOT, qui acquièrent des données avec une résolution spatiale de plus en plus fine et des modes d'acquisition variés (multi/hyperspectral en optique, polarimétrique en radar, ...). La dimension temporelle est également à l'origine de cette inflation avec des fréquences d'observation de plus en plus élevées. Ces instruments de télédétection spatiale permettent ainsi d'obtenir des séries d'images avec une bonne répétitivité et une couverture globale. Cette dernière est un atout majeur par rapport aux mesures *in situ* comme les balises *Global Positioning System* (GPS), et aux images optiques terrestres ou aéroportées. À titre d'exemple, dans le cadre du programme Copernicus de l'Agence spatiale européenne (en anglais *European Space Agency*) (ESA), le lancement du satellite Sentinel-1A en 2014, suivi par son jumeau Sentinel-1B en 2016, permet d'avoir des observations sur n'importe quelle zone de la Terre, tous les 6 jours en Europe et sur des sites "à risque", et tous les 12 jours ailleurs. Ces satellites, disposant de capteurs actifs de type Radar à Synthèse d'Ouverture (en anglais *Synthetic Aperture Radar*) (SAR), permettent d'observer la Terre de jour comme de nuit, et ceci dans n'importe quelle condition météorologique. Plusieurs domaines de recherche peuvent bénéficier de cette mission, par exemple, la surveillance des surfaces de la Terre pour les risques liés aux glissements de terrain, aux éruptions volcaniques ou bien encore aux déplacements des glaciers. Du côté des satellites optiques, la série des satellites Landsat, développée par l'Administration nationale de l'aéronautique et de l'espace des États-Unis (en anglais *National Aeronautics and Space Administration*) (NASA), est un exemple patent de continuité d'observation. Depuis 1972, 8 satellites se sont succédés, fournissant ainsi plusieurs millions d'images multispectrales. L'instrument principal embarqué sur son tout dernier satellite Landsat-8, nommé *Operational Land Imager* (OLI), acquiert des images en neuf bandes spectrales allant du visible au moyen infrarouge. Ces images constituent des ressources uniques pour l'étude des changements climatiques, la surveillance de l'utilisation des sols (urbanisation, déforestation), la cartographie, le suivi des cultures ou bien encore l'évaluation des dégâts après un tremblement de terre ou une inondation. Par ailleurs, grâce aux politiques actuelles des agences spatiales internationales qui vont vers la libre distribution des données satellitaires (e.g., données Landsat de la NASA depuis 2009 ou données Sentinel de l'ESA depuis 2014), la contrainte liée à leur coût d'obtention a été levée et de plus en plus d'études scientifiques s'appuient sur les observations disponibles.

La finalité applicative de nos travaux est l'étude des évolutions des vitesses d'écoulement à la surface des glaciers, qui sont des marqueurs importants dans le contexte du changement climatique (Strozzi *et al.*, 2008; Akbari *et al.*, 2014; Dehecq *et al.*, 2015). Ce type de travaux s'appuie de plus en plus souvent sur différentes techniques permettant d'estimer les déplacements à partir des images satellitaires (e.g., Hanssen (2001); Vernier *et al.* (2011)). Plus précisément, une Série Temporelle d'Images Satellitaires (STIS), qui couvre une même zone géographique à différentes dates, est d'abord construite. Celle-ci caractérise tout point de la zone observée à chaque date d'acquisition par un scalaire ou un nombre complexe représentant la mesure des ondes électromagnétiques réfléchies par la Terre, ces ondes pouvant être émises par le soleil (images optiques) ou le satellite lui-même (images radar). Sur la base d'une STIS, il est ensuite possible de construire une Série Temporelle de Champs de

Déplacements (STCD) dont chacun des champs exprime, en tout point de la zone observée, des déplacements estimés entre deux dates. En plus des mesures de déplacement qui peuvent être des scalaires ou des vecteurs, ces séries contiennent généralement des indices exprimant la confiance associée à chaque mesure de déplacement. Ces indices peuvent être obtenus en s'appuyant sur des informations liées au calcul des champs lui-même (e.g., Giles *et al.* (2009)) ou à la cohérence spatiotemporelle des champs produits (e.g., Dehecq *et al.* (2015)).

1.2 La fouille de données pour les STIS et les STCD

Considérant le volume très important (e.g., plusieurs téraoctets par jour pour Sentinel-1) et la variété des données satellitaires (e.g., 13 bandes spectrales en lumière visible et proche infrarouge pour Sentinel-2), il est aujourd'hui impossible d'imaginer une analyse manuelle des STIS disponibles ou de leurs produits tels que les STCD. Pour tirer parti de ces données parfois gratuites et particulièrement utiles pour la surveillance et la gestion des territoires, des techniques de fouille de données doivent être considérées. Ces dernières font partie intégrante du processus d'Extraction de Connaissances à partir des Données (ECD) au sein duquel, après sélection et prétraitement, elles permettent de produire des modèles ou motifs dont l'interprétation sert de support à l'extraction de connaissances (Fayyad *et al.*, 1996) (cf. Figure 1.1). Ces étapes sont déroulées de manière itérative, jusqu'à obtention de résultats pertinents, en collaboration permanente avec l'utilisateur final. Les étapes principales de l'ECD peuvent être résumées comme suit :

- *Sélection* : cette étape consiste à choisir les données sur lesquelles la fouille sera effectuée.
- *Prétraitement* : les données brutes peuvent être bruitées, manquantes, non structurées, ou dans un format non approprié pour une analyse automatique. Cette étape consiste à une suppression partielle du bruit, une inférence de données manquantes, ou encore une transformation des données dans un format standard.
- *Transformation* : la quantité des variables disponibles dans les données sélectionnées peut parfois rendre l'analyse très difficile, complexe, voire impossible. Une sélection des variables pertinentes pour représenter les données en fonction de l'objectif est ainsi très importante afin de faciliter l'analyse et la rendre plus performante dans certains cas. Cette tâche peut se faire avec des techniques de réduction de dimensions pour réduire le nombre des variables, ou encore pour trouver des représentations invariantes. À l'échelle de chacune des variables, les valeurs peuvent être aussi quantifiées/labellisées afin d'avoir un niveau sémantique plus élevé que les mesures originales.
- *Fouille de données* : cette étape, constituant le cœur du processus ECD, consiste à choisir les méthodes et les motifs/modèles ciblés, à déterminer les paramètres éventuels des algorithmes d'extraction et extraire les motifs/modèles des données.
- *Interprétation* : la sortie de l'étape précédente nécessite une interprétation de l'utilisateur final/l'expert dans le but de comprendre les motifs/modèles obtenus, de les lier à de possibles caractéristiques des données ou à des propriétés des phénomènes/objets observés.

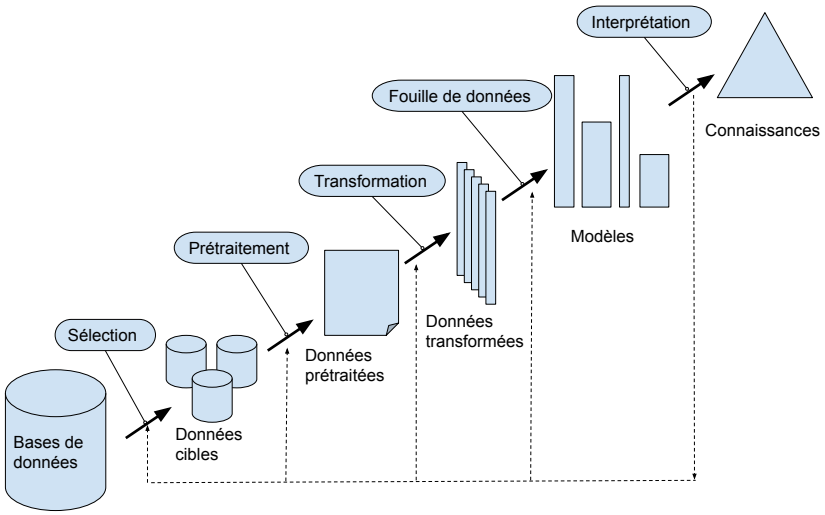


FIGURE 1.1 – Processus d'Extraction de Connaissances à partir des Données (ECD)

Quatre grandes familles de techniques de fouille de données utiles dans le cadre de l'analyse de STIS peuvent être constituées : la classification, le *clustering*, la détection de changements et la recherche de motifs fréquents.

— *Classification* : le principe de ces méthodes est d'assigner à une zone donnée représentée par un pixel ou un ensemble de pixels une classe. Cette assignation est réalisée au moyen d'un classifieur construit et entraîné sur des données pour lesquelles de telles classes sont disponibles. Les méthodes de classification sont ainsi *supervisées* par la connaissance des experts du domaine. Les techniques les plus récentes s'appuient sur des réseaux neuronaux profonds (*deep neural networks*) et obtiennent des performances de très haut niveau pour l'analyse de STIS, en particulier lorsque des images optiques sont utilisées dans le but d'identifier des classes d'occupation des sols comme les forêts, les zones urbaines ou les zones agricoles. Les travaux de Mauro *et al.* (2017) ou bien encore Ienco *et al.* (2017) sont pionniers en ce domaine.

— *Clustering* : afin de ne pas contraindre l'exploration de ces données à l'existence d'une vérité terrain ou d'une connaissance utilisateur qui est variable par essence (dans le temps, d'une personne à l'autre), l'utilisation de méthodes de fouille de données non supervisées constitue un axe de recherche essentiel. Le *clustering*, aussi appelé *classification non supervisée*, a pour objectif de regrouper les pixels similaires dans un sous-ensemble nommé *cluster* tout en s'assurant que les *clusters* extraits soient suffisamment distincts les uns des autres. Par exemple, dans les travaux de Khiali *et al.* (2018), toutes les images de la série sont segmentées et les segments couvrant le plus de pixels possible de façon complémentaire sont identifiés comme des objets de référence. Pour chaque objet de référence, un graphe dirigé acyclique est construit de façon à représenter l'évolution temporelle de la zone correspondante. Chaque graphe est ensuite synthétisé sous la forme d'un vecteur contenant autant de dimensions que de dates d'acquisition. La distance entre ces vecteurs est alors mesurée grâce à la distance euclidienne et des algorithmes de *clustering* hiérarchique ou spectral sont mis en œuvre pour identifier des *clusters* de type forêt ou vigne. Il est également possible de se ramener à des traitements au niveau du pixel. Par exemple, dans Petitjean *et al.* (2012), une telle approche est adoptée, permettant de grouper les pixels en fonction de leurs évolutions à l'aide d'un algorithme de type *k-means*. Pour ce faire, les valeurs des pixels

sont transformées en symboles exprimant par exemple la biomasse présente au sol, et, pour chaque pixel, une séquence de symboles est construite de façon à représenter son évolution. Grâce à des techniques d'alignement de séquences de type *time warping*, ces évolutions n'ont pas forcément besoin d'être synchronisées. En revanche, la nature des évolutions est contrainte par l'utilisation de distances de *clustering* dédiées. Une approche pixel est également utilisée dans Guyet et Nicolas (2016) où les pixels sont regroupés en fonction de leurs évolutions intra et interannuelles. Les évolutions intra-annuelles sont tout d'abord *clusterisées* par *k-means* en s'appuyant directement sur les valeurs des pixels. Puis, l'évolution interannuelle de chaque pixel est représentée par une séquence de *clusters* intra-annuels. Ces séquences sont finalement *clusterisées* par *k-médoïdes*, de façon à obtenir un *clustering* complet d'une STIS. L'utilisation de réseaux neuronaux en apprentissage profond pour le *clustering* de STIS reste un problème ouvert et prometteur comme mentionné par Ball *et al.* (2018).

— *Détection de changements* : comme son nom l'indique, la détection de changements vise à déterminer les zones affectées par des changements. Pour cela, les approches supervisées et non supervisées sont mobilisées. Un recensement des techniques employées a récemment été produit par Aminikhanghahi et Cook (2017). Par essence, ces méthodes sont dédiées à une tâche précise, la détection de changements, et ne peuvent pas être utilisées dans une perspective de découverte de connaissances où le type des évolutions présentes dans les STIS fait partie des informations à extraire.

— *Recherche de motifs fréquents* : la recherche de motifs fréquents (*frequent pattern mining*) a été introduite par Agrawal *et al.* (1993) pour analyser une base de données de transactions liées à des ventes de produits. Le principe est d'extraire de façon non supervisée des régularités traduisant les ensembles de produits achetés fréquemment (*itemsets fréquents*) et les dépendances les plus probables entre ces ensembles de produits, exprimées sous forme de *règles d'association*. Ces motifs ont été adaptés et utilisés dans Romani *et al.* (2013) pour extraire d'une STIS des relations entre les phénomènes météorologiques et l'évolution phénologique des champs de canne à sucre. Par exemple, il a été souvent observé une règle d'association indiquant que si les sols sont humides alors les champs de canne à sucre sont bien développés sur la même période et un peu après, et ceci avec une haute probabilité. Les travaux de Julea *et al.* (2011) ont quant à eux étendu les *motifs séquentiels fréquents* tels que ceux définis par Agrawal et Srikant (1995) de façon à extraire les évolutions partagées par un nombre suffisant de pixels ayant tendance à être groupés spatialement. Les évolutions sont ainsi qualifiées par des motifs Séquentiels Fréquents Groupés (SFG) et sont exprimées sous la forme d'une sous-séquence telle que "*biomasse importante* $\rightarrow$ *biomasse absente* $\rightarrow$ *traces de biomasse*". Ce motif indique que des pixels sont caractérisés par une biomasse importante à une date pouvant différer d'un pixel à l'autre, puis la biomasse disparaît avant de se reconstituer quelque temps plus tard, à des dates qui peuvent aussi différer. Il peut par exemple correspondre à des pixels affectés par un feu de forêt. Les motifs SFG donnent accès à une information plus fine que celle qui pourrait être obtenue à l'aide de méthodes de *clustering*. En effet, chaque pixel pouvant être décrit par un ou plusieurs motifs SFG, les groupes de pixels associés à chacun des motifs peuvent donc s'inclure et se chevaucher les uns les autres. De tels motifs peuvent néanmoins être nombreux. Il peut alors être utile de les "*clusteriser*" comme dans les travaux de Rigotti *et al.* (2014) ou de les classer comme proposé par Méger *et al.* (2015). Enfin, les motifs SFG peuvent être utilisés pour fouiller des STIS (e.g., Julea *et al.* (2011)) ou des STCD (e.g., Pericault *et al.* (2015)).

À notre connaissance, seuls les motifs SFG ont été utilisés à ce jour pour dépouiller aussi bien les STIS que les STCD. Les autres méthodes de fouille de données dédiées aux STIS pourraient également être utilisées pour traiter les STCD. Néanmoins, quelle que soit la méthode considérée, aucune d'entre elles ne prend aujourd'hui en compte les indices de confiance disponibles au sein des STCD.

1.3 Contributions

Ce travail de thèse est soutenu par la région Auvergne-Rhône-Alpes, dans le cadre de l'ARC 6, une des 8 Communautés de Recherche Académique, qui est dédiée aux Technologies de l'Information et de la Communication et aux Usages Informatiques Innovants. Il a été mis en œuvre grâce à la collaboration de deux laboratoires, le LISTIC et le LIRIS. L'objectif principal consiste à développer des méthodes de fouilles de données pour l'analyse de STCD obtenues à partir des images satellitaires, avec, pour finalité applicative, l'étude des évolutions des vitesses d'écoulement des glaciers dans un contexte de changement climatique. Afin d'encourager la découverte de connaissances à partir de telles données et d'introduire le moins possible de connaissance experte, la contribution de cette thèse porte sur des méthodes non supervisées de fouille de données à base de motifs SFG tels que définis dans Julea *et al.* (2011). Ces motifs permettent d'extraire des informations plus fines que celles apportées par un *clustering* (cf. Section 1.2) et leur potentiel pour l'analyse de STCD a été mis en évidence par Pericault *et al.* (2015). Néanmoins, ces derniers travaux font face actuellement à deux limitations principales :

1. Il n'y pas de prise de compte des indices de confiance. Dans le contexte spécifique des données satellitaires où les acquisitions sont parfois très bruitées et où les calculs de déplacement peuvent fournir des résultats avec un niveau faible de confiance, cette considération s'avère très importante afin d'assurer un résultat fiable pour l'utilisateur final.
2. La technique de classement des motifs extraits utilisée est basée sur une approche par *swap* randomisation proposée par Méger *et al.* (2015). Elle évalue les motifs individuellement sans prendre en compte la façon dont les motifs se complètent spatialement et temporellement. Les motifs ainsi sélectionnés peuvent donc être redondants et masquer d'autres évolutions de déplacements intéressantes en reléguant ces dernières au second plan dans le classement produit.

Notre première contribution adresse le problème lié à la prise en compte des indices de confiance. Pour cela, nous définissons tout d'abord des mesures de fiabilité à l'échelle des motifs séquentiels et de leurs occurrences. Ces mesures permettent de sélectionner les motifs portés par des données qui sont en moyenne suffisamment fiables. Nous proposons ensuite un algorithme d'extraction sous contrainte de fiabilité, en nous inspirant de travaux tels que ceux de Pei *et al.* (2007) sur l'extraction de motifs séquentiels sous contraintes. Notre algorithme s'appuie notamment sur une recherche efficace des occurrences les plus fiables par programmation dynamique et sur un élagage de l'espace de recherche fourni par une stratégie de *push* partiel, ce qui permet de considérer des STCD conséquentes. Cette nouvelle méthode a été implémentée sur la base du prototype existant *SITS-P2miner*, développé au sein du LISTIC et du LIRIS pour extraire et classer des motifs SFG.

Notre deuxième contribution adresse la deuxième limitation : la sélection de motifs pertinents et complémentaires pour décrire les évolutions de STCD. L'approche proposée est basée sur un critère informationnel et inspirée de travaux tels que ceux de Lam *et al.* (2014) sur la sélection de motifs séquentiels qui compressent les données et ceux de Muzammal et Raman (2015) sur l'extraction de motifs séquentiels dans des bases probabilistes. Elle permet de prendre en compte à la fois les indices de confiance et la façon dont les motifs sélectionnés se complètent spatialement et temporellement. Pour ce faire, les indices de confiance sont interprétés comme des probabilités, et les STCD comme des bases de données probabilistes dont les distributions ne sont que partiellement connues. Le gain informationnel associé à un motif est alors défini en fonction de la capacité de ses occurrences à compléter/affiner les distributions caractérisant les données. Sur cette base, une heuristique est proposée afin de sélectionner des motifs informatifs et complémentaires. Cette méthode permet de fournir un ensemble de motifs faiblement redondants et donc plus faciles à interpréter que ceux fournis par la *swap* randomisation. Elle a été implémentée au sein d'un prototype dédié.

Les deux propositions sont évaluées à la fois quantitativement et qualitativement en utilisant une STCD de référence couvrant des glaciers du Groenland et construite à partir de données optiques par Tedstone *et al.* (2015). Une autre STCD que nous avons construite à partir de données radar couvrant le massif du Mont-Blanc est également utilisée. Outre le fait d'être construites à partir de données et de techniques de télédétection différentes, ces séries se différencient drastiquement en termes d'indices de confiance, la série couvrant le massif du Mont-Blanc se situant à des niveaux de confiance très faibles.

1.4 Organisation du mémoire

La suite de ce manuscrit présente tout d'abord l'état de l'art des domaines liés à cette thèse, notamment le calcul et l'analyse des champs de déplacement à partir des images satellitaires et les méthodes de fouille de données pouvant être utilisées afin d'extraire les motifs portés par des données fiables et de sélectionner des motifs complémentaires spatialement et temporellement. Cet état de l'art se décompose en deux chapitres :

— Le chapitre 2 donne une vision globale des caractéristiques des données satellitaires disponibles actuellement, à savoir les données optiques et SAR. Il détaille ensuite les deux méthodes principales pour mesurer les déplacements à partir de ces types de données, l'une qui utilise les informations sur l'amplitude (données optiques ou radar), et l'autre qui utilise les informations sur la phase (données radar uniquement). Chacune de ces méthodes requiert des configurations différentes, et répond ainsi à des cas de figure spécifiques. Finalement, nous présentons différentes méthodes couramment utilisées pour analyser les STCD, tout en détaillant leurs avantages et leurs inconvénients.
— Le chapitre 3 présente l'état de l'art sur les méthodes de fouille de données permettant de sélectionner les motifs séquentiels afin d'analyser des STCD. Une première partie est tout d'abord consacrée à l'extraction de motifs séquentiels sous contraintes tandis qu'une deuxième partie est dédiée à la sélection de motifs sur critères informationnels. Ce type de sélection qui permet de se concentrer sur des motifs complémentaires à l'échelle du jeu de données peut notamment faciliter l'interprétation des résultats obtenus en évitant de fournir à l'expert des motifs redondants et trop nombreux.

Les deux chapitres suivants sont consacrés aux propositions faites pour prendre en compte les indices de confiance lors de l'extraction et de la sélection de motifs séquentiels à partir de STCD.

— Le chapitre 4 détaille la première contribution scientifique de cette thèse où il est proposé de sélectionner les motifs portés par des données qui sont en moyenne suffisamment fiables. L'algorithme d'extraction sous contrainte de fiabilité est détaillé, en particulier en ce qui concerne la recherche efficace des occurrences les plus fiables et l'élagage de l'espace de recherche.
— Le chapitre 5 propose une nouvelle méthode de sélection des motifs sur critère informationnel prenant en compte les indices de confiance et la façon dont les motifs sélectionnés se complètent spatialement et temporellement. L'interprétation probabiliste des indices de confiance est expliquée et le calcul du gain informationnel associé à chaque motif est présenté.

Les chapitres 4 et 5 illustrent les méthodes proposées à travers des expériences sur des données réelles permettant d'évaluer qualitativement et quantitativement la pertinence des propositions faites. Enfin, nous concluons avec le chapitre 6 et présentons les perspectives ouvertes par ce travail.

Chapitre 2

Calcul et analyse de STCD à partir de STIS

Sommaire

2.1 Introduction

Ce chapitre introduit dans un premier temps les techniques d'acquisition des images satellitaires. Les deux principaux types d'images satellitaires, optique et Radar à Synthèse d'Ouverture (en anglais *Synthetic Aperture Radar*) (SAR), ainsi que leurs avantages et leurs inconvénients seront présentés. Ensuite, les méthodes de calcul de champs de déplacements à partir des images satellitaires, afin de construire des Séries Temporelles de Champs de Déplacements (STCD), seront détaillées. En plus, les indices de confiance associés aux estimations de déplacement seront également présentés. Finalement, nous aborderons les deux principales approches permettant d'analyser ces STCD : l'une est basée sur les techniques de régularisation et d'agrégation des informations, l'autre utilise les techniques non supervisées de fouille de données.

2.2 Données satellitaires

Les Satellites d'Observation de la Terre (SOT) sont utilisés depuis plus de 40 ans pour effectuer des observations géophysiques et géographiques de la Terre depuis l'espace. Ces observations sont utilisées à des fins telles que la mesure de paramètres physiques pour le suivi des océans et de l'atmosphère, la détection de changements de couverture aux sols, la connaissance et la prédiction des effets du changement climatique. D'un point de vue technique, les SOT ont des caractéristiques variables pour répondre à ces différents besoins, notamment sur la résolution spatiale, la fréquence de passage, la couverture, les modes d'acquisition. Actuellement, les deux grandes familles d'instruments d'acquisition d'images embarqués à bord des satellites sont les capteurs passifs et les capteurs actifs. Les premiers détectent l'énergie naturelle (rayonnement) émise ou réfléchie par l'objet ou la scène observée. La lumière solaire réfléchie est la source de radiation la plus souvent mesurée par ces capteurs. Les instruments passifs comprennent : les capteurs optiques opérant dans le visible ou l'infrarouge, les spectromètres, etc. Quant aux capteurs actifs, ils fournissent leur propre source d'énergie pour éclairer les objets qu'ils observent. En effet, un capteur actif émet un rayonnement dans la direction de la cible à étudier, et détecte ensuite le rayonnement réfléchi. Les principaux instruments actifs sont : les altimètres laser, les lidars, les radars et les sondeurs. Parmi ces techniques d'acquisition, nous allons considérer dans cette thèse seulement deux instruments permettant de produire des images bidimensionnelles de la surface de la Terre : les capteurs optiques et les radars à synthèse d'ouverture (SAR).

2.2.1 Imagerie optique

L'imagerie optique est une technologie passive, c'est-à-dire que le capteur ne reçoit que la réflexion du rayonnement solaire pour former l'image. C'est pour cette raison que ce type d'imagerie n'est fonctionnelle que pendant la journée et nécessite de bonnes conditions météorologiques (ciel dégagé) pour obtenir de bons résultats. Par contre, son avantage principal repose sur la facilité d'interprétation par rapport à l'imagerie SAR, grâce à son mode de fonctionnement similaire à la vision humaine. Par ailleurs, elle peut être utilisée dans certains cas comme une source supplémentaire à l'imagerie SAR afin d'obtenir un meilleur résultat (Vaglio Laurin *et al.*, 2013; Forkuor *et al.*, 2014; Reiche *et al.*, 2015). Ce type d'imagerie est divisé notamment suivant le nombre de bandes spectrales disponibles : données mono-

spectrales (une seule bande), multispectrales (quelques bandes), superspectrales (quelques dizaines de bandes) et hyperspectrales (une centaine de bandes). Il est à noter que pour chaque satellite, la résolution spatiale varie selon les bandes spectrales. Par exemple, pour les satellites Sentinel-2A et 2B, les bandes 2 (490 nm), 3 (560 nm), 4 (665 nm), 8 (842 nm) ont une résolution spatiale de 10 m tandis que les autres bandes ont une résolution spatiale de 20 ou 60 m. Les principaux satellites qui fournissent les images optiques sont : Landsat [1] (à partir de 1972), SPOT [2] (à partir de 1986), Pléiades [3] (à partir de 2011), Sentinel 2 [4] (à partir de 2015), etc. Ces satellites peuvent également être groupés par tranche de résolution : basse résolution $10 - 100$ m (Landsat, Sentinel-2, SPOT-1 $\rightarrow$ SPOT-4), haute résolution $1 - 10$ m (SPOT-5 $\rightarrow$ SPOT-7, ALOS-1), très haute résolution < 1 m (WorldView, KOMPSAT-3, Pleiades 1A, Pleiades 1B). Parmi ces satellites, les données de Landsat et de Sentinel-2 sont accessibles gratuitement pour tous. Cela constitue un grand avantage pour les chercheurs dans différents domaines, de la géographie à la surveillance des phénomènes naturels.

Par exemple, la figure 2.1 montre une image en *fausse couleur* [5] du glacier Nordenskiöld (sur le côté droit de l'image) prise par le satellite Sentinel-2A en septembre 2017. Pendant cette période de l'année, la végétation, constituée majoritairement de l'herbe et de petites plantes, apparaît en rouge sur cette image. Des zones en bleue claire dans l'eau contiennent des sédiments fins produits par l'abrasion des glaciers frottant contre les rochers, appelés "lait des glaciers".

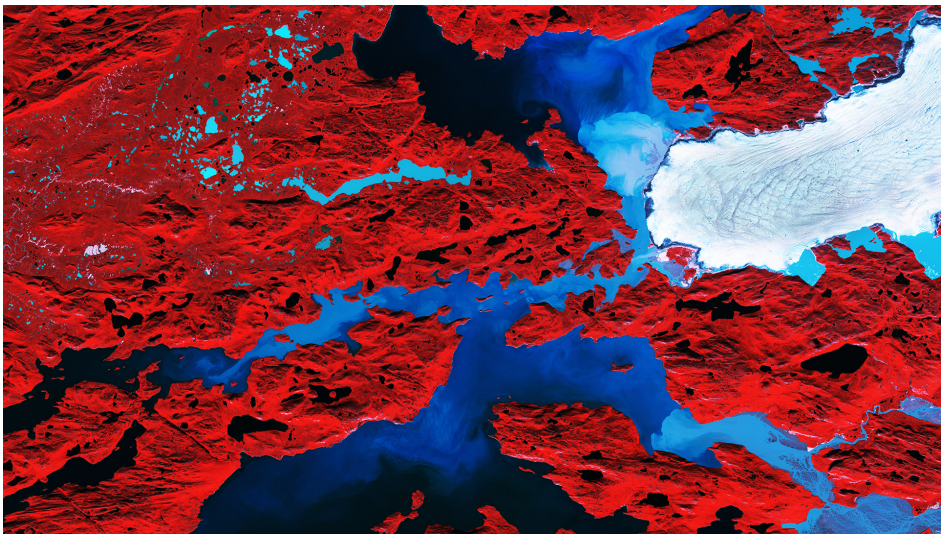


FIGURE 2.1 – Image Sentinel-2 du glacier Nordenskiöld, Groenland (source ESA)

L'estimation de déplacement s'effectue en comparant les images observant les mêmes zones géographiques à des dates différentes. Afin de faciliter ce calcul, nous sommes généralement ramenés à réduire la dimensionnalité des images multispectrales. Par exemple, les bandes synthétiques telles que la première composante de l'Analyse en Composantes Principales (ACP) peuvent être utilisées (Scambos *et al.*, 1992). Heid et Kääb (2012) ont pour leur part

1. http ://landsat.usgs.gov/
2. https ://spot.cnes.fr/
3. https ://pleiades.cnes.fr/
4. https ://earth.esa.int/web/guest/missions/esa-operational-eo-missions/sentinel-2
5. Une représentation en fausse couleur consiste en une image avec des couleurs différentes des couleurs réelles pour rendre les différences plus visibles.

sélectionné simplement la bande panchromatique pour avoir une meilleure résolution par rapport aux autres bandes.

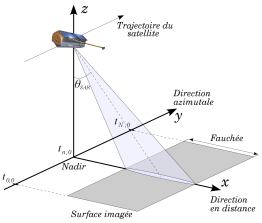
2.2.2 Imagerie radar

Contrairement à l'imagerie optique, le radar à synthèse d'ouverture (SAR) est une technologie active. Pour créer une image SAR, l'antenne portée par le satellite, alignée le long de sa trajectoire transmet latéralement des faisceaux d'impulsions d'ondes électromagnétiques vers la surface de la Terre et en reçoit les échos (cf. Figure 2.2). Les images brutes sont ensuite construites à partir des signaux rétrodiffusés, contenant deux informations importantes : le délai et l'intensité. Ces signaux dépendent de plusieurs facteurs : distance entre l'antenne et le sol, rugosité et propriété diélectrique de la surface (Massonnet et Souyris, 2008). Comme les satellites se déplacent en même temps que l'acquisition, les satellites SAR utilisent les informations relatives au déplacement du satellite pour construire une image de haute résolution avec une antenne de taille raisonnable. Les satellites qui ont pour but de surveiller des glaciers, des déplacements de la surface de la Terre ou des immeubles disposent de bandes d'acquisition spécifiques et se trouvent à une Orbite Terrestre Basse (en anglais *Low Earth Orbit*) (LEO) pour avoir une bonne résolution. Par exemple, les satellites Sentinel-1 ont une hauteur d'orbite d'environ 700 km au-dessus de la Terre. Les images sont acquises selon des orbites ascendantes et descendantes [6]. Les capteurs fonctionnent généralement en plusieurs modes, e.g., Spotlight, StripMap, interférométrique, ScanSAR, Wide Swath, Wave Mode. Pour un satellite donné, chaque mode correspond à une résolution et une couverture spatiale spécifique. Par exemple, les images acquises en mode Stripmap des satellites Sentinel-1 ont des fauchées (i.e., la largeur des vues) de 80 km avec une résolution spatiale d'environ 5×5 m, tandis que celles acquises en mode interférométrique ont des fauchées de 250 km avec une résolution de 5×20 m.

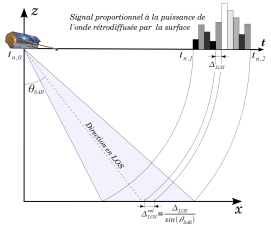
En fonctionnement, l'antenne émet des faisceaux latéralement avec des angles d'incidence différents, notés θ_{SAR}. Par exemple, pour le module *Advanced Synthetic Aperture Radar* (ASAR) du satellite Envisat, θ_{SAR} varie entre 15°et 45,2° [7]. Concernant la polarisation des impulsions, l'idée est de transmettre les vecteurs de champs électromagnétiques horizontaux (H) ou verticaux (V) et de recevoir les signaux rétrodiffusés qui peuvent être soit horizontaux, soit verticaux, soit les deux. À titre d'exemple, des images obtenues avec une émission des signaux horizontaux et une réception des signaux verticaux sont notées HV. Il y a principalement deux types de polarisation : simple (HH ou VV) et double (HH+HV ou VV+VH), les données "quad pol" (HH, VV, HV et VH) étant plus rares (satellites ALOS et RADARSAT-2). La géométrie des images SAR conduit à des pixels repérés selon des coordonnées en distance (*range*) [8] et en azimut [9] (cf. Figure 2.2).

Il y a différents formats d'image satellitaire comme les images brutes (*raw image*), les images *Single Look Complex* (SLC) et les images *Ground Range Detected* (GRD). Les images brutes contiennent principalement les informations sur les échos, la calibration de l'instrument, l'orbite et l'altitude. Les images SLC sont obtenues à partir des images brutes en faisant la focalisation SAR avec l'utilisation des données de l'orbite et d'altitude du satellite. Une image SLC est un signal complexe bidimensionnel, $z = Ae^{j\phi}$, avec A l'amplitude du signal et

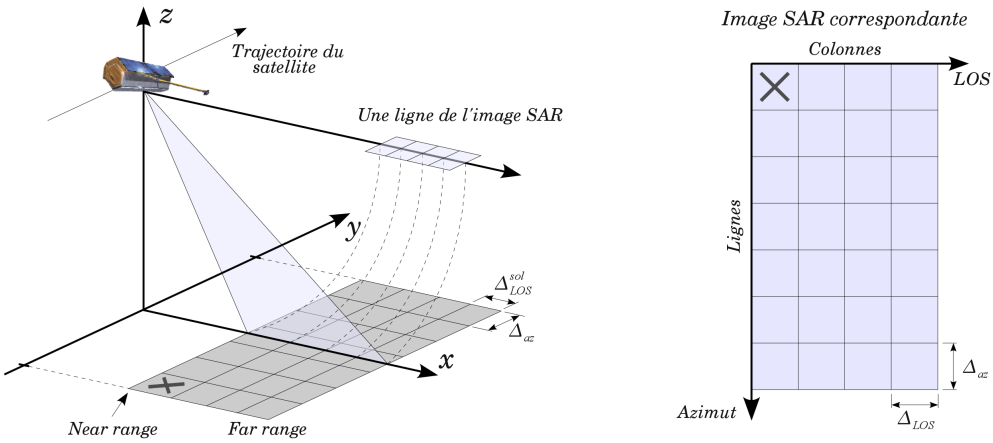
6. Ascendant : le satellite passe de l'hémisphère sud à l'hémisphère nord et inversement pour descendant
7. http ://envisat.esa.int/handbooks/asar/CNTR1-1-5.html
8. Direction en ligne de visée (en anglais *Line Of Sight*) (LOS)
9. Direction le long de la trajectoire du satellite



(a) Géométrie d'acquisition



(b) Construction d'une ligne : le chronogramme a un lien direct avec la position sol



(c) Correspondance entre les pixels SAR et les parcelles de terrain au sol

FIGURE 2.2 – Principe de construction des images SAR (source Fallourd (2012))

ϕ le déphasage. L'amplitude correspond à la force de rétrodiffusion qui dépend de la pente locale du terrain, l'humidité du sol, la rugosité, etc. (Massonnet et Souyris, 2008). La phase contient les informations liées à la distance entre l'antenne et la cible, et aux propriétés de rétrodiffusion des diffuseurs élémentaires. Quant aux images GRD, elles sont "multilooked" [10] et projetées en géométrie du sol. Contrairement aux images SLC, les images GRD ne contiennent pas les informations de phase.

Les satellites SAR peuvent être groupés par leur longueur d'onde électromagnétique, découpée en 7 bandes. Parmi ces bandes, les suivantes sont les plus utilisées : bande L : $15-30$ cm (JERS-1, ALOS-1, ALOS-2), bande C : $3.75-7.5$ cm (ERS-1, ERS-2, ENVISAT, RADARSAT-1, RADARSAT-2, Sentinel-1) et bande X : $2.4-3.75$ cm (COSMO-SkyMed 1 $\rightarrow$ COSMO-SkyMed 4, TerraSAR-X, TanDEM-X).

Une des caractéristiques des images SAR est la forte dépendance entre l'intensité et la pente locale du terrain, ce qui donne l'impression de relief, notamment sur les montagnes (cf. Figure 2.3). En effet, les versants face au capteur réfléchissent les ondes électromagnétiques vers l'antenne plus fortement que les versants opposés. Par conséquent, les premiers apparaissent plus clairs que les deuxièmes.

10. Une technique qui regroupe les pixels voisins en faisant la moyenne, permettant de réduire le bruit

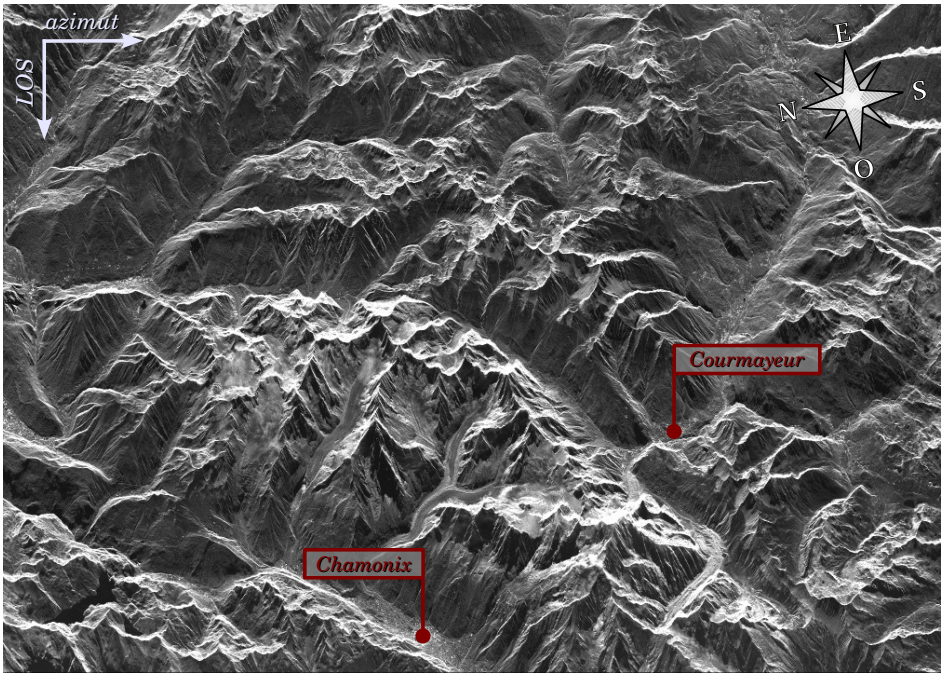


FIGURE 2.3 – Image acquise par le satellite TerraSAR-X en passe descendante sur le massif du Mont-Blanc (Fallourd, 2012)

2.2.3 Avantages et inconvénients des images satellitaires

2.2.3.1 Avantages

Un des avantages majeurs des techniques de télédétection spatiale par rapport aux mesures *in situ* repose sur leur couverture globale. En effet, cela permet d'étudier des zones de très grandes tailles, par exemple les glaciers de la chaîne du Pamir-Karakoram-Himalaya, s'étalant sur plus de 3000 km d'ouest en est. En plus, nous pouvons ainsi avoir accès à des données sur des zones difficiles d'accès, pour des raisons géographiques ou politiques. Effectuer de telles études avec les mesures *in situ* traditionnelles serait très difficile, voire impossible.

En parallèle, les orbites des satellites sont programmées de façon à ce que l'on puisse avoir des acquisitions sur n'importe quelle zone géographique sous le même angle de vision et avec les mêmes configurations à des intervalles réguliers, e.g., la constellation Sentinel-1A et 1B avec une répétitivité de 6 jours en Europe et sur des sites "à risque". Une telle configuration permet un suivi temporel de façon homogène des régions d'intérêt. En effet, cette disponibilité automatique est indispensable non seulement pour suivre des évolutions sur le long terme comme les glaciers, mais également des changements brusques comme les éruptions volcaniques, les glissements de terrain. Pour les premières, une série temporelle d'images acquises à des intervalles réguliers peut être facilement construite pour chercher les évolutions intéressantes (Petitjean *et al.*, 2012; Fallourd *et al.*, 2011; Reiche *et al.*, 2015; Nguyen *et al.*, 2017). Quant aux deuxièmes, nous portons plus d'attention sur les changements ayant lieu entre deux images acquises avant et après l'événement (Tobita *et al.*, 2001; Chadwick Jr *et al.*, 2006).

2.2.3.2 Inconvénients

Précision Malgré de nombreux développements dans le domaine, les techniques de télédétection par satellite ne peuvent pas dans un futur proche égaler les mesures *in situ* en termes de précision. En effet, une mesure *Global Positioning System* (GPS) permet de suivre des déplacements avec une précision millimétrique (Huss *et al.*, 2007; Andrews *et al.*, 2014), difficilement atteignable par télédétection dans le cadre de déplacements sur les glaciers ou des glissements de terrain. Lors du processus d'acquisition des images satellitaires, plusieurs sources d'erreur peuvent être introduites. Par exemple, les images SAR souffrent de façon inhérente du bruit multiplicatif, appelé *speckle* (cf. Figure 2.4). Celui-ci est le résultat des fluctuations aléatoires des ondes retournées par de nombreux objets, appelés *scatterer*, présents sur la surface correspondant à une cellule de résolution qui donne un pixel de l'image. De son côté, une bonne exploitation de l'imagerie optique nécessite, par principe, de bonnes conditions météorologiques, i.e., pas de nuage, bon éclairage solaire. Ces deux techniques de télédétection souffrent également les perturbations atmosphériques et éventuellement des bruits électroniques du capteur. C'est pour ces raisons que l'utilisation des images satellitaires demande beaucoup d'efforts sur l'exploitabilité des données. Par conséquent, les mesures *in situ* (si disponibles) et les connaissances d'experts sont souvent utilisées en complément avec les techniques de télédétection afin d'évaluer la fiabilité de celles-ci.

FIGURE 2.4 – Présence du *speckle* dans une image SAR sur une zone agricole de la Vallée du Tibre, Italie (source ESA)

Géométrie Pour les satellites SAR, leur angle d'incidence θ_{SAR} varie généralement entre 20° et 50°. Ce facteur a un impact direct sur les images obtenues en termes de géométrie. En effet, sur les régions avec une topographie complexe comme les glaciers, les collines, il est inévitable d'avoir des distorsions géométriques sur des images acquises. Il est difficile, voire impossible dans certaines configurations, de corriger ces distorsions sans connaissance préalable de la topographie de la zone d'intérêt, via par exemple un Modèle Numérique de Terrain (MNT). Les différentes distorsions sont montrées sur la figure 2.5. Pour les versants face au satellite, on a :

— des zones de compression (1) pour $0° < \alpha_{pente} < \theta_{SAR}$ (avec α_{pente} l'angle trigonométrique de la pente par rapport à l'axe horizontal) : l'image est comprimée dans la direction de la ligne de visée ;
— des zones de recouvrement (2) pour $\alpha_{pente} = \theta_{SAR}$: tous les échos rétrodiffusés sont concentrés en un seul point ;
— des zones de repliement (3) pour $\theta_{SAR} < \alpha_{pente} < 90°$: l'objet est imagé dans le sens inverse donnant l'impression d'être l'image miroir de l'objet réel et vient recouvrir les points équidistants.

Pour les versants opposés, on observe :

— des zones de dilatation (4) pour $-\theta_{SAR} < \alpha_{pente} < 0°$: l'image est étirée dans la direction de la ligne de visée ;
— des zones d'ombre (5) pour $-90° < \alpha_{pente} < -\theta_{SAR}$ et les régions situées derrière : ces régions ne sont pas éclairées par le satellite, mais sont présentes sur l'image sous forme de pixels noirs.

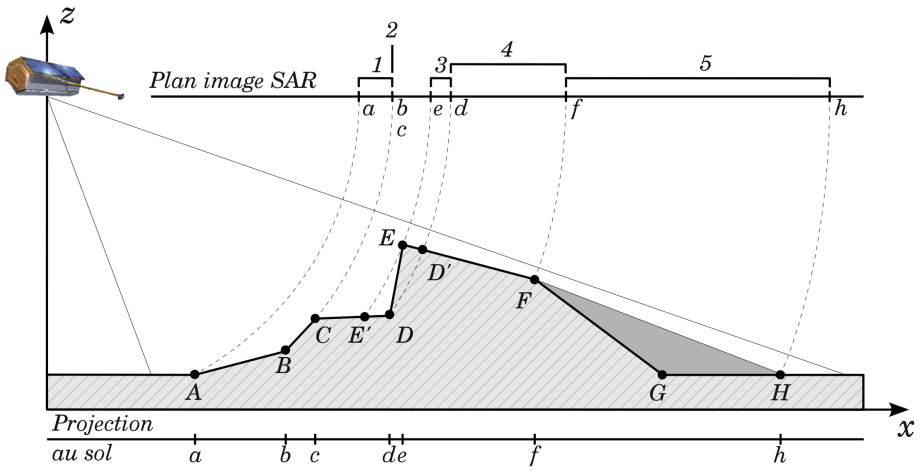


FIGURE 2.5 – Schéma des distorsions géométriques de l'imagerie SAR : (1) zone de compression, (2) zone de recouvrement, (3) zone de repliement, (4) zone de dilatation et (5) zone d'ombre (FG) et d'ombre portée (GH) (source Fallourd (2012))

Perte de similarité Afin d'avoir une estimation de déplacement satisfaisante, il est indispensable que les images satellitaires acquises à différentes dates sur une même zone géographique gardent une bonne similarité entre elles. C'est-à-dire qu'il y a un certain niveau de ressemblance entre deux images consécutives permettant de retrouver les mêmes détails, les mêmes formes d'une image à une autre. Or, les changements de surface, dus aux déplacements rapides des glaciers, aux glissements brutaux de terrain, aux conditions météorologiques, ou aux précipitations sont très critiques pour l'imagerie satellitaire. En plus, les images optiques sont notamment sensibles aux variations de couleur et d'albédo de la surface. De leur côté, les images SAR peuvent facilement perdre la cohérence interférométrique entre deux acquisitions dans ces conditions, surtout pour des images acquises avec un écart temporel trop important ou avec une grande distance entre les orbites (Strozzi *et al.*, 2002).

2.3 Méthodes de mesure de déplacement à partir des images satellitaires

Pour mesurer les déplacements ayant lieu entre deux images, capturant la même zone géographique avec les mêmes configurations (e.g., le même capteur, le même mode de fonctionnement, avec une légère différence de l'orbite), le principe de base repose sur la comparaison entre ces deux images. Cette comparaison peut se faire à deux niveaux : l'intensité (disponibles en optique et SAR - aussi appelé dans ce cas-ci *amplitude*) et la phase (seulement pour les images SAR). Le choix des méthodes dépend du type de données disponibles (optique ou SAR), de l'amplitude de déplacement estimée entre les deux dates d'acquisition, etc. Par exemple, l'interférométrie différentielle, qui utilise principalement les informations sur la phase, permet généralement de mesurer les déplacements avec une haute précision, de l'ordre millimétrique (Hooper *et al.*, 2012). Il n'est pour autant pas toujours possible d'utiliser cette technique parce qu'elle requiert une bonne cohérence entre deux images SAR. Or, suite à des changements brusques de la zone d'intérêt (e.g., glissements de terrain, grandes constructions), des configurations inadaptées de l'orbite, la cohérence peut diminuer fortement, empêchant une bonne estimation de déplacement avec cette méthode. La technique dite *offset tracking* est une alternative à l'interférométrie différentielle, car l'utilisation de celle-ci dans le contexte des déplacements glaciaires est souvent limitée par la perte de cohérence. En effet, *offset tracking* est moins contraignante vis-à-vis des configurations d'acquisition, et plus adaptée pour de grands déplacements, mais sa précision n'est pas comparable avec les méthodes d'interférométrie différentielle.

2.3.1 Interférométrie différentielle

L'Interférométrie Différentielle SAR (en anglais *Differential Interferometry SAR*) (DInSAR) utilise les informations de déphasage disponibles dans l'imagerie SAR pour calculer les déplacements ayant lieu entre deux dates d'acquisition sur la zone d'intérêt.

La mesure DInSAR utilise deux images SAR acquises à des dates différentes avec les mêmes conditions, i.e., même capteur, même configuration, etc. Entre ces deux acquisitions, bien que le satellite repasse sur la même orbite, sa trajectoire n'est pas exactement la même, mais diffère d'une base B. La base perpendiculaire $B_\perp$ est la projection de B sur la normale à la ligne de visée de l'image maître (cf. Figure 2.6).

Avec deux images SAR préalablement recalées, nous pouvons construire l'interférogramme en multipliant, pixel par pixel, la première image avec le conjugué complexe de la deuxième image (Massonnet et Feigl, 1998) :

$$z_1 z_2^* = A_1 e^{j\phi_1} A_2 e^{-j\phi_2} = A_1 A_2 e^{j(\phi_1 - \phi_2)} = A_1 A_2 e^{j\phi} \tag{2.1}$$

L'amplitude interférométrique sera ainsi l'amplitude de la première image multipliée par celle de la deuxième, i.e., $A_1 A_2$, tandis que la phase interférométrique ϕ est la différence de phase entre les deux images, i.e., $\phi_1 - \phi_2$. En faisant référence à la figure 2.6, pour le point P, la phase interférométrique dépend de la différence de trajet des ondes électromagnétiques $d = R_M - R_S$:

$$\phi(P) = \frac{4\pi}{\lambda}\left(R_M - R_S\right) = \frac{4\pi}{\lambda} d \tag{2.2}$$

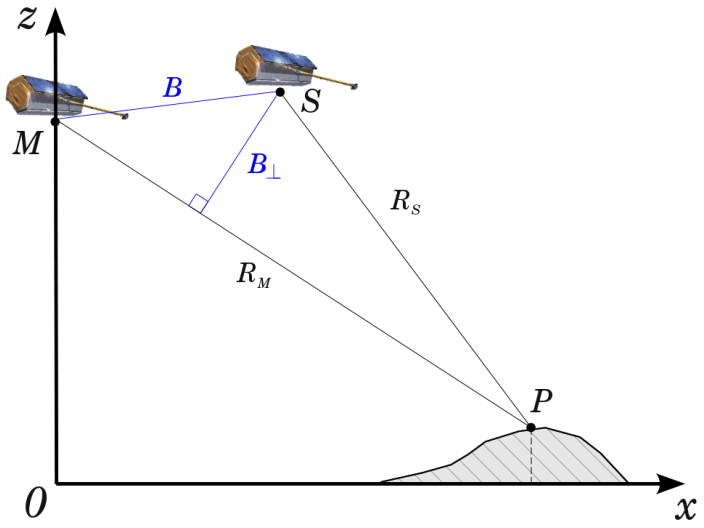


FIGURE 2.6 – Configuration entre deux acquisitions SAR (source Fallourd (2012))

où λ est la longueur d'onde.

Une fois l'interférogramme construit, il peut être utilisé pour deux applications différentes. Dans le cas où les deux images sont acquises simultanément, sous des angles légèrement différents, les caractéristiques de rétrodiffusion étant quasi identiques, la phase interférométrique dépend alors seulement de la différence entre les chemins parcourus par des ondes électromagnétiques. Cette configuration "stéréo" est utilisée pour construire des MNT, à travers les différentes missions (SRTM, TanDEM-X). La deuxième configuration consiste à utiliser deux images à des moments différents pour estimer les changements. En effet, deux images ainsi captées avec un écart temporel apportent des informations sur le déplacement ayant lieu entre les deux dates d'acquisition.

Notre étude concerne seulement la deuxième configuration. Dans ce cas précis, la phase interférométrique ϕ dépend en réalité de la topographie, de la différence des orbites entre les passages des satellites, du déplacement entre les deux dates d'acquisitions, des perturbations atmosphériques et du bruit. Elle s'exprime comme suit :

$$\phi = \phi_{orb} + \phi_{topo} + \phi_{disp} + \phi_{atm} + \phi_{bruit} \tag{2.3}$$

où ϕ_{orb} dénote la phase orbitale, ϕ_{topo} dénote la phase topographique, ϕ_{disp} dénote la phase liée au déplacement recherché, ϕ_{atm} dénote la phase atmosphérique et ϕ_{bruit} correspond au bruit.

Par conséquent, l'information sur le déplacement n'est pas accessible directement à partir de la phase interférométrique. Afin d'en déduire la phase liée au déplacement ϕ_{disp}, il faut arriver à estimer les autres éléments constituant ϕ. La phase orbitale ϕ_{orb} peut être extraite en utilisant les informations précises de l'orbite lors des passages des satellites, alors que la phase topographique ϕ_{topo} peut être éliminée en utilisant les MNT. La phase atmosphérique ϕ_{atm}, introduite par des variations temporelles et spatiales de température, pression, et humidité atmosphérique, peut se confondre avec les déplacements entre deux acquisitions. Différentes approches sont proposées afin d'éliminer cet élément de la phase interférométrique en utilisant les informations supplémentaires comme les modèles météorologiques (Foster *et al.*,

2006; Doin *et al.*, 2009; Jolivet *et al.*, 2011, 2014), les mesures GPS (Lofgren *et al.*, 2010), ou des observations multi-spectrales (e.g., les données du capteur *Medium Resolution Imaging Spectrometer* (MERIS) embarqué sur le satellite Envisat ou *Moderate Resolution Imaging Spectroradiometer* (MODIS) embarqué sur les satellites Terra et Aqua) (Li *et al.*, 2009; Zhenhong Li *et al.*, 2009). Enfin, la phase liée au bruit ϕ_{bruit} peut être corrigée par des filtres avant, pendant et après la construction de l'interférogramme (Suo *et al.*, 2010; Chao *et al.*, 2013).

Finalement, les interférogrammes, dont la phase est comprise entre $-\pi$ et π, doivent être déroulés afin d'obtenir les déplacements. Cette étape, appelée *déroulement de phase*, s'effectue par deux approches principales : les méthodes locales et les méthodes globales. Les méthodes locales effectuent le déroulement de phase spatialement en deux dimensions (Chen et Zebker, 2000) ou en ajoutant la dimension temporelle dans les approches 3-D (Hooper et Zebker, 2007). Quant aux méthodes globales, les techniques d'optimisation basées sur des *graph cut* (Bioucas-Dias et Valadão, 2005; Ferraioli *et al.*, 2009) ou sur une descente de gradient (Ghiglia et Romero, 1994) sont utilisées. Le déroulement de phase est une étape très coûteuse en termes de ressources de calcul. Chen et Zebker (2002) ont proposé une méthodologie dans laquelle un interférogramme de grande taille est partitionné en un ensemble de plusieurs petites régions qui sont ensuite déroulées individuellement, avant d'être assemblées avec un calcul d'optimisation sur les *offsets* entre les régions.

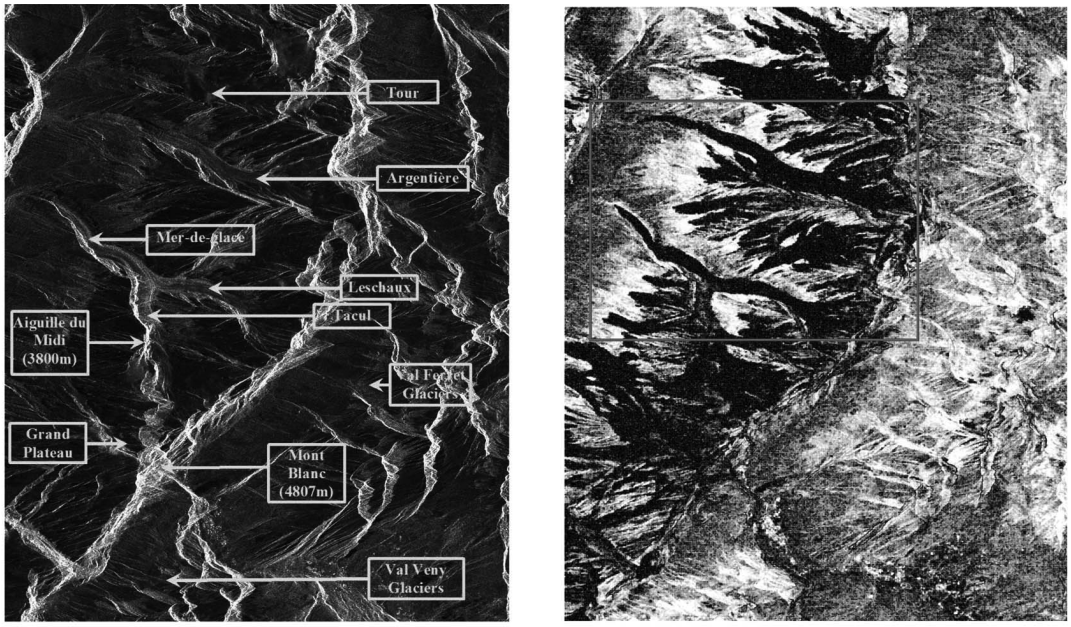
Comparées aux méthodes d'*offset tracking*, les techniques DInSAR ont une meilleure précision, de l'ordre centimétrique, voire millimétrique (Hooper *et al.*, 2012). Elle est ainsi particulièrement adaptée pour des déplacements fins et à condition de disposer des images avec une bonne cohérence interférométrique, notée γ. Étant mesurée par la corrélation entre deux images complexes (cf. équation 2.4), la cohérence interférométrique détermine le degré de similarité entre l'image maître et l'image esclave.

$$\gamma_{(m,n)} e^{i\varphi_{(m,n)}} = \frac{\sum_{(i,j)\in\mathbf{F}} z_1(i,j) z_2^*(i,j)}{\sqrt{\sum_{(i,j)\in\mathbf{F}} \left(|z_1(i,j)|^2\right)} \sqrt{\sum_{(i,j)\in\mathbf{F}} \left(|z_2(i,j)|^2\right)}} \tag{2.4}$$

où $\mathbf{F}$ dénote le voisinage qui constitue la fenêtre d'estimation de la phase et la cohérence au pixel (m, n). La cohérence γ, dont la valeur varie entre 0 et 1, peut être utilisée comme un indice de confiance de la phase interférométrique (Hanssen, 2001). En effet, une zone de cohérence élevée correspond à une zone où les réponses des ondes électromagnétiques sont stables lors de différentes acquisitions. Une forte cohérence assure ainsi la fiabilité de l'interférogramme et inversement. Comme évoqué précédemment, à cause d'une perte de cohérence importante dans les zones glaciaires due aux changements de surface des glaciers, en particulier en été (cf. Figure 2.7), les techniques DInSAR y sont difficilement appliquées.

La limitation principale des techniques DInSAR repose sur le bruit lié aux changements des propriétés de rétrodiffusion. En effet, la valeur de l'amplitude et de la phase d'une cellule d'image SAR est constituée de la somme des éléments de rétrodiffusion élémentaires qui se retrouvent sur la zone de terrain associée. Lorsqu'une zone contient un élément de rétrodiffusion dominant, comme les bâtiments, les blocs rocheux conséquents ayant une bonne orientation par rapport au satellite, la réponse obtenue sur la cellule d'image correspondante est plutôt stable même s'il y a de petits changements de terrain ou de la trajectoire du satellite. Quant aux autres zones n'ayant pas d'éléments de rétrodiffusion dominants, la réponse devient très sensible aux petits changements, rendant l'image SAR difficile à interpréter (Zebker et Villasenor, 1992). Cet effet, appelé *décorrélation du speckle*, dépend également de la longueur d'onde électromagnétique, les capteurs émettant les longueurs d'onde plus grandes (bandes

L, P) étant moins sensibles par rapport aux autres capteurs. En ce qui nous concerne, si la décorrélation est importante, l'estimation du déplacement par DInSAR devient moins fiable. Différents travaux sont faits pour diminuer l'effet d'une décorrélation élevée, en utilisant les filtres sur les images SAR (Gatelli *et al.*, 1994), les filtres sur l'interférogramme (Goldstein et Werner, 1998), ou l'estimation des fréquences locales (Trouvé *et al.*, 1998). La technique de *multilooking* qui consiste à regrouper les éléments voisins de l'interférogramme permet également de renforcer le signal principal tout en réduisant le bruit de phase (Eppler et Rabus, 2012). Même si ces approches permettent d'avoir des interférogrammes plus fiables que les interférogrammes bruts, leur résultat devient insatisfaisant lorsque nous avons une décorrélation très élevée.



(a) Amplitude d'une image ERS-1 en passe descendante (17/08/1991)

(b) Cohérence interférométrique (avec 3 jours d'écart temporel)

FIGURE 2.7 – Perte de cohérence sur les zones glaciaires en été du massif du Mont-Blanc (d'après (Trouvé *et al.*, 2007))

Il existe des techniques DInSAR utilisant les séries temporelles au lieu de seulement une paire d'images afin d'atténuer les limitations, en particulier les perturbations atmosphériques. La première technique basée sur un suivi des réflecteurs permanents, appelée *Persistent Scatterers* (PS), utilise les pixels dont la rétrodiffusion est consistante sur toutes les images (Hooper Andrew *et al.*, 2004; Hooper *et al.*, 2007). La deuxième technique, appelée *Small BASeline* (SBAS), utilise, quant à elle, seulement les paires d'images acquises avec une différence faible de trajectoire du satellite et un petit écart temporel, dans le but de minimiser la décorrélation (Lanari *et al.*, 2004). Cette technique s'adresse mieux aux cellules d'images n'ayant pas d'éléments de rétrodiffusion dominants tandis que la première est mieux adaptée pour les cellules d'images dominées par un seul élément de rétrodiffusion (Hooper *et al.*, 2012). Hooper (2008); Ferretti *et al.* (2011) ont proposé de nouvelles approches combinant ces deux techniques (PS et SBAS) afin d'obtenir des résultats plus fiables.

2.3.2 Offset tracking

Les méthodes d'*offset tracking* cherchent à estimer les déplacements ayant lieu entre deux images, maître et esclave, en appariant leurs pixels. Pour cela, elles utilisent les *fonctions de similarité*, qui peuvent être divisées en deux catégories : cohérence et corrélation. La première mesure, utilisée sur les signaux complexes, s'applique uniquement sur les images SAR (Derauw, 1999; Strozzi *et al.*, 2002), tandis que la deuxième s'applique sur les deux types d'images : optique et SAR. En effet, la mesure de corrélation utilise soit l'information d'amplitude pour l'imagerie SAR (Strozzi *et al.*, 2002; Nakamura *et al.*, 2007; Strozzi *et al.*, 2008; Giles *et al.*, 2009; Fallourd *et al.*, 2011; Vernier *et al.*, 2011; Raucoules *et al.*, 2013; Singleton *et al.*, 2014), soit des bandes synthétiques (e.g., l'indice de glace, l'indice de végétation par différence normalisée (en anglais *Normalized Difference Vegetation Index*) (NDVI)) ou réelles pour l'imagerie optique (Scherler *et al.*, 2008; Heid et Kääb, 2012; Dehecq *et al.*, 2015). À l'heure actuelle, l'utilisation des mesures de corrélation est plus répandue que la cohérence. Lorsque les premières sont utilisées, la technique est aussi appelée *corrélation d'amplitude*.

On considère deux images préalablement recalées, l'image maître I_m et l'image esclave I_s. Pour chaque pixel (i, j), une fenêtre Ω_m dans l'image maître centrée sur ce point est définie. Le vecteur de déplacement $\vec{V}^d(i, j)$ est obtenu en cherchant le maximum de la fonction de similarité $sim(p, q)$ entre Ω_m et une autre fenêtre Ω_s de la même taille dans l'image esclave centrée sur $(i + p, j + q)$. À partir d'un *a priori* sur le déplacement, la fenêtre de recherche $\Delta(i, j)$ (cf. Figure 2.8) pour chaque position (i, j) est définie avec :

$$(i + p, j + q) \in \Delta(i, j) = [i + p_{min}, j + p_{max}] \times [i + q_{min}, j + q_{max}] \tag{2.5}$$

L'estimation du vecteur de déplacement est :

$$\vec{V}^d(i, j) = (p_{opt}, q_{opt}) = \underset{(p,q)}{\operatorname{argmax}}\, sim(p, q) \tag{2.6}$$

avec $sim(p, q)$ la fonction de similarité entre les deux imagettes décalées de (p, q) et (p_{opt}, q_{opt}) les deux composantes du déplacement optimal en x et en y dans le plan de l'image.

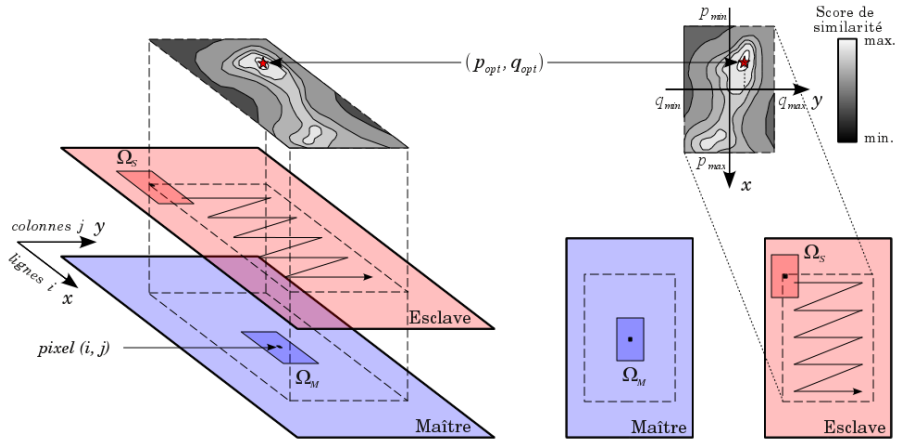


FIGURE 2.8 – Recherche du maximum de similarité (source Fallourd (2012))

La taille des fenêtres est choisie en fonction des *a priori* sur le type de déplacements attendu et les configurations d'acquisition.

Quant à la fonction de similarité, outre la mesure de cohérence γ, plusieurs fonctions de corrélation peuvent être utilisées, dont la plus simple est l'inter-corrélation (*Cross-Correlation*) :

$$CC(p,q) = \sum_{(k,l)\in\Omega_m} I_m(k,l) I_s(k+p, l+q) \tag{2.7}$$

Cette fonction de similarité dépend directement de la valeur des pixels. Plus elle est grande, plus le résultat est grand. En niveaux de gris, deux zones claires corrélant mal peuvent avoir une valeur de similarité plus forte que deux zones bien corrélées. C'est pour cela que la normalisation est nécessaire :

$$NCC(p,q) = \frac{\sum\limits_{(k,l)\in\Omega_m} I_m(k,l) I_s(k+p, l+q)}{\sqrt{\sum\limits_{(k,l)\in\Omega_m} I_m(k,l)^2} \sqrt{\sum\limits_{(k,l)\in\Omega_m} I_s(k+p, l+q)^2}} \tag{2.8}$$

Par définition, cette fonction ne mesure que la similarité de valeur entre deux images. Il arrive parfois que deux images soient très similaires malgré le fait que l'une est sombre et l'autre est claire, e.g., en cas de changement de conditions météorologiques. Dans ce cas, cette mesure sera très faible tandis qu'elle devrait être élevée. C'est pour cela que finalement, une fonction de similarité appelée *inter-corrélation centrée normée* est souvent utilisée :

$$ZNCC(p,q) = \frac{\sum\limits_{(k,l)\in\Omega_m} \left(I_m(k,l) - \bar{I}_m\right)\left(I_s(k+p, l+q) - \bar{I}_s\right)}{\sqrt{\sum\limits_{(k,l)\in\Omega_m} \left|I_m(k,l) - \bar{I}_m\right|^2} \sqrt{\sum\limits_{(k,l)\in\Omega_m} \left|I_s(k+p, l+q) - \bar{I}_s\right|^2}} \tag{2.9}$$

où $\bar{I}_m$ et $\bar{I}_s$ indiquent respectivement la valeur moyenne des pixels dans l'image maître I_m et dans l'image esclave I_s.

Cette fonction de similarité $ZNCC$ varie entre -1 et 1. La valeur 1 correspond à une corrélation totale (deux zones sont identiques), 0 à une décorrélation totale et -1 à une anti-corrélation totale (opposition).

Ces fonctions de similarité sont tout d'abord destinées aux images optiques qui ne contiennent que du bruit additif. Pour leur part, les images SAR ont un bruit multiplicatif (*speckle*), qui, en théorie, rend ces fonctions de corrélation moins adaptées. Cependant, la fonction $ZNCC$ peut être utilisée pour les images SAR lorsque le suivi du déplacement ne peut pas s'effectuer avec la méthode DInSAR (Nicolas *et al.*, 2012).

En utilisant la position du maximum de similarité sur la fenêtre de recherche, la précision de la mesure sera pixellique. Il est possible d'augmenter cette précision en utilisant des fonctions d'interpolation. D'après la définition, nous pouvons constater que la fonction de similarité est une fonction discrète sur la fenêtre de recherche. Pour ramener la précision au niveau sous-pixellique, il est possible de rendre cette fonction continue autour de son maximum. De cette façon, on atteint une précision de l'ordre de $\frac{1}{10}$ de pixel (Tobita *et al.*, 2001; Pathier *et al.*, 2006). L'interpolation peut se faire par plusieurs techniques :

— l'interpolation par 2 paraboles du second degré, une dans la direction x et une dans la direction y,

— l'interpolation par une fonction de type paraboloïde du second degré de deux variables (x, y) :

$$P(x, y) = ax^2 + by^2 + cxy + dc + ey + f \tag{2.10}$$

Christy (1998) a montré que la deuxième approche donne le meilleur compromis précision/rapidité de calcul.

Afin de diminuer encore l'erreur de l'estimation, Werner *et al.* (2005) ont appliqué un suréchantillonnage sur les images SLC avant d'utiliser la corrélation d'amplitude, permettant ainsi de garantir une précision de l'ordre de $\frac{1}{30}$ de pixel. De leur côté, Casu *et al.* (2011) combine les estimations de déplacement sur différents couples d'images SAR ayant une petite différence de trajectoire du satellite (approche de type SBAS) pour obtenir une série temporelle de champs de déplacements avec une précision similaire.

2.4 Indices de confiance associés aux champs de déplacements

Un indice de confiance, noté ρ, associé à chaque estimation de déplacement, exprime le niveau auquel nous pouvons faire confiance à notre estimation. Plus cet indice est grand, plus on est confiant et inversement. La méthode utilisée pour déterminer ρ dépend de plusieurs facteurs, notamment le type de données (optique ou SAR), la méthode de calcul des champs (*offset tracking* ou DInSAR) et le type de déplacement (régulier (glacier, subsidence), ou brusque (tremblement de terre, glissement de terrain)). Dans le cadre de calcul de déplacement à partir des images satellitaires, cet indice de confiance s'avère très important afin de pouvoir analyser au mieux les résultats obtenus. Comme montré précédemment, par rapport à des mesures *in situ* qui peuvent être considérées comme vérité terrain grâce à leur niveau de confiance très élevé, les techniques de télédétection appliquées sur les images satellitaires ne fournissent que les "estimations" de déplacement, qui ne sont pas aussi fiables à cause du bruit, des perturbations, ou bien des problèmes techniques des capteurs. Nous allons dans cette section détailler les méthodes principales disponibles aujourd'hui dans la littérature afin de fournir des indices de confiance pour les estimations de déplacement.

Il est important de noter que dans cette section, nous allons parler des indices de confiance liés aux déplacements et non pas de l'indice de confiance que l'on rencontre dans la fouille des règles d'association, une notion très populaire dans la communauté de fouille de données.

2.4.1 Confiances mono-couples

Comme son nom l'indique, l'indice de confiance mono-couple se calcule à partir de seulement deux images satellitaires utilisées pour estimer le champ de déplacements. Suivant les techniques utilisées, cet indice peut être basé sur les propriétés intrinsèques des fonctions utilisées lors du calcul de déplacement.

Dans le cas de l'*offset tracking*, les mesures caractérisant la fonction de similarité, e.g., le Rapport Signal sur Bruit (en anglais *Signal to Noise Ratio*) (SNR) (Strozzi *et al.*, 2002, 2008; Scherler *et al.*, 2008), la hauteur du pic (Nakamura *et al.*, 2007), la largeur à mi-hauteur du pic (Giles *et al.*, 2009), peuvent exprimer la fiabilité associée à la mesure.

En considérant la valeur du pic de similarité comme le signal et les autres valeurs dans la

fenêtre de recherche comme bruit, le SNR peut être ainsi utilisé comme un indice de confiance (Strozzi *et al.*, 2002, 2008; Scherler *et al.*, 2008). Plus la valeur du SNR est élevée, plus on aura confiance dans l'estimation de déplacement.

La hauteur du pic peut également être utilisée pour prédire si nous pouvons faire confiance aux mesures de déplacement. Par exemple, pour la fonction $ZNCC$ appliquée aux données SAR, Nakamura *et al.* (2007) ont montré de façon empirique que pour une valeur de S_{max} supérieure ou égale à 0.2, nous pouvons estimer que la mesure de déplacement correspondante est suffisamment fiable.

Finalement, la largeur à mi-hauteur du pic peut aussi être un indicateur de confiance (Giles *et al.*, 2009). Cette mesure s'estime à partir de la fonction d'interpolation sur les valeurs de similarité. Si le pic est élevé comparé aux pixels voisins, la fonction d'interpolation aura une forte courbure au sommet. Cela se traduit généralement par une faible valeur de la largeur à mi-hauteur.

Quant à l'interférométrie différentielle, les confiances associées aux mesures sont difficilement prises en compte. En effet, en raison de la complexité de la chaîne de traitement, la propagation de la confiance s'avère très difficile. Néanmoins, la cohérence interférométrique γ peut être considérée comme un indice de confiance du calcul de déplacement (e.g., Hanssen (2001)). Plus la valeur de γ est forte, plus on aura la confiance à l'interférogramme. Cette mesure souffre cependant du bruit et du biais, qui affectent l'estimation de la cohérence elle-même.

Ces méthodes ont comme point commun de s'appuyer directement sur les caractéristiques des mesures qui ont servi pendant le processus du calcul des champs de déplacements. Une autre possibilité est de tirer parti des propriétés statistiques au niveau spatial de chacun des champs de déplacements. Par exemple, dans le cas des glaciers et des glissements de terrain, les zones voisines doivent généralement avoir une certaine stationnarité au niveau de la direction et de l'amplitude du déplacement. Par exemple, Scherler *et al.* (2008); Harant *et al.* (2011) assument que les déplacements sont dans la direction de la plus grande pente. Ainsi, des mesures statistiques angulaires comme la variance ou l'écart-type peuvent être utilisées pour former un indice de confiance pour ce type de déplacement.

2.4.2 Confiances multi-couples

Dans le cadre des déplacements réguliers comme les glaciers, les propriétés statistiques des champs de déplacements peuvent être exploitées en vue d'estimer des indices de confiance non seulement à l'échelle spatiale, mais également à l'échelle temporelle. En effet, l'indice de confiance peut être exprimé intuitivement par la *cohérence temporelle directionnelle* des vecteurs de déplacement. Avec l'hypothèse d'un déplacement dans la direction de plus grande pente (Scherler *et al.*, 2008; Harant *et al.*, 2011), la direction de déplacement d'une région du glacier aux différentes dates doit être similaire. Dehecq *et al.* (2015) proposent une mesure de cohérence des vecteurs de vitesse $CVV_{x,y}$ en chaque pixel tout au long de la STCD. Elle est construite dans l'idée de la stationnarité temporelle du déplacement. Plus les vecteurs pointent vers la même direction, plus les estimations de déplacement sur la zone (x, y) sont cohérentes. Pour chaque position (x, y), la mesure de cohérence des vecteurs de vitesse est calculée en utilisant l'équation 2.11 :

$$CVV_{x,y} = \frac{\|\sum_{t=1}^{n} \vec{v}_{x,y,t}\|}{\sum_{t=1}^{n} \|\vec{v}_{x,y,t}\|} \tag{2.11}$$

où n est le nombre de champs de déplacements, et $\vec{v}_{x,y,t}$ est le vecteur de vitesse aux coordonnées (x, y) et à la date t.

Suivant l'inégalité triangulaire, la valeur de CVV est comprise dans l'intervalle $[0, 1]$. Plus les vecteurs de déplacement pointent vers des directions aléatoires, plus CVV sera proche de 0 et inversement. La valeur de confiance maximale, i.e., 1, est atteinte lorsque tous ces vecteurs pointent dans la même direction. La figure 2.9 montre par exemple la cohérence des vecteurs de vitesse calculée à partir des champs de déplacements sur des glaciers du Groenland, et obtenus durant la période 1985 - 1987 dans les travaux de Tedstone *et al.* (2015).

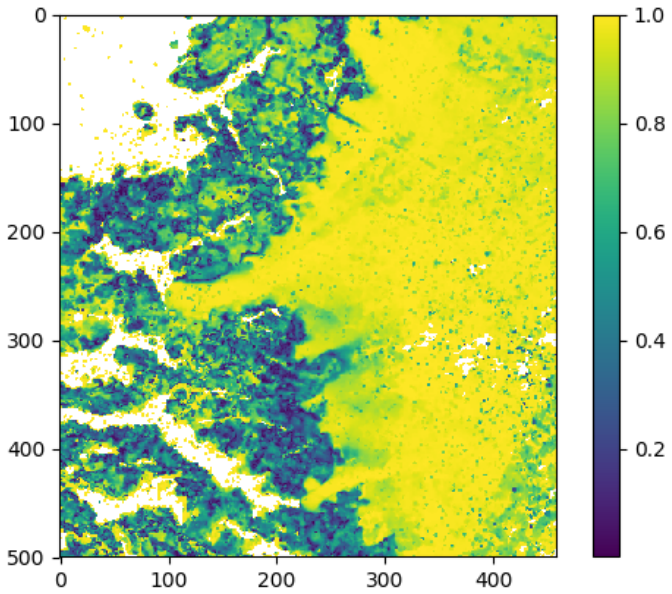


FIGURE 2.9 – Cohérence des vecteurs de vitesse (CVV) sur des glaciers du Groenland durant la période 1985 - 1987

2.5 Méthodes d'analyse des séries de champs de déplacements

Une fois la STCD et les indices de confiance associés obtenus, il est nécessaire de les analyser dans le but d'en tirer des connaissances sur les phénomènes intéressants. Il est à noter que les mesures de déplacement peuvent représenter différentes grandeurs et être sous différentes formes. Elles peuvent être des vecteurs (2-D, 3-D) ou encore des valeurs scalaires. Celles-ci peuvent être obtenues à partir des vecteurs en utilisant les opérateurs comme la norme, la projection. Les mesures peuvent, quant à elles, exprimer tout simplement les déplacements entre deux dates, mais également d'autres grandeurs dérivées comme la vitesse, l'accélération. Le choix dépend de l'objectif de l'analyse, des mesures disponibles, ainsi que des caractéristiques des mesures. Par exemple, en enlevant la valeur moyenne temporelle (ou médiane pour les données présentant des valeurs aberrantes) des mesures de vitesse, les

changements comme accélération, ralentissement, anomalie peuvent être mis en évidence. De la même manière, d'autres opérateurs peuvent être utilisés, par exemple, une division par l'écart-type pour s'affranchir des variations d'amplitudes. Concernant les techniques d'analyse des STCD, elles peuvent être divisées en deux catégories principales :

1. régularisation et agrégation des informations,
2. quantification et approche par motifs séquentiels.

La première approche est basée sur une simplification des données, en utilisant par exemple les opérations d'agrégation ou un dépouillage des données suivant un transect particulier, dans le but de faciliter leur interprétation. Cette approche est simple et efficace pour certains types de déplacement particuliers, mais demande beaucoup d'expertise pour mener à bien l'analyse. La deuxième approche, plus complexe, utilise les techniques de fouille de données spatio-temporelles. Dans le contexte de cette thèse, seulement les techniques de type non-supervisé sont considérées, car contrairement aux méthodes supervisées, elles fonctionnent sur les données non-étiquetées et sont capables d'extraire toutes les évolutions de déplacements présentes dans les données. En outre, les évolutions à extraire doivent satisfaire quelques *a priori* introduits par l'utilisateur.

Dans cette section, nous allons détailler ces deux approches dans le contexte d'analyse des STCD.

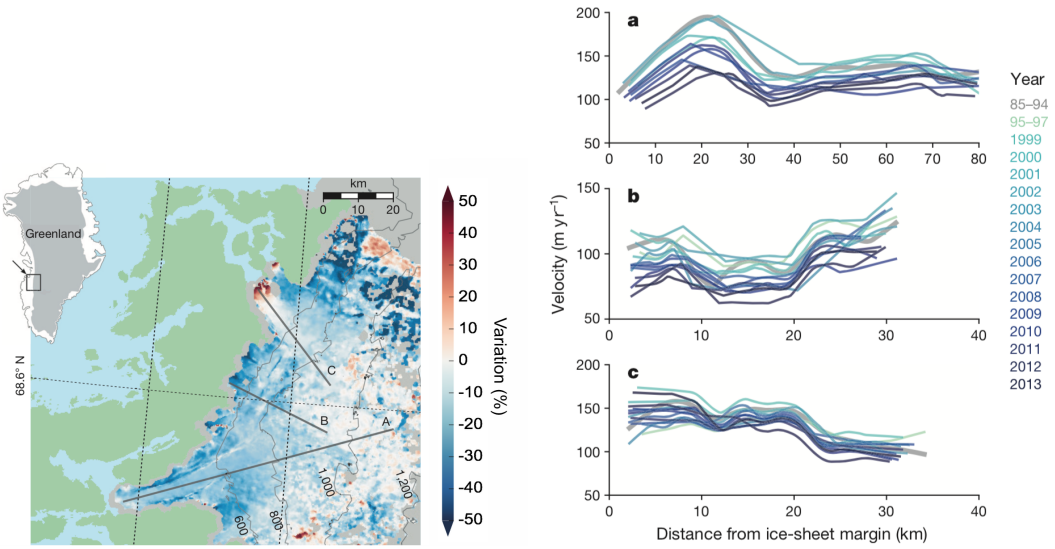
Il est à noter qu'il existe d'autres techniques non-supervisées qui sont pour l'instant utilisées afin d'analyser les Série Temporelle d'Images Satellitaires (STIS). Par exemple, Heas et Datcu (2005) extraient les régions d'intérêt sur toutes les images avant d'extraire les évolutions de ces régions. Khiali *et al.* (2018) considèrent des segments sur différentes images pour construire des graphes représentant des évolutions temporelles. Les algorithmes de *clustering* sont ensuite appliqués sur ces graphes pour identifier des *clusters* de type forêt ou vigne. Les techniques de *clustering* peuvent être également utilisées pour extraire les évolutions en prenant en compte directement les pixels (e.g., Ketterlin et Gançarski (2007); Gueguen et Datcu (2007); Petitjean *et al.* (2012)). À notre connaissance, ces approches, bien qu'elles puissent être adaptées pour analyser les STCD, ne sont pas pour l'instant appliquées sur les données de déplacement. C'est pour cette raison qu'elles ne seront pas présentées de façon plus détaillée dans la suite de cette section.

2.5.1 Régularisation et agrégation des informations

Une des stratégies fréquemment employées pour analyser des STCD consiste à simplifier fortement le contenu, soit temporel, soit spatial, de la série, afin d'extraire les informations principales. Au niveau temporel, la simplification peut se faire en utilisant des opérateurs simples comme le déplacement moyen, accumulé (Raucoules *et al.*, 2013; Chadwick Jr *et al.*, 2006), minimal, maximal (Chadwick Jr *et al.*, 2006), ou bien encore l'écart-type de déplacement (Fallourd, 2012), de chaque pixel au cours du temps. Par conséquent, on obtient pour chaque opérateur une seule valeur par pixel avec laquelle on peut produire une seule carte avec une échelle de valeurs. Par exemple, la figure 2.10a montre la variation de déplacement, en pourcentage, de la période 2007-2014 comparée à la période 1985-1994. Ces cartes, simples à produire et à interpréter, ne peuvent néanmoins pas représenter la dynamique du phénomène observé avec un niveau de détail satisfaisant, notamment lorsque l'évolution de déplacements est complexe.

La deuxième approche s'appuie sur une simplification spatiale, en analysant seulement

un nombre réduit de pixels sur des zones particulières, comme le long d'une faille sismique, ou le long d'un transect (Fallourd *et al.*, 2011; Casu *et al.*, 2011; Raucoules *et al.*, 2013; Hooper *et al.*, 2012; Tedstone *et al.*, 2015; Altena *et al.*, 2018). Par exemple, la figure 2.10b montre les estimations de vitesse le long de trois transects sur les glaciers du Groenland (cf. Figure 2.10a), pour différentes périodes entre 1985 et 2013 (Tedstone *et al.*, 2015).



(a) Variation de vitesses entre 2007-2014 et 1985-1994

(b) Vitesse des glaciers le long des trois transects A, B, C (fig. a)

FIGURE 2.10 – Évolution de vitesses des glaciers du Groenland (source (Tedstone *et al.*, 2015))

En réduisant le nombre de pixels à analyser, nous pouvons également suivre l'évolution temporelle de façon détaillée, date par date, pour chacun des pixels sélectionnés. Afin d'obtenir la tendance de déplacements en fonction du temps, une régression, souvent linaire, peut également être appliquée sur l'ensemble de pixels choisis (Ponton *et al.*, 2014).

Ces approches de simplification de la série fonctionnent bien dans des conditions particulières, sur certains types de déplacement, e.g., vitesse constante ou avec un profil linéaire, en s'appuyant fortement sur l'expertise de l'utilisateur pour choisir de bonnes zones d'intérêt. Elles demandent ainsi beaucoup de travail manuel pendant l'étape de dépouillage et d'inspection. De plus, un autre inconvénient principal réside dans la perte d'information, soit spatiale, soit temporelle. Malgré ces limitations, vu sa simplicité technique, cette stratégie est actuellement très populaire dans la communauté de télédétection.

2.5.2 Quantification et approche par motifs séquentiels

Une autre approche pour analyser les STCD, basée sur les techniques de fouille de données, consiste à extraire les évolutions de déplacements intéressantes sous forme de motifs séquentiels (Pericault *et al.*, 2015). Dans cette approche, les valeurs de déplacement, considérées comme étant scalaires, sont associées à des symboles de plus hauts niveaux sémantiques. Par exemple, les déplacements *lents* dont la vitesse est inférieure à 10 m/an pourraient être associés à un niveau 1, les déplacements *moyens* dont la vitesse est comprise entre 10 et 20 m/an pourraient être, quant à eux, associés à un niveau 2, et ainsi de suite. L'objec-

tif consiste ensuite à extraire les motifs séquentiels [11] qui se trouvent dans les séquences de symboles décrivant les niveaux de déplacement à chaque pixel au cours du temps.

Il est à noter que seuls les motifs satisfaisant des critères d'intérêt sont extraits. Par exemple, le critère de *fréquence minimale* spécifie que les motifs extraits doivent se trouver dans au moins un certain nombre de séquences. Il permet ainsi de se concentrer uniquement sur les phénomènes qui se produisent assez souvent dans la série. En outre, Pericault *et al.* (2015) ont montré que l'utilisation du critère de *connexité spatiale minimale*, introduite dans Julea *et al.* (2011), s'avère également très efficace dans le contexte des STCD. Ce critère garantit que les motifs extraits apparaissent d'une manière groupée spatialement.

Les motifs extraits, bien qu'ils satisfassent tous les critères d'intérêt, sont souvent très nombreux (Méger *et al.*, 2015). C'est pour cette raison qu'il est nécessaire d'utiliser les techniques de classement des motifs afin de sélectionner seulement une liste réduite de motifs. Méger *et al.* (2015) ont par exemple introduit une technique de classement des motifs basée sur la randomisation des données et une mesure d'information mutuelle. Cette technique prend en compte le jeu de données dans sa globalité, en regardant s'il est probable ou pas que les motifs extraits apparaissent également dans une série randomisée. La randomisation s'effectue de façon à ce que la distribution des symboles en ligne et en colonne reste identique entre la série randomisée et l'originale. Une telle randomisation permet ainsi de casser la connectivité des événements tout en gardant les fréquences spatiales et temporelles des symboles. Les motifs sont classés en fonction d'une mesure d'information mutuelle entre la position des séquences où se trouve chacun des motifs dans la base de données originale et la base randomisée. Une valeur faible de celle-ci souligne des phénomènes singuliers alors qu'une valeur élevée exprime des phénomènes importants qui ne peuvent pas être détruits par la randomisation.

L'approche par motifs séquentiels est utilisée pour analyser les STIS (Julea *et al.*, 2011; Méger *et al.*, 2011; Julea *et al.*, 2012; Méger *et al.*, 2015) ainsi que les STCD (Pericault *et al.*, 2015). Cependant, dans ces travaux, le classement des motifs s'effectue de façon individuelle, avec des redondances, et sans prendre en compte la façon dont les motifs sélectionnés se complètent. En plus, d'après notre connaissance bibliographique, les méthodes d'analyse des STCD actuellement utilisées ne permettent pas une prise en compte des indices de confiance, souvent associés aux champs de déplacements. Cet inconvénient peut devenir crucial puisque l'indice de confiance est un des éléments les plus importants des estimations de déplacement obtenues à partir des images satellitaires qui contiennent souvent beaucoup de bruit, des incohérences temporelles et spatiales.

Il est à noter que dans cette section, nous avons décrit de façon succincte l'approche. Plus de détails seront donnés dans le chapitre suivant.

2.6 Conclusions

Le développement en continu des techniques de télédétection, de l'acquisition des images avec les satellites optiques ou SAR, à des méthodes de calcul de champs de déplacements, fournit une source de données très importante pour analyser l'évolution de déplacements de n'importe quelle zone géographique sur des périodes différentes. Ces données, souvent associées à des indices de confiance, ne sont pour autant pas faciles à interpréter, compte

11. Chaque motif séquentiel est une suite de symboles, par exemple $1 \rightarrow 2 \rightarrow 3$.

tenu de leur grand volume et de leur complexité. Des méthodes actuelles peuvent demander beaucoup de travail manuel avec une certaine expertise sur la zone étudiée. De plus, elles ne permettent pas à la fois d'avoir une interprétation complète sur la zone, et en même temps de tirer parti des indices de confiance disponibles sur chaque point de mesure. C'est pour cette raison que nous allons travailler sur un processus d'analyse automatique des STCD avec une approche par motifs séquentiels qui prend en compte les indices de confiance et la complémentarité informationnelle par une sélection plus *ensembliste*.

Chapitre 3

Critères de sélection des motifs pour l'analyse des STCD

Sommaire

3.1 Introduction

Comme montré dans le chapitre précédent, les techniques de fouille de données ont la capacité d'extraire les évolutions de déplacements présentes dans les données et qui satisfont certains critères d'intérêt définis par l'utilisateur final. L'ensemble de motifs séquentiels extraits peut en outre nous fournir un résumé de la série en révélant des informations sur les évolutions de déplacements de façon très détaillée (avec une localisation spatiale et temporelle). L'extraction, effectuée de façon non supervisée, sans introduction d'*a priori* sur les types de déplacement et sur la localisation dans l'espace et dans le temps, permet ainsi de s'affranchir de la frontière des connaissances de l'expert. Ceci est primordial pour découvrir de nouvelles informations enfouies dans les bases de données.

Quant aux indices de confiance associés aux mesures de déplacement, les techniques d'extraction de motifs séquentiels sous contraintes, permettant par exemple d'extraire seulement les motifs dont l'indice de confiance est supérieur à un seuil, peuvent être appliquées à notre problème.

De plus, pour faciliter l'interprétation de l'ensemble de motifs extraits, Méger *et al.* (2015) ont proposé un classement de motifs basé sur une technique de *swap* randomisation de la base de séquences (cf. Annexe A). Or, cette technique s'effectue de façon individuelle, sans la prise en compte de la façon dont les motifs sélectionnés se complètent. Un tel classement peut fournir une liste de motifs avec redondance d'un point de vue informationnel. Par conséquent, il est nécessaire d'appliquer des techniques de classement qui fonctionnent sous une sémantique *ensembliste* avec la prise en compte de la manière dont les motifs se complètent dans la base de données.

Les méthodes existantes sont présentées dans ce chapitre consacré à la littérature concernant les techniques d'extraction de motifs séquentiels sous contraintes et à celles concernant la sélection de motifs séquentiels complémentaires au niveau informationnel.

3.2 Extraction de motifs séquentiels sous contraintes

Les techniques d'extraction de motifs séquentiels utilisent généralement des contraintes qui permettent de choisir les motifs satisfaisant certains intérêts applicatifs. Généralement, grâce à leurs propriétés spécifiques, ces contraintes nous amènent à une réduction de l'espace de recherche, permettant de diminuer de façon considérable les ressources de calcul.

3.2.1 Définitions préliminaires des motifs séquentiels

Nous abordons dans cette partie les définitions préliminaires de la fouille de données concernant les motifs séquentiels, dans le but de faciliter la lecture des parties qui suivent.

Définition 3.1 (*Itemset*, Événement). Soit $I = \{i_1, i_2, \ldots, i_n\}$, un ensemble de n symboles distincts appelés *items* et munis d'un ordre total [1]. Un *itemset* α de taille l est un ensemble non vide constitué par l *items* provenant de I. Dans le contexte de données temporelles, un *événement* est formé en associant un *itemset* à une *date d'apparition*, noté (t, α).

1. Cet ordre total, optionnel, est pourtant indispensable dans certaines situations, par exemple pour définir les préfixes/suffixes d'un motif (cf. Définition 3.5).

Nous utiliserons dans les exemples les lettres majuscules pour noter les *items*. Un événement, par exemple $(10, \{A, C\})$, sera noté $(10, AC)$ pour alléger la notation. L'ordre total sur les *items* sera quant à lui simplement l'ordre alphabétique.

Définition 3.2 (Séquence d'événements). Une *séquence d'événements* (ou *séquence*) s de *longueur* L est une liste contenant L événements, triés par ordre croissant du temps, notée $s = \langle (t_1, \alpha_1), (t_2, \alpha_2), \ldots, (t_L, \alpha_L) \rangle$.

Définition 3.3 (Sous-séquence / Sur-séquence). Une séquence $s_1 = \langle (t_1, \alpha_1), (t_2, \alpha_2), \ldots, (t_{L_1}, \alpha_{L_1}) \rangle$ est appelée *sous-séquence* d'une séquence $s_2 = \langle (t'_1, \alpha'_1), (t'_2, \alpha'_2), \ldots, (t'_{L_2}, \alpha'_{L_2}) \rangle$, avec $L_1 \leq L_2$, noté $s_1 \prec s_2$, s'il existe des entiers $1 \leq i_1 < i_2 < \cdots < i_{L_1} \leq L_2$ tels que $\alpha_1 \subseteq \alpha'_{i_1}, \ldots, \alpha_{L_1} \subseteq \alpha'_{i_{L_1}}$. Dans ce cas, s_2 est appelée *sur-séquence* de s_1.

Définition 3.4 (Motif séquentiel). Un motif séquentiel de longueur m est une suite de m *itemsets*, et il est représenté de façon suivante : $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \cdots \rightarrow \beta_m$.

Définition 3.5 (Préfixe, suffixe d'un motif). Soit un motif $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \cdots \rightarrow \beta_m$, un motif $\beta' = \beta'_1 \rightarrow \beta'_2 \rightarrow \cdots \rightarrow \beta'_n$, avec $n \leq m$, est un *préfixe* de β si les conditions suivantes sont satisfaites : (1) $\beta'_i = \beta_i$ pour $i = 1 \ldots n-1$; (2) $\beta'_n \subseteq \beta_n$; et (3) tous les *items* dans l'ensemble $(\beta_n \setminus \beta'_n)$ sont après ceux dans β'_n dans l'ordre total des *items*. Le *suffixe* d'un motif β par rapport à un de ses préfixes β' est le motif β privé de β'.

Par exemple, les motifs A, $A \rightarrow B$ sont les préfixes du motif $A \rightarrow BC$. BC est le suffixe du même motif par rapport au préfixe A, et C est son suffixe par rapport au préfixe $A \rightarrow B$.

Définition 3.6 (Occurrence d'un motif). On dit qu'un motif $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \cdots \rightarrow \beta_m$ apparaît dans une séquence $s = \langle (t_1, \alpha_1), (t_2, \alpha_2), \ldots, (t_L, \alpha_L) \rangle$, avec $m \leq L$, lorsqu'il existe des entiers $1 \leq i_1 < i_2 < \cdots < i_m \leq L$ tels que $\beta_1 \subseteq \alpha_{i_1}, \ldots, \beta_m \subseteq \alpha_{i_m}$. Dans ce cas, s est dite *couverte* par β, et $\langle t_{i_1}, t_{i_2}, \ldots, t_{i_m} \rangle$ représente une *occurrence* de β dans s.

Définition 3.7 (Base de séquences). Une *base de séquences* D est un ensemble de couples (sid, s), où s est une séquence et sid est son identifiant.

Définition 3.8 (Support, couverture). Le *support* d'un motif β dans une base de séquences D est le nombre de séquences de D couvertes par β, noté $support(\beta)$. Dans le contexte des Séries Temporelles d'Images Satellitaires (STIS) et des Séries Temporelles de Champs de Déplacements (STCD), nous utilisons également la notion de *couverture* d'un motif β, définie comme l'ensemble des séquences couvertes par β et notée $cover(\beta)$.

3.2.2 Classes de contraintes majeures

Nous introduisons dans cette section les classes de contraintes souvent utilisées en fouille de données. Les contraintes de support minimal/maximal sont les contraintes les plus rencontrées dans les applications de ce domaine.

Définition 3.9 (Contraintes de support minimal/maximal). Étant donnés la base de séquences D, le motif β et les seuils $minSupp, maxSupp \in \mathbb{N}$, la *contrainte de support minimal* a une valeur vraie si $support(\beta) \geq minSupp$ et une valeur fausse dans le cas contraire. De façon similaire, la condition devient $support(\beta) \leq maxSupp$ pour la *contrainte de support maximal*. Ces deux contraintes sont communément appelées *contraintes de support*. Un motif satisfaisant la contrainte de support minimal est appelé *motif fréquent*.

Les algorithmes d'extraction de motifs exploitent souvent la contrainte de support minimal. Cette contrainte peut parfois rencontrer des limites, du point de vue applicatif et technique. Du point de vue applicatif, les motifs la satisfaisant sont souvent très nombreux et peu d'entre eux sont vraiment intéressants, surtout lorsque le seuil de support minimal est petit. L'intégration d'autres contraintes est dans ce cas nécessaire pour à la fois réduire le nombre de motifs obtenus et se concentrer sur les motifs pertinents suivant les critères définis par l'utilisateur. Du point de vue technique, même si l'utilisation de la contrainte de support permet de réduire l'espace de recherche, il est souvent souhaitable de la combiner avec d'autres contraintes de façon active (avec une intégration des contraintes dans le processus d'extraction) afin d'avoir des résultats encore plus concis avec une consommation des ressources plus réduite également.

Dans la littérature (e.g., Bonchi et Lucchese (2005); Pei *et al.* (2007)), les différentes contraintes peuvent être vues de deux points de vue : applicatif et technique. D'un point de vue applicatif, la plupart des contraintes étudiées, autres que la contrainte de support, peuvent se répartir en 9 catégories :

— Catégorie 1 : Les contraintes sur les *items* spécifiant les *items* qui doivent apparaître ou non dans les motifs. Il y a deux contraintes principales de ce type : les contraintes d'inclusion et les contraintes d'exclusion.
Par exemple, la contrainte $C_{item}(\beta) \equiv (\forall i : 1 \leq i \leq len(\beta), \beta[i] \subseteq X)$, avec β un motif séquentiel, $len(\beta)$ sa longueur, $\beta[i]$ le $i^{\text{ème}}$ élément de β et $X \subseteq I$ un *itemset* fixé, est une contrainte sur les *items* de type inclusion.

— Catégorie 2 : Les contraintes de longueur spécifiant la longueur exacte, maximale ou minimale des motifs.
Par exemple, l'utilisateur peut s'intéresser uniquement aux motifs dont la longueur est au moins égale à 10. Dans ce cas, la contrainte est notée $C_{len}(\beta) \equiv (len(\beta) \geq 10)$.

— Catégorie 3 : Les contraintes de largeur spécifiant le nombre exact, maximal ou minimal d'*items* qui construisent les éléments des motifs.
Dans plusieurs applications de fouilles de données, telles que les séquences d'achats, en considérant chaque transaction (e.g., passage à la caisse d'un supermarché) comme un événement, ce dernier peut effectivement contenir 1 ou plusieurs articles (*items*). Dans cette application, l'utilisateur peut par exemple extraire seulement les motifs séquentiels dont *itemset* contient au moins 10 articles avec une contrainte $C_{width}(\beta) \equiv (\forall i : 1 \leq i \leq len(\beta), len(\beta[i]) \geq 10)$, avec $len(\beta[i])$ le nombre d'*items* que contient $\beta[i]$.

— Catégorie 4 : Les contraintes de sur-motif spécifiant qu'un motif doit contenir au moins un sous-motif parmi un ensemble donné. Ces contraintes s'expriment par exemple sous la forme suivante : $C_{pat}(\beta) = (\exists \beta' \in X : \beta' \prec \beta)$, avec X un ensemble de sous-motifs.
Par exemple, dans le cadre des STCD, nous pouvons nous intéresser aux zones dont l'évolution de déplacements contient une décélération, i.e., un niveau *élevé* suivi d'un niveau *moyen* ou *faible*, ou un niveau *moyen* suivi d'un niveau *faible* (lors d'une discrétisation par 3 niveaux). L'ensemble X contient alors trois sous-motifs, $X = \{C \rightarrow B, C \rightarrow A, B \rightarrow A\}$, avec les symboles A, B, C qui représentent respectivement les déplacements faibles, moyens, et forts.

— Catégorie 5 : Les contraintes d'agrégats simples, qui sont des contraintes exprimées

par une fonction d'agrégation de base, e.g., la somme, le maximum et le minimum. Dans le contexte des séquences d'achats, l'utilisateur pourrait par exemple s'intéresser uniquement aux motifs contenant des *items* dont la valeur minimale du prix est supérieure à un seuil. Ce type de contrainte peut aussi fonctionner dans les cas où un ordre est défini sur les *items*. Par exemple, avec les mêmes notations que l'exemple pour les contraintes de sur-motif, le motif $\beta = B \rightarrow C \rightarrow B$ satisfait la contrainte imposant que le déplacement minimal corresponde au moins à un niveau moyen.

— Catégorie 6 : Les contraintes d'agrégats difficiles (*tough aggregate constraints* en anglais), qui sont des contraintes exprimées par d'autres fonctions d'agrégation, comme la moyenne, l'écart type. Ces contraintes se distinguent par rapport aux contraintes d'agrégats simples par le fait qu'elles sont beaucoup plus difficiles à prendre en compte pendant l'étape d'extraction.
Par exemple, pour les séquences d'achats, une contrainte de ce type peut consister à imposer une valeur minimale/maximale sur le prix moyen des transactions dans chaque motif séquentiel.

— Catégorie 7 : Les contraintes d'expression régulière $\mathcal{C}_{ER}$ spécifiant la forme syntaxique des motifs qui peuvent être formés à partir des *items* en utilisant des opérateurs d'expression régulière, tels que la disjonction ou la fermeture de Kleene (e.g., Albert-Lorincz et Boulicaut (2003)). Chaque contrainte d'expression régulière peut être représentée par un automate fini déterministe. Un motif séquentiel β satisfait $\mathcal{C}_{ER}$ lorsque β est accepté par l'automate fini déterministe correspondant.
Par exemple, la contrainte d'expression régulière $(A*)(B|C)C$ sélectionne les motifs qui commencent par 0 ou plusieurs A, suivi(s) de B ou C, suivi d'un C.

— Catégorie 8 : Les contraintes de durée, qui elles spécifient la durée minimale ou maximale entre le premier et le dernier événement des occurrences d'un motif séquentiel.

— Catégorie 9 : Et enfin, les contraintes de *gap* spécifiant la durée maximale ou minimale entre deux événements consécutifs d'un motif séquentiel.

Parmi ces familles de contraintes, les contraintes de durée et de *gap* définissent également la façon de calculer le support du motif. En effet, ces contraintes s'appliquent directement sur chacune des occurrences qui se trouvent dans chaque séquence. Ainsi, le support est incrémenté seulement s'il y existe au moins une occurrence dans la séquence qui satisfait la contrainte. Ces contraintes sont aussi appelées contraintes *temporelles*. Quant aux autres contraintes, en s'appliquant cette fois-ci sur le motif lui-même ou sur l'ensemble des occurrences, elles ne modifient pas le calcul du support. Ces contraintes sont regroupées en deux classes : (1) les contraintes *syntaxiques* qui regroupent les contraintes sur les *items*, sur la longueur, sur la largeur, les contraintes de sur-motif, les contraintes d'expression régulière ; et (2) les contraintes *d'agrégats simples et complexes*.

Sur le plan technique, les contraintes sont classées en fonction de leurs propriétés qui peuvent s'avérer utiles lors du processus d'extraction. Ces propriétés déterminent ainsi les approches algorithmiques les mieux adaptées afin de les prendre en compte. À l'heure actuelle, les propriétés intéressantes listées dans la littérature sont la monotonie, l'anti-monotonie, la *succinctness* (Ng *et al.*, 1998), la convertibilité (Pei *et al.*, 2004a), la préfixe-monotonie (Pei *et al.*, 2007) et la *loose* anti-monotonie (Bonchi et Lucchese, 2005).

Définition 3.10 (Monotonie). Une contrainte est dite monotone si pour tout motif la

satisfaisant, tous ses sur-motifs la satisfont également.

Par exemple, la contrainte de longueur minimale est monotone puisque la longueur d'un motif est au moins égale à celle de tous ses sous-motifs.

Définition 3.11 (Anti-monotonie). Une contrainte est dite anti-monotone si pour tout motif la satisfaisant, tous ses sous-motifs la satisfont également.

La contrainte de support minimal est anti-monotone puisque le support d'un motif ne peut pas dépasser celui de ses sous-motifs.

Définition 3.12 (Succinctness). Une contrainte est dite succincte si la spécification qui la définit permet de générer directement les motifs la satisfaisant.

Par exemple, la contrainte de longueur maximale est une contrainte succincte puisque l'on peut générer directement de façon exhaustive l'ensemble des motifs dont la longueur est inférieure à un nombre défini.

Les propriétés d'anti-monotonie, de monotonie, et de succinctness des contraintes généralement utilisées sont listées dans la table 3.1.

Catégorie de contraintes		Anti-mono	Mono	Succ
Item	$C_{item}(\beta) \equiv (\forall i : 1 \leq i \leq len(\beta), \beta[i]\ \theta\ X)(\theta \in \{\subseteq, \supseteq\})$	Oui	Non	Oui
	$C_{item}(\beta) \equiv (\forall i : 1 \leq i \leq len(\beta), \beta[i] \cap X \neq \emptyset)$	Oui	Non	Oui
	$C_{item}(\beta) \equiv (\exists i : 1 \leq i \leq len(\beta), \beta[i]\ \theta\ X)(\theta \in \{\subseteq, \supseteq\})$	Non	Oui	Oui
	$C_{item}(\beta) \equiv (\exists i : 1 \leq i \leq len(\beta), \beta[i] \cap X \neq \emptyset)$	Non	Oui	Oui
Longueur	$C_{len}(\beta) \equiv len(\beta) \leq l$	Oui	Non	Oui
	$C_{len}(\beta) \equiv len(\beta) \geq l$	Non	Oui	Oui
Largeur	$C_{width}(\beta) \equiv width(\beta) \leq l$	Oui	Non	Oui
	$C_{width}(\beta) \equiv width(\beta) \geq l$	Non	Oui	Oui
Sur-motif	$C_{pat}(\beta) = (\exists \beta' \in X : \beta' \prec \beta)$	Non	Oui	Oui
Agrégats simples	$max(\beta) \leq v; min(\beta) \geq v$	Oui	Non	Oui
	$max(\beta) \geq v; min(\beta) \leq v$	Non	Oui	Oui
Agrégats difficiles	$sum(\beta)\ \theta\ v, \theta \in \{\leq, \geq\}$	Non	Non	Non
	$avg(\beta)\ \theta\ v, \theta \in \{\leq, \geq\}$	Non	Non	Non
Expressions régulières	C_{ER}	Non	Non	Non
Durée	$maxspan \equiv Dur(\beta) \leq \Delta t$	Oui	Non	Non
	$minspan \equiv Dur(\beta) \geq \Delta t$	Non	Oui	Non
Gap	$gapmin \equiv Gap(\beta) \geq \Delta t$	Oui	Non	Non
	$gapmax \equiv Gap(\beta) \leq \Delta t$	Non	Non	Non

TABLE 3.1 – Propriétés des contraintes souvent utilisées (Pei *et al.*, 2007)

Définition 3.13 (Convertible). Une contrainte est dite convertible anti-monotone sur les *itemsets* lorsqu'il y a un ordre R des *items* tel que chaque fois qu'un *itemset* X satisfait cette contrainte, tous les préfixes de X (pour l'ordre R) la satisfont également. De façon analogue, une contrainte est dite convertible monotone lorsqu'il y a un ordre R des *items* tel que chaque fois qu'un *itemset* X ne satisfait pas cette contrainte, tous les préfixes de X ne la satisfont pas non plus.

Par exemple, pour des *itemsets* représentant des articles achetés simultanément par un client, la contrainte qui précise que le prix moyen des articles doit être inférieur à 100 euros est une contrainte convertible anti-monotone. En effet, en définissant R comme l'ordre numérique croissant du prix des articles, nous observons qu'à chaque fois qu'un article est ajouté au motif, le prix moyen ne doit pas diminuer.

Définition 3.14 (Préfixe-monotonie). Une contrainte est dite préfixe anti-monotone si pour chaque motif séquentiel β satisfaisant cette contrainte, tous ses préfixes la satisfont également. Une contrainte est dite préfixe monotone si pour chaque motif séquentiel β satisfaisant cette contrainte, tous les motifs ayant β comme préfixe la satisfont également.

Par exemple, pour des séquences dont chaque *item* représente la température maximale d'une date dans l'ordre chronologique, la contrainte précisant que la température maximale doit être inférieure à 30°C est une contrainte préfixe anti-monotone.

Nous pouvons constater facilement qu'une contrainte anti-monotone est aussi préfixe anti-monotone, et qu'une contrainte monotone est aussi préfixe monotone. L'inverse n'est pas vrai. Par exemple, la contrainte *gapmax* est préfixe anti-monotone même si elle n'est ni anti-monotone ni monotone (Pei *et al.*, 2007).

Définition 3.15 (*Loose* Anti-monotonie). Une contrainte est dite *loose* anti-monotone sur les *itemsets* lorsque pour chaque *itemset* X la satisfaisant tel que $|X| \geq 2$, il existe $i \in X$ tel que $X \setminus \{i\}$ la satisfait aussi.

Par exemple, la contrainte de la variance minimale/maximale des prix des articles achetés par un client est *loose* anti-monotone (Bonchi et Lucchese, 2005).

3.2.3 Algorithmes d'extraction de motifs séquentiels fréquents

Avant de regarder l'état de l'art sur les méthodes d'extraction de motifs séquentiels sous contraintes, nous allons tout d'abord présenter de façon très succincte quelques approches proposées dans la littérature pour extraire les motifs séquentiels fréquents. Celles-ci sont des briques de base permettant de comprendre les techniques d'extraction sous contraintes, qui sont généralement inspirées de ces techniques.

Il y a un large éventail d'algorithmes d'extraction des motifs séquentiels fréquents, qui sont divisés principalement en trois catégories : approches *Apriori*, approches par listes d'occurrences, et approches par projections.

Ces approches s'appuient toutes sur la propriété d'anti-monotonie de la contrainte de support minimal. Si un motif β n'est pas fréquent, aucun de ses sur-motifs ne peut l'être. Nous pouvons alors réduire le nombre de motifs à parcourir tout en assurant le même résultat qu'une recherche exhaustive. Ces approches se différencient principalement sur l'utilisation de structures de données et d'algorithmes dédiés.

L'idée générale de la première approche, dite *Apriori*, se trouve dans la génération des motifs séquentiels candidats en assemblant des motifs séquentiels fréquents qui sont déjà trouvés, suivie d'une vérification de la contrainte de support minimal des motifs candidats. Elle explore l'espace des motifs en largeur, c'est-à-dire que les motifs de longueur $i+1$ sont considérés seulement après avoir extrait tous les motifs fréquents de longueur i. Grâce à la propriété d'anti-monotonie de la contrainte de support, les motifs candidats de longueur $i+1$, qui sont les sur-motifs d'au moins un motif fréquent de longueur i, sont les seuls motifs de longueur $i+1$ pouvant être fréquents. Ainsi, nous garantissons une exploration complète en ne regardant que les candidats qui sont en général beaucoup moins nombreux que tous les motifs de la même longueur. Une fois les candidats générés, une phase de comptage du support est alors effectuée afin de vérifier s'ils sont fréquents ou pas. Agrawal et Srikant sont les premiers à introduire le problème d'extraction des motifs séquentiels fréquents, et ils ont proposé 3

algorithmes de type *Apriori* pour le résoudre : *AprioriAll*, *AprioriSome* et *DynamicSome* (Agrawal et Srikant, 1995). Ensuite, les mêmes chercheurs ont proposé l'algorithme *Generalized Sequential Pattern* (GSP) (Srikant et Agrawal, 1996), qui est beaucoup plus rapide, grâce notamment à une génération plus efficace des motifs candidats. Dans cette approche, deux motifs fréquents β et β' d'une longueur i peuvent être fusionnés pour créer des candidats lorsque la condition suivante est satisfaite : la suppression d'un *item* du premier *itemset* dans β donne le même motif obtenu en supprimant un *item* du dernier *itemset* dans β'. Dans le cas le plus simple où chaque *itemset* contient seulement 1 *item*, par exemple $\beta = A \rightarrow B \rightarrow C$, $\beta' = B \rightarrow C \rightarrow D$ [2], le motif candidat est alors $A \rightarrow B \rightarrow C \rightarrow D$. Cette génération permet d'obtenir une liste exhaustive et non répétitive de candidats. L'algorithme s'exécute de façon itérative jusqu'à ce qu'aucun motif candidat ne soit généré. Masseglia *et al.* (1998) ont proposé *Prefix tree for Sequential Pattern* (PSP), dont le principe de fonctionnement est inspiré de GSP, avec des améliorations au niveau de la structure de données pour gérer les motifs candidats et les motifs fréquents. Masseglia *et al.* ont montré que l'utilisation de la structure *prefix-tree* est beaucoup plus efficace par rapport à la structure *hash-tree* utilisée dans GSP, surtout lorsque le seuil de support minimal est petit.

Ces méthodes de type *Apriori* demandent une lecture de toute la base de séquences à chaque phase de comptage du support. Cela les rend très gourmandes en termes d'accès disque. Une solution consiste à stocker la base de séquences en mémoire vive, sous la forme de listes d'occurrences. Comme son nom l'indique, une liste d'occurrences contient pour chacun des *items* l'identifiant des séquences qui le contiennent ainsi que ses dates d'apparition dans chacune de ces séquences. Zaki (2001) a proposé *Sequential PAttern Discovery using Equivalence classes* (SPADE) qui utilise les listes d'occurrences pour extraire les motifs séquentiels fréquents. Il faut seulement deux lectures des données sur disque : une pour trouver tous les *items* fréquents et une pour les motifs fréquents de longueur 2. En effet, les supports des motifs candidats de longueur $k+1$ peuvent être obtenus en utilisant des opérations de jointure sur les listes d'occurrences des motifs fréquents de longueur k. Ayres *et al.* (2002) ont proposé *Sequential PAttern Mining* (SPAM) avec un encodage des listes d'occurrences par des vecteurs de bits. Ceci s'avère très efficace lorsque les motifs apparaissent dans un grand nombre de séquences. Dans ce cas, les listes d'occurrences deviennent très grandes, empêchant ainsi de comparer de façon efficace (avec l'utilisation des listes d'occurrences standard) les éléments de deux listes d'occurrences lors des opérations de jointure. L'utilisation de vecteurs de bits permet de réduire de façon significative l'espace de mémoire utilisé pour stocker les listes d'occurrences et facilite les opérations de jointure. L'inconvénient principal des algorithmes de ces deux premières approches (de type *Apriori* ou par listes d'occurrences) réside dans l'étape de génération des candidats, où un grand nombre de motifs peut être généré lorsque le seuil de support minimal est petit.

La troisième famille de méthodes, basée sur les projections de la base de données, évite la génération des motifs candidats. Leur idée générale est de construire des projections de la base de données, afin d'extraire des motifs sur chacune des projections. Une projection de la base de données par rapport à un préfixe est composée des sous-séquences construites à partir de la fin de la première occurrence du préfixe dans chaque séquence. Par exemple, pour une séquence $s = \langle (t_1, BD), (t_2, AC), (t_3, B), (t_4, ABC) \rangle$, la projection de s par rapport à un préfixe $B \rightarrow A$ sera $\langle (t_2, \_C), (t_3, B), (t_4, ABC) \rangle$. $\_C$ indique que le dernier élément dans le préfixe, dans ce cas A, forme un événement en le combinant avec C. Les algorithmes de ce type sont *Frequent pattern-projected Sequential pattern mining* (FreeSpan) (Han *et al.*, 2000) et *Prefix-projected Sequential pattern mining* (PrefixSpan) (Pei *et al.*, 2004b). Leur avantage

2. En supprimant l'*item* A dans β et l'*item* D dans β', nous obtenons le même motif $B \rightarrow C$.

principal par rapport aux algorithmes précédents est qu'ils génèrent moins de candidats, et évitent de considérer par exemple les motifs qui n'ont pas d'occurrences. La construction des projections est l'étape la plus coûteuse dans cette approche. Ce problème peut être atténué en utilisant soit une projection à deux niveaux pour réduire la taille et le nombre des bases de données projetées, soit une méthode de pseudo-projection lorsqu'une base de données projetée peut être entièrement contenue dans la mémoire principale (Pei *et al.*, 2004b).

3.2.4 Algorithmes d'extraction de motifs sous contraintes

Comme évoqué précédemment, des contraintes supplémentaires sont généralement imposées aux motifs séquentiels. Une solution simple consiste à utiliser un des algorithmes d'extraction de motifs séquentiels fréquents, et ensuite à vérifier pour chacun des motifs extraits s'il satisfait les contraintes introduites. Cette solution, qui ne demande aucune modification dans l'algorithme d'extraction, est souvent une solution très inefficace d'un point de vue technique. En effet, le nombre de motifs satisfaisant les contraintes est généralement très petit par rapport nombre total de motifs en sortie de la phase d'extraction brute. Par conséquent, les algorithmes actuellement utilisés consistent à incorporer des contraintes directement dans les algorithmes d'extraction. Le choix de l'algorithme d'extraction à utiliser ainsi que le niveau de modifications correspondant dépendent fortement du type de contraintes utilisées.

Par exemple, pour la contrainte de durée *maxspan* (cf. Table 3.1), les algorithmes de type *Apriori* peuvent être utilisés avec une vérification de la contrainte pendant la phase de comptage du support. En effet, étant donnés deux motifs β et β' tels que β est un sous-motif de β', nous pouvons facilement constater que le nombre d'occurrences de β satisfaisant la contrainte *maxspan* doit être au moins égal à celui de β'. Par conséquent, si β n'est pas fréquent, β' ne l'est pas non plus. Une telle incorporation de la contrainte dans les algorithmes *Apriori*, avec une modification mineure de ceux-ci, permet ainsi de traiter moins de motifs à chaque étape, rendant le processus plus performant que la solution de base.

D'une manière générale, si une contrainte est anti-monotone, monotone ou succincte, des approches simples, mais efficaces (e.g., Ng *et al.* (1998)) peuvent être appliquées pour prendre en compte cette contrainte en utilisant n'importe quel algorithme d'extraction de motifs séquentiels fréquents parmi ceux présentés dans la section 3.2.3. Les contraintes anti-monotones sont prises en compte de la même manière que l'exemple précédent pour *maxspan*, c'est-à-dire que la vérification pour savoir si un motif la satisfait s'effectue lors du comptage du support. Pour les contraintes monotones, par définition, lorsqu'un motif β les satisfait, tous ses sur-motifs les satisfont également. Par conséquent, nous pouvons réduire considérablement le nombre de vérifications nécessaires lors de l'extraction. Quant aux contraintes succinctes, comme la liste exhaustive des motifs qui les satisfont peut être générée directement à partir de leur formulation, l'exploration peut être guidée directement vers ces motifs. Ainsi, il n'est pas nécessaire de vérifier ces contraintes de façon itérative pendant le processus d'extraction.

Cependant, certaines classes importantes de contraintes, telles que les contraintes d'expressions régulières et les contraintes de la valeur moyenne, de la somme, ne rentrent pas dans ce cadre. L'incorporation de ces contraintes dans les algorithmes d'extraction est plus complexe. La suite de cette section est dédiée aux algorithmes proposés pour prendre en compte ces types de contraintes de façon efficace.

Dans le cadre des contraintes d'expression régulière $\mathcal{C}_{ER}$, Garofalakis *et al.* (1999) ont proposé une famille de quatre algorithmes, appelée *Sequential Pattern mIning with Regular*

expressIon consTraints (SPIRIT). Des *relaxations* sur les contraintes d'expression régulière sont faites afin d'obtenir des contraintes anti-monotones. Les techniques de relaxation de contraintes sont utilisées pour avoir des contraintes qui sont moins strictes [3] que la contrainte originale et qui a des propriétés intéressantes permettant d'accélérer l'extraction. Les algorithmes SPIRIT utilisent une structure algorithmique similaire à celle des algorithmes *Apriori*. C'est-à-dire qu'ils explorent l'espace de recherche en largeur, de façon itérative, avec un paradigme *générer et tester*. Les nouvelles contraintes relaxées, étant anti-monotones, sont très faciles à incorporer dans l'extraction. À l'issue de l'extraction, les motifs satisfaisant les contraintes relaxées nécessitent d'être vérifiés vis-à-vis de la contrainte originale. La différence entre les quatre algorithmes se situe notamment dans la construction des contraintes relaxées, chacune étant plus stricte que la précédente. Le premier algorithme SPIRIT (N) ("N" pour "Naïve") élague uniquement les motifs candidats contenant un ou plusieurs éléments qui n'apparaissent pas dans la contrainte $\mathcal{C}_{ER}$. C'est la contrainte la moins stricte parmi les quatre. Le second, SPIRIT (L) ("L" pour "Légale"), nécessite que le motif candidat β soit *légal* vis-à-vis d'un état de l'automate fini déterministe $\mathcal{A}_{ER}$ associé à $\mathcal{C}_{ER}$. Un motif β est dit *légal* par rapport à un état e de $\mathcal{A}_{ER}$ lorsque toutes les transitions d'état dans $\mathcal{A}_{ER}$ sont définies en parcourant la séquence de transitions des éléments de β depuis e (Garofalakis *et al.*, 1999). Le troisième, SPIRIT (V) ("V" pour "Valide"), élague tous les motifs candidats qui ne valident aucun état de $\mathcal{A}_{ER}$. Un motif β valide un état e de $\mathcal{A}_{ER}$ si β est légal par rapport à e et l'état final obtenu en parcourant la séquence de transitions des éléments de β à partir de e est un *état d'acceptation* [4] de $\mathcal{A}_{ER}$. Le quatrième, SPIRIT (R) ("R" pour "Régulier"), incorpore la contrainte $\mathcal{C}_{ER}$ tout au long du processus d'extraction en comptant le support uniquement pour les motifs candidats acceptés par $\mathcal{A}_{ER}$. Autrement dit, le motif candidat doit être valide par rapport à l'état initial de $\mathcal{A}_{ER}$. Garofalakis *et al.* (1999) ont montré que SPIRIT (R) est très performant lorsque la contrainte $\mathcal{C}_{ER}$ est suffisamment stricte. Dans le cas inverse, le nombre de candidats générés par SPIRIT (R) explose et il devient beaucoup moins performant. C'est SPIRIT (V) qui est recommandé dans la plupart des cas.

Cette approche, bien qu'intéressante, n'est pas systématique pour toutes les contraintes. Au lieu de considérer des techniques de relaxation pour les incorporer dans les algorithmes *Apriori*, comme cela a été le cas pour SPIRIT, d'autres algorithmes, comme notamment cSPADE (Zaki, 2000), Prefix-growth (Pei *et al.*, 2007), ont été proposés pour prendre en compte directement les contraintes dans le processus d'extraction.

L'algorithme cSPADE (Zaki, 2000), une extension de SPADE, permet de prendre en compte les contraintes suivantes : (1) contraintes de longueur/largeur maximale, (2) contraintes de *gap* minimal/maximal, (3) contraintes de fenêtre temporelle maximale [5], (4) contraintes sur les *items*. Comme SPADE, cSPADE utilise une approche par listes d'occurrences. Parmi les contraintes qu'il supporte, seule la contrainte de *gap* maximal n'est pas anti-monotone. Par exemple, avec la contrainte $gap\_max = 2$, le motif $\beta = A \rightarrow B$ n'apparaît pas dans la séquence $s = \langle(1, A)(2, X)(3, S)(4, B)\rangle$, tandis que son sur-motif $\beta' = A \rightarrow X \rightarrow B$ y apparaît. Une nouvelle stratégie pour joindre les listes d'occurrences a été développée dans cSPADE pour cette contrainte.

Dans le cas de Prefix-growth (Pei *et al.*, 2007), l'algorithme s'appuie sur les techniques basées sur les projections de la base de données, comme PrefixSpan (Pei *et al.*, 2004b). Prefix-

3. Une contrainte $\mathcal{C}$ est dite moins stricte qu'une contrainte $\mathcal{C}'$ si tout motif β satisfaisant $\mathcal{C}'$ satisfait également $\mathcal{C}$. L'ensemble des motifs satisfaisant $\mathcal{C}'$ est ainsi un sous-ensemble de l'ensemble des motifs satisfaisant $\mathcal{C}$.

4. Aussi appelé *état final* ou *état terminal*.

5. Ce type de contraintes impose que le motif apparaisse dans une fenêtre de temps.

growth est construit sur la propriété de préfixe-monotonie qui regroupe les contraintes préfixe monotones et préfixe anti-monotones. Puisqu'une contrainte anti-monotone est aussi préfixe anti-monotone, et qu'une contrainte monotone est aussi préfixe monotone, cette propriété offre un cadre qui englobe beaucoup de contraintes : *item*, longueur, largeur, sur-motif, agrégat simple, durée, *gapmin*. En plus, Pei *et al.* (2007) ont montré également que les contraintes d'expression régulière satisfont la propriété de préfixe-monotonie, et parmi les principaux types de contraintes, seules les contraintes d'agrégats difficiles (e.g., contraintes sur la valeur moyenne, sur la somme) ne font pas partie de cette catégorie. L'exploration de l'espace de recherche s'effectue en profondeur et les motifs séquentiels fréquents sont extraits de façon récursive dans les projections de la base de données. Les motifs satisfaisant la contrainte préfixe-monotone sont ensuite gardés, et utilisés à leur tour comme préfixe pour construire de nouvelles projections. De cette façon, l'algorithme est capable de prendre en compte directement la plupart des contraintes, en élaguant les motifs indésirables le plus tôt possible.

Pour les contraintes d'agrégats difficiles, telles que les contraintes sur valeurs moyennes, qui ne satisfont pas la propriété de préfixe-antimonotonie, des extensions des approches par projection, tirant parti des propriétés spécifiques de ces contraintes, ont été réalisées, comme par exemple la version révisée de Prefix-growth (Pei *et al.*, 2007) et l'algorithme proposé par Chen *et al.* (2008).

3.3 Sélection de motifs séquentiels sur complémentarité informationnelle

Un des inconvénients majeurs de l'extraction des motifs fréquents est l'explosion du nombre de ces motifs, surtout lorsque le seuil de support minimal est petit ou lorsque les contraintes additionnelles ne sont pas suffisamment sélectives. Par conséquent, l'ensemble des motifs extraits dans sa globalité ne s'avère généralement pas intéressant en termes d'interprétabilité. Afin de ne sélectionner qu'un ensemble réduit des motifs extraits, tout en gardant sa capacité à résumer la base de données, différentes techniques sont proposées pour les *itemsets* comme le tiling (Geerts *et al.*, 2004; Guns *et al.*, 2013), les motifs skyline (Goyal *et al.*, 2008; Soulet *et al.*, 2011), les motifs maximaux (Gouda et Zaki, 2001; Burdick *et al.*, 2005), les motifs clos (Pasquier *et al.*, 1999) et les motifs optimaux de compression (Vreeken *et al.*, 2011). Ces approches ont été étendues dans le cadre des motifs séquentiels comme Raissi *et al.* (2006); García-Hernández *et al.* (2006); Luo et Chung (2004) pour les motifs maximaux, Yan *et al.* (2003); Wang *et al.* (2007) pour les motifs clos, et Tatti et Vreeken (2012); Lam *et al.* (2014); Ibrahim *et al.* (2016) pour les motifs optimaux de compression. Parmi ces approches, la dernière, qui est une approche informationnelle basée sur la capacité de compresser la base de données en utilisant l'ensemble de motifs choisis, s'avère une direction de recherche très prometteuse. En effet, cette approche sélectionne un ensemble réduit de motifs qui permet de compresser au mieux le jeu de données. De cette façon, les motifs choisis sont souvent fréquents et se complètent également très bien au niveau de leurs localisations dans les données. L'ensemble final des motifs sélectionnés est ainsi moins redondant et permet également une interprétation plus facile. Cette section sera consacrée à la bibliographie concernant les algorithmes de sélection de motifs séquentiels par critère de compression.

3.3.1 Principe de "Longueur de Description Minimale"

Supposons que nous disposions d'un schéma de compression, noté C, permettant de compresser une base de données D en utilisant un ensemble de motifs noté $\mathcal{H}$ et appelé modèle. De façon informelle, le principe de Longueur de Description Minimale (en anglais *Minimal Description Length*) (MDL) indique que le meilleur modèle $\mathcal{H}$, parmi un ensemble de modèles $\mathbf{H}$, est celui qui conduite à la plus courte description des données, c'est-à-dire celui qui compresse le mieux les données (Grünwald, 2007). La longueur de description est calculée comme suit :

$$L_C^{\mathcal{H}}(D) = L_C(\mathcal{H}) + L_C(D|\mathcal{H}) \tag{3.1}$$

où : $L_C(\mathcal{H})$ est la longueur de description de $\mathcal{H}$, et $L_C(D|\mathcal{H})$ est la longueur de description des données encodées par $\mathcal{H}$.

Cette expression est appelée la longueur de description en deux parties, qui est différente de la longueur de description "raffinée" où le modèle et les données sont codés ensemble (Grünwald, 2007). Dans le contexte de la fouille de données, nous souhaitons isoler le modèle, qui est la partie contenant les motifs "intéressants". En outre, bien que la version raffinée ait des fondations théoriques plus solides, elle ne peut être calculée que pour certains cas particuliers (Tatti et Vreeken, 2012). C'est pour ces raisons que seulement la longueur de description en deux parties est utilisée dans les algorithmes de sélection de motifs.

L'objectif final consiste à trouver le modèle qui minimise la longueur de description :

$$\mathcal{H}_{opt} = \underset{\mathcal{H} \in \mathbf{H}}{\operatorname{argmin}} \left(L_C(\mathcal{H}) + L_C(D|\mathcal{H}) \right) \tag{3.2}$$

Il existe un lien direct entre la compression de données et l'extraction de motifs. La première cherche à compresser les données en se servant de régularités qui y sont présentes. Ces régularités correspondent généralement à des motifs qui possèdent des propriétés intéressantes. Par exemple, un algorithme de compression peut découvrir que le motif $A \rightarrow B$ fait partie des régularités permettant de réduire la longueur de description. La recherche de motifs fréquents pourrait, quant à elle, retrouver ce même motif dans la même base de données.

Cependant, l'objectif de la compression et celui de l'extraction de motifs sont très différents l'un de l'autre. En effet, la compression des données cherche les régularités dans le seul but de réduire la taille de la représentation. Par principe, ces régularités ne sont pas nécessairement intéressantes du point de vue de la fouille de données et de ses applications. C'est pour cette raison qu'il faut adapter le principe MDL à des critères d'intérêt de la fouille de données, en utilisant par exemple un schéma d'encodage C tel que les motifs sélectionnés soient complémentaires et peu redondants entre eux.

3.3.2 Schéma d'encodage

À partir de l'équation 3.1, nous pouvons constater que l'élément le plus important dans le principe de MDL est le schéma d'encodage C. En effet, il permet de calculer la longueur de description et détermine le score de chaque modèle.

Vreeken *et al.* (2011); Lam *et al.* (2014) utilisent des schémas d'encodage basés sur des *tables de codage*. Une table de codage $\mathcal{H}$ est techniquement un dictionnaire qui contient des motifs permettant de reconstruire les séquences de données. Pour encoder chaque séquence

s, les occurrences de tous les motifs de $\mathcal{H}$ dans s sont remplacées par les pointeurs associés dans $\mathcal{H}$.

Dans le contexte des *itemsets*, puisque la notion "séquentiel" n'existe pas, la tâche principale consiste à identifier quels éléments dans $\mathcal{H}$ sont utilisés pour encoder chaque transaction. Afin de garantir l'existence et l'unicité d'un ensemble d'éléments pour encoder de façon complète chaque transaction, Vreeken *et al.* (2011) adopte une table de codage contenant au moins les *itemsets* composés d'un seul *item*, et impose que l'ensemble des motifs de la table forme un code préfixe, c'est-à-dire qu'aucun motif ne soit préfixe d'un autre motif (Cover et Thomas, 2006).

Dans le cas des schémas d'encodage pour les données séquentielles, la position des événements dans les séquences doit être prise en compte afin de garantir la reconstruction de données à partir de données compressées. Lam *et al.* (2014) ont utilisé un schéma d'encodage qui consiste à préciser la date d'apparition dans la séquence des événements de chaque motif utilisé pour l'encodage. À titre d'exemple, considérons le dictionnaire $\mathcal{H}$ et la base de données D dans la table 3.2. La séquence s_1 peut être encodée comme suit : $s_1 = [\beta_1|1, 7][\beta_2|3, 4, 9][\beta_3|8]$, où $[\beta_1|1, 7]$ encode les événements $(1, A)$ et $(7, B)$ en utilisant le motif β_1 du dictionnaire, $[\beta_2|3, 4, 9]$ encode les événements $(3, C)$,$(4, A)$ et $(9, B)$ en utilisant le motif β_2, et le dernier événement $(8, C)$ est encodé par $[\beta_3|8]$. Lam *et al.* (2012) utilisent une technique d'encodage pour les bases de séquences de mots où chaque mot est considéré comme un événement. Les dates d'apparition pour ce type de données n'existent pas réellement, et elles sont généralement considérées comme consécutives. Par exemple, la séquence s_2 peut représenter une séquence de mots [6], et peut être encodée par : $[\beta_1|E(2)][\beta_3][\beta_2|E(1), E(1)]$. $[\beta_1|E(2)]$ indique une occurrence de β_1 dont les dates d'apparition du premier et du deuxième événement sont séparées de deux unités de temps, i.e., $(1, A), (3, B)$. L'encodage se poursuit dans l'ordre chronologique. $[\beta_3]$ indique ainsi que le symbole C se place à la première position libre, i.e., $(2, C)$. De la même façon, $[\beta_2|E(1), E(1)]$ encode les événements restants, i.e., $(4, C), (5, A), (6, B)$.

Base de données D	Dictionnaire $\mathcal{H}$
$s_1 = (1, A)(3, C)(4, A)(7, B)(8, C)(9, B)$	$\beta_1 = A \rightarrow B$
$s_2 = (1, A)(2, C)(3, B)(4, C)(5, A)(6, B)$	$\beta_2 = C \rightarrow A \rightarrow B$
	$\beta_3 = C$

Table 3.2 – Exemple d'une base de données et un dictionnaire

Ces schémas d'encodage montrent la différence entre les objectifs de la compression et de la fouille de données. En effet, en les utilisant, nous pouvons nous retrouver dans des situations où les données compressées sont plus volumineuses que les données originales. L'objectif principal n'est pas de compresser au mieux les données mais de trouver un modèle qui contient des motifs intéressants. Par conséquent, il n'est pas essentiel que les données "compressées" soient moins volumineuses que les données d'origine.

3.3.3 Longueur de description

Chaque motif présent dans le dictionnaire est associé à un codage unique dont la longueur permet de déterminer au final la taille de la description de la base de données encodée par ce dictionnaire. Intuitivement, afin d'obtenir une bonne compression, plus les motifs sont

6. Chaque mot est ici représenté abstraitement par un des symboles dans l'ensemble $\{A, B, C\}$.

fréquents, plus leur longueur de codage est petite et inversement. Selon Grünwald (2007), la longueur de description optimale d'un motif β peut être obtenue en utilisant l'entropie de Shannon, et calculée comme suit :

$$E(\beta) = -\log\left(\frac{f(\beta)}{F}\right) \tag{3.3}$$

où $f(\beta)$ indique le nombre d'utilisations de β pour encoder la base de données D, et F indique le nombre total d'utilisations de tous les motifs dans le dictionnaire $\mathcal{H}$.

Le problème d'optimisation (équation 3.2) devient le suivant :

$$\mathcal{H}_{opt} = \underset{\mathcal{H}}{\operatorname{argmin}}\left(L_C(\mathcal{H}) + \sum_{\beta\in\mathcal{H}}\left(-log\frac{f(\beta)}{F} \times f(\beta) + g(\beta)\right)\right) \tag{3.4}$$

où $g(\beta)$ indique le coût additionnel pour encoder les dates d'apparition des événements dans les occurrences.

Algorithmes pour sélectionner les motifs

Suivant Lam *et al.* (2012), le problème de la recherche du dictionnaire optimal pour encoder une base de données est NP-difficile. C'est pour cette raison que les approches actuelles consistent à utiliser des solutions heuristiques en sélectionnant de façon itérative et gloutonne les motifs. Plus précisément, à partir d'un ensemble de motifs candidats, généralement l'ensemble des motifs fréquents, chaque itération ajoute dans le dictionnaire le motif qui forme le meilleur nouveau dictionnaire, c'est-à-dire celui qui compresse le mieux les données. Ces itérations s'effectuent jusqu'à ce qu'un nombre de motifs défini par l'utilisateur soit obtenu. Ce principe est utilisé dans le cadre des *itemsets* (Vreeken *et al.*, 2011) ainsi que pour les motifs séquentiels (Lam *et al.*, 2014). Ces algorithmes requièrent l'extraction des motifs séquentiels fréquents, une étape qui consomme souvent beaucoup de ressources de calcul. Afin de limiter ce coût, des algorithmes d'extraction directe ont également été proposés (Vreeken *et al.*, 2011; Lam *et al.*, 2014). Ces algorithmes réalisent simultanément l'extraction de motifs et la compression de données.

Les travaux de Lam *et al.* (2014) sont basés sur Lam *et al.* (2012), qui avait précédé de peu une proposition de Tatti et Vreeken (2012), basée elle aussi sur des algorithmes heuristiques simples de type glouton pour ajouter aux motifs sélectionnés, à chaque itération, le motif permettant de compresser le mieux les séquences. Afin de calculer la longueur de l'encodage d'une séquence s avec l'utilisation d'une table de codage $\mathcal{H}$, Tatti et Vreeken (2012) utilisent deux séquences de codage, C_p et C_g, qui contiennent respectivement les codes des motifs et les codes des *gaps*. Par exemple, pour une séquence $s = \langle(1, A)(2, B)(3, D)(4, C)(5, A)\rangle$ et une table de codage contenant les motifs $\beta_1 = A$, $\beta_2 = A \to D \to C$, $\beta_3 = B$, s sera encodé par les séquences de codage suivantes :

$$C_p = \hat{\beta}_2\hat{\beta}_3\hat{\beta}_1$$
$$C_g = g\bar{g}\bar{g}$$

où $\hat{\beta}$ indique le code associé à β, g indique le code pour un *gap* entre les éléments consécutifs du motif et $\bar{g}$ est le code pour l'absence de *gap*. En utilisant à la fois les deux séquences C_p, C_g

et la table de codage $\mathcal{H}$ associée, Tatti et Vreeken ont montré que nous pouvons facilement retrouver la séquence s. En commençant par une séquence vide s_k, nous rencontrons en premier le code du motif β_2 en lisant C_p, le premier élément de β_2 (*item* A dans ce cas) est alors ajouté à s_k. Pour un motif non-singleton comme β_2, comme il peut y avoir des *gaps* entre les éléments consécutifs dans une occurrence, nous continuons la lecture sur la séquence C_g pour déterminer s'il existe des *gaps* ou pas. Si un *gap* (codé par g) est rencontré, les éléments du motif suivant dans la séquence C_p sont ajoutés à leur tour dans s_k. Dans notre cas, nous ajoutons l'*item* B à la fin de s_k. Le processus se poursuit ensuite de façon similaire jusqu'à régénération de s. Dans les cas où la base de données D contient plusieurs séquences, cette technique d'encodage nécessite de coder également la longueur de chaque séquence en mesurant la longueur d'encodage de C_p et C_g à partir de l'entropie de Shannon.

Plus récemment, Ibrahim *et al.* (2016) ont proposé des motifs du type motifs séquentiels appelés *épisodes série avec intervalles fixes* (*fixed interval serial episodes* en anglais) pour encoder les séquences. Un tel motif sera par exemple $\beta = A \xrightarrow{2} B \xrightarrow{3} C$, où les valeurs $2, 3$ représentent les *gaps* que doivent respecter les événements d'une occurrence. Soit la séquence $s = \langle (1, A)(2, A)(3, B)(4, E)(5, A)(6, B)(7, B)(8, D)(10, C), (11, E) \rangle$, les occurrences de β sont $\langle 1, 3, 6 \rangle$ et $\langle 5, 7, 10 \rangle$. De cette façon, l'occurrence d'un motif peut être représentée seulement par la date d'apparition du premier symbole, ensuite, la position des autres symboles peut être déduite facilement. Ainsi, le coût d'encodage des *gaps* pour chaque occurrence disparaît. Par conséquent, ce type d'épisode permet de compresser mieux la base de données par rapport aux approches précédentes, lorsque l'on ne tient pas compte de la taille d'encodage $\mathcal{H}$.

3.4 Conclusions

Dans ce chapitre, nous avons étudié les techniques d'extraction de motifs séquentiels sous contraintes ainsi que les techniques de sélection de motifs séquentiels complémentaires au niveau informationnel. L'extraction de motifs séquentiels sous contraintes présente des intérêts non seulement applicatifs mais également techniques. En effet, elle permet de prendre en compte les critères d'intérêt que peut introduire l'utilisateur final pour extraire les motifs potentiellement plus pertinents. Du point de vue technique, en prenant en compte des propriétés spécifiques comme la monotonie, le processus d'extraction peut devenir beaucoup plus rapide en optimisant l'utilisation des ressources de calcul. La sélection de motifs séquentiels complémentaires au niveau informationnel en utilisant le principe MDL est, quant à elle, un outil performant permettant de sélectionner seulement un ensemble de taille raisonnable de motifs. En utilisant des schémas d'encodage adaptés ainsi que des algorithmes heuristiques efficaces, cette approche fournit généralement des motifs qui représentent bien les données, et qui sont complémentaires. Ces critères sont indispensables pour une interprétation plus aisée des motifs par l'utilisateur final. Nous observons pour autant que les méthodes existantes ne permettent pas d'exploiter les indices de confiance disponibles. Les deux chapitres suivants présentent nos propositions pour (1) extraire seulement les motifs qui sont fiables en prenant en compte les indices de confiance exprimés sous forme de contraintes et (2) sélectionner un ensemble de motifs complémentaires selon des critères informationnels basés sur les indices de confiance.

Chapitre 4

Extraction de motifs SFG avec prise en compte des indices de confiance

Sommaire

4.1 Introduction

Les indices de confiance associés à des mesures de déplacement nécessitent d'être considérées pendant le processus d'extraction afin de sélectionner seulement les évolutions de déplacements auxquelles nous pouvons accorder une certaine fiabilité. Or, la technique de fouille de données basée sur les motifs séquentiels actuellement utilisée pour analyser les Séries Temporelles de Champs de Déplacements (STCD) (Pericault *et al.*, 2015) n'est pas capable de les prendre en compte.

Les techniques d'extraction de motifs sous contraintes vues au chapitre précédent pourraient être appliquées dans ce contexte, en utilisant des contraintes basées sur les indices de confiance. Ce chapitre est ainsi consacré à notre proposition pour intégrer ces contraintes au sein de l'extraction des motifs Séquentiels Fréquents Groupés (SFG) (Julea *et al.*, 2011), dans le but d'extraire seulement les motifs qui occupent des zones de données disposant d'un niveau de confiance suffisant. Pour ce faire, des indices de confiance d'un motif séquentiel sont introduits sur différents niveaux : chaque occurrence, chaque séquence dans laquelle le motif apparaît, et finalement la base de séquences. Afin de déterminer de façon efficace l'indice de confiance de chacun des motifs, un algorithme basé sur la programmation dynamique est également proposé. Néanmoins, l'espace de recherche reste très grand pour nos jeux de données habituels, avec des configurations d'extraction standard. Afin d'atténuer ce problème, une technique de *push* partiel utilisant les propriétés des indices de confiance est proposée pour réduire l'espace de recherche, tout en préservant la correction de l'algorithme d'extraction. Finalement, pour valider nos propositions, les expériences sur deux Séries Temporelles de Champs de Déplacements (STCD) sont présentées : une sur le Groenland provenant de données optiques, et une sur le massif du Mont-Blanc provenant de données Radar à Synthèse d'Ouverture (en anglais *Synthetic Aperture Radar*) (SAR).

4.2 Représentation des STCD

Dans ce travail de thèse, les mesures de déplacement sont estimées exclusivement à partir des images satellitaires. Nous avons vu, dans le chapitre 2, les techniques de télédétection permettant de le faire. Ces mesures peuvent être de différents types (le vecteur de déplacement ayant lieu entre deux dates d'acquisition, la vitesse de déplacement), et dans des directions différentes (en *range* et en azimut, ou en ligne de visée (en anglais *Line Of Sight*) (LOS)). Sans perte de généralité, chaque mesure de déplacement sera considérée comme un vecteur, dont on dispose à la fois de la direction et de l'amplitude. Une STCD se compose de plusieurs champs de déplacements, dont chacun correspond à une période distincte, sur une même zone géographique. Ces champs de déplacements sont des matrices 2D de la même taille, avec une superposition parfaite des cellules, i.e., les termes de mêmes indices représentent les informations sur les déplacements d'une même zone, mais à différentes dates.

Dans ces travaux de thèse, nous nous intéressons non seulement aux mesures de déplacement elles-mêmes, mais également à leur niveau de confiance. En effet, comme nous l'avons vu dans le chapitre 2, chaque mesure de déplacement peut être associée à un indice de confiance, qui peut être issu du calcul du déplacement lui-même, ou obtenu *a posteriori* en qualifiant la cohérence spatiotemporelle des champs, ou encore déduite à partir de la nature physique du type de déplacement. Par principe, cet indice nous indique à quel point il est possible d'avoir confiance dans l'estimation de déplacement fournie. De façon générale, pour une donnée de

déplacement définie, plus son indice de confiance est élevé, plus l'estimation du déplacement est fiable.

Dans la suite de cette section, nous allons définir de façon plus précise les STCD sous la forme de *bases de séquences*.

Définition 4.1 (Indice de confiance). Un *indice de confiance*, noté ρ, est une mesure associée à chaque donnée de déplacement. Plus ρ est grand, plus nous sommes confiants vis-à-vis de la donnée. Sans perte de généralité, nous supposons que ρ est compris entre 0 et 1. Ces limites correspondent respectivement aux situations où l'on n'a aucune confiance et où l'on est parfaitement sûr de la mesure fournie.

Comme nous l'avons vu dans le chapitre 2, plusieurs méthodes existent pour estimer les indices de confiance dans le contexte des STCD, dépendant à la fois de la nature des données et en même temps de la technique d'estimation de déplacement. Cependant, dans ce travail, afin de préserver la généralité de notre approche, nous ne fixons pas la méthode de calcul des confiances.

Définition 4.2 (Événement). Soit $I = \{i_1, i_2, ..., i_n\}$ un ensemble contenant n symboles appelés *items* (dans ce qui suit, les items seront en général représentés par des nombres entiers). Ces symboles correspondent aux valeurs discrètes que peut prendre une mesure de déplacement. Comme chaque mesure de déplacement est considérée comme un vecteur, les grandeurs comme l'amplitude ou l'angle peuvent être utilisées pour la discrétisation. Un *événement* est un triplet (t, α, ρ) où t correspond à la date d'événement, $\alpha \in I$ est le symbole représentant le déplacement estimé, et $\rho \in [0, 1]$ est l'indice de confiance associé à l'événement.

Définition 4.3 (Séquence d'événements). Une *séquence d'événements* représente les déplacements date par date sur un pixel de coordonnées (x, y) de la STCD. Elle est notée $seq(x, y) = \langle (t_1, \alpha_1, \rho_1), (t_2, \alpha_2, \rho_2), \ldots, (t_n, \alpha_n, \rho_n) \rangle$, où (x, y) correspond aux coordonnées de la séquence, et les dates d'événement sont dans l'ordre chronologique $(t_1 < t_2 < \cdots < t_n)$.

Définition 4.4 (Base de séquences, STCD symbolique). Une *base de séquences*, notée D, est un ensemble de paires (sid, s) où s représente une séquence et sid correspond à son identifiant. Dans le contexte des STCD, les coordonnées (x, y) peuvent représenter l'identifiant de la séquence associée. Comme les estimations de déplacement sont discrétisées en utilisant des symboles, cette base de séquences D représente une STCD dite *symbolique*.

4.3 Les motifs SFG

Dans cette section, nous rappelons les définitions des motifs Séquentiels Fréquents Groupés (SFG) et des cartes de Localisation Spatio-Temporelle (LST) utilisés dans Julea *et al.* (2011) et Méger *et al.* (2015).

Définition 4.5 (Motif séquentiel). Un *motif séquentiel* β est un *m-uplet* $\langle \beta_1, \beta_2, ..., \beta_m \rangle$ d'éléments appartenant à l'ensemble des *items* I, avec m la longueur de β. Un tel motif sera également noté $\beta_1 \rightarrow \beta_2 \rightarrow ... \rightarrow \beta_m$.

Définition 4.6 (Occurrence, date d'occurrence). Soit $seq(x, y) = \langle (t_1, \alpha_1, \rho_1), (t_2, \alpha_2, \rho_2), \ldots, (t_n, \alpha_n, \rho_n) \rangle$ une séquence d'événements et $\beta = \beta_1 \rightarrow \beta_2 \rightarrow ... \rightarrow \beta_m$ un motif séquentiel. S'il existe $i_1 < i_2 < \cdots < i_m$ tels que $\beta_1 = \alpha_{i_1}, \beta_2 = \alpha_{i_2}, \ldots, \beta_m = \alpha_{i_m}$, $o = \langle t_{i_1}, t_{i_2}, \ldots, t_{i_m} \rangle$ est alors une *occurrence* de β dans $seq(x, y)$. Dans ce cas, nous appelons $t_{i_1}, t_{i_2}, \ldots, t_{i_m}$ les *dates d'occurrence*.

Définition 4.7 (Couverture, support). Lorsqu'il existe au moins une occurrence d'un motif β dans une séquence d'événements $seq(x, y)$, cette dernière est dite *couverte* par β. L'ensemble des séquences couvertes par β représente sa *couverture*, notée $cover(\beta)$. Le *support* de β dans la base de séquences D, noté $support(\beta)$, est le nombre de séquences dans D couvertes par β, i.e., $support(\beta) = |cover(\beta)|$.

Définition 4.8 (Motif séquentiel fréquent). Soit σ un entier naturel non nul qualifié de *seuil de support*. Soit β un motif séquentiel, β est un *motif séquentiel fréquent* lorsque $support(\beta) \geq \sigma$. On peut également définir un *seuil de support relatif* $\sigma_r \in [0, 1]$. Dans ce cas, le motif séquentiel β est fréquent si $\frac{support(\beta)}{|D|} \geq \sigma_r$, où $|D|$ représente le nombre de séquences dans D.

Définition 4.9 (Connexité locale (CL)). Pour une STCD symbolique, on considère une fonction $isCovered(seq(x, y), \beta)$ qui, pour une séquence $seq(x, y)$ et un motif séquentiel β, renvoie 1 si $seq(x, y)$ est couverte par β et 0 dans le cas contraire. Si β couvre $seq(x, y)$, alors sa *connexité locale* en (x, y) est :

$$CL((x, y), \beta) = \left[\sum_{i=-1}^{i=1} \sum_{j=-1}^{j=1} isCovered(seq(x + i, y + j), \beta) \right] - 1$$

La valeur $CL((x, y), \beta)$ est égale au nombre de pixels couverts par β dans le voisinage immédiat de (x, y) (parmi ces 8 plus proches voisins).

Définition 4.10 (Connexité moyenne (CM)). La *connexité moyenne* du motif β est définie comme suit :

$$CM(\beta) = \frac{\sum_{(x,y) \in cover(\beta)} CL((x, y), \beta)}{support(\beta)}$$

Cette mesure donne pour les pixels couverts par β le nombre moyen de pixels, dans leurs 8 plus proches voisins, aussi couverts par β.

Définition 4.11 (Motif Séquentiel Fréquent Groupé (SFG)). Soit D une base de séquences, β un motif séquentiel fréquent dans D et κ un seuil de connexité moyenne. Le motif β est qualifié de motif *Séquentiel Fréquent Groupé (SFG)* si $CM(\beta) \geq \kappa$. Cette contrainte garantit qu'au niveau spatial, les séquences couvertes par β ne soient pas trop isolées en moyenne.

Définition 4.12 (Sous-motif). Soit $\beta = \beta_1 \to \beta_2 \to \cdots \to \beta_k$ et $\beta' = \beta'_1 \to \beta'_2 \to \cdots \to \beta'_{k'}$ deux motifs séquentiels tels que $k' < k$. Le motif β' est un *sous-motif* de β s'il existe des entiers $1 \leq i_1 < i_2 < \cdots < i_{k'} \leq k$ tels que $\beta'_1 = \beta_{i_1}, \beta'_2 = \beta_{i_2}, \ldots, \beta'_{i_{k'}} = \beta_{i_k}$. De façon informelle, β' est un sous-motif de β s'il peut être obtenu à partir de β en enlevant des symboles. Dans ce cas, la notation suivante est utilisée : $\beta' \prec \beta$.

L'opérateur "$\prec$", appelé *opérateur d'inclusion*, peut être également utilisé pour indiquer qu'une occurrence est *incluse* dans une autre occurrence, comme ci-dessous :

Définition 4.13 (Inclusion des occurrences). Soit $o = \langle t_1, t_2, \ldots, t_m \rangle$ et $o' = \langle t'_1, t'_2, \ldots, t'_{m'} \rangle$ deux occurrences de deux motifs séquentiels dans une même séquence telles que $m' < m$. o' est dite *incluse* dans o, notée $o' \prec o$, lorsqu'il existe des entiers $1 \leq i_1 < i_2 < \cdots < i_{m'} \leq m$ tels que $t'_1 = t_{i_1}, t'_2 = t_{i_2}, \ldots, t'_{m'} = t_{i_{m'}}$.

Définition 4.14 (Motif maximal). Un motif β dans une collection de motifs $\mathcal{C}$ est appelé *motif maximal* lorsque β n'est sous-motif d'aucun motif dans $\mathcal{C}$, i.e., $\nexists \beta' \in \mathcal{C} : \beta \prec \beta'$. Sélectionner les motifs maximaux permet de se concentrer uniquement sur les évolutions les plus spécifiques parmi les motifs extraits.

Par exemple, les motifs $2 \to 1$, $1 \to 2$ et $2 \to 2$ sont inclus dans $2 \to 1 \to 2$. Par conséquent, $2 \to 1 \to 2$ est le seul motif maximal de l'ensemble $\{2 \to 1, 1 \to 2, 2 \to 2, 2 \to 1 \to 2\}$.

Définition 4.15 (Carte de Localisation Spatio-Temporelle (LST)). Pour visualiser l'emplacement spatio-temporel d'un motif β, les cartes de LST sont utilisées. La carte LST du motif β est simplement une image où un pixel (x, y) est noir si la séquence associée n'est pas couverte par β, et est affiché avec une couleur indiquant la date de fin au plus tôt de la première occurrence de β dans $seq(x, y)$ dans le cas contraire.

4.4 Propositions

D'après notre étude bibliographique (cf. Section 2.5), l'approche par motifs SFG a montré sa capacité à représenter les évolutions de déplacements dans les STCD. Par contre, la méthode actuellement utilisée pour analyser les STCD ne prend pas en compte les indices de confiance associés aux mesures de déplacement. Dans cette section, nous allons présenter nos contributions pour adapter les motifs SFG aux données *incertaines*[1] au travers de trois aspects : (1) des *mesures de fiabilité*, (2) un algorithme basé sur la programmation dynamique pour un calcul efficace de ces mesures, et (3) une technique de *push* partiel pour réduire l'espace de recherche.

4.4.1 Mesures de fiabilité

Afin de définir une mesure de fiabilité globale (à l'échelle de la base de séquences) pour chaque motif séquentiel, nous avons besoin de différentes mesures à des niveaux plus bas, i.e., celui de l'occurrence et de la séquence.

Définition 4.16 (Fiabilité de l'occurrence). À l'échelle de l'occurrence, la mesure de fiabilité adopte un point de vue conservateur, associant à une occurrence o la confiance minimale des événements formant o dans la séquence concernée. Elle est notée ρ_{occ}, et définie par :

$$\rho_{occ}(seq(x, y), o) = \min \{\rho(x, y, t) \mid t \textit{ in tuple } o\}$$

où $\rho(x, y, t)$ représente l'indice de confiance de l'événement à la date t et dans la séquence $seq(x, y)$.

Pour une occurrence donnée, cette mesure représente tout simplement la confiance minimale dont nous disposons sur chacun de ses éléments.

Définition 4.17 (Fiabilité d'un motif dans une séquence). À l'échelle de la séquence, la mesure de fiabilité d'un motif β dans une séquence $seq(x, y)$ vise à représenter dans quelle

1. Les données incertaines sont dans notre cas les données qui contiennent des indices de confiance.

mesure $seq(x, y)$ contient une occurrence de bonne fiabilité du motif β. Si β couvre l'emplacement (x, y), cette mesure est alors calculée comme suit :

$$\rho_{pat}(seq(x, y), \beta) = \max_{o \in \mathcal{O}(seq(x,y),\beta)} \{\rho_{occ}(seq(x, y), o)\}$$

où $\mathcal{O}(seq(x, y), \beta)$ représente l'ensemble des occurrences de β dans $seq(x, y)$. La valeur de ρ_{pat} représente donc la meilleure fiabilité des occurrences dans la séquence. Lorsque l'ensemble $\mathcal{O}(seq(x, y), \beta)$ est vide, la mesure de fiabilité sera fixée à 0.

Exemple 4.4.1. La figure 4.1 représente la mesure de fiabilité du motif $\beta = 1 \rightarrow 3 \rightarrow 2$ dans la séquence $s = \langle(1, \mathbf{1}, 0.5), (2, \mathbf{3}, 0.8), (3, \mathbf{2}, 0.2), (4, \mathbf{1}, 0.6), (5, \mathbf{2}, 0.4), (6, \mathbf{3}, 0.7), (7, \mathbf{2}, 0.1)\rangle$. La ligne verte représente la mesure de fiabilité β dans les sous-séquences de s qui commencent par le premier événement de s et qui finissent à des instants t différents, notées $s_{\rightarrow t}$ (avec $t = 1, 2, \ldots, 7$).

Pour $t = 1$ et $t = 2$, nous observons que les sous-séquences correspondantes, $s_{\rightarrow 1} = \langle(1, \mathbf{1}, 0.5)\rangle$, $s_{\rightarrow 2} = \langle(1, \mathbf{1}, 0.5), (2, \mathbf{3}, 0.8)\rangle$, ne sont pas couvertes par β. Par conséquent, la mesure de fiabilité du motif β dans ces sous-séquences est fixée à la valeur minimale de confiance, $\rho_{pat}(s_{\rightarrow 1}, \beta) = \rho_{pat}(s_{\rightarrow 2}, \beta) = 0$. À partir du troisième événement de s, nous observons au moins une occurrence de β. Par exemple, pour $s_{\rightarrow 3}$ et $s_{\rightarrow 4}$, il y a une seule occurrence de β, i.e., $\mathcal{O}(s_{\rightarrow 3}, \beta) = \{\langle 1, 2, 3\rangle\} = \mathcal{O}(s_{\rightarrow 4}, \beta)$. D'après la définition 4.16, la mesure de fiabilité de cette occurrence vaut 0.2 (la valeur de confiance du 3^{e} événement de la séquence). Pour $s_{\rightarrow 5}$ et $s_{\rightarrow 6}$, deux occurrences de β sont observées : $\mathcal{O}(s_{\rightarrow 5}, \beta) = \{\langle 1, 2, 3\rangle, \langle 1, 2, 5\rangle\} = \mathcal{O}(s_{\rightarrow 6}, \beta)$. La 2^{e} occurrence, $\langle 1, 2, 5\rangle$, a une valeur de fiabilité de 0.4. La fiabilité du motif dans ces deux sous-séquences $s_{\rightarrow 5}$ et $s_{\rightarrow 6}$ est alors de 0.4 d'après la définition 4.17. C'est aussi la valeur de fiabilité du motif dans la séquence s parce qu'on ne peut trouver aucune autre occurrence dont la fiabilité est meilleure. En effet, pour la sous-séquence $s_{\rightarrow 7}$ (qui est la séquence elle-même), les occurrences possibles contenant le dernier événement $(7, \mathbf{2}, 0.1)$ n'auront en aucun cas une mesure de fiabilité qui dépasse l'indice de confiance associé à cet événement, c'est-à-dire 0.1.

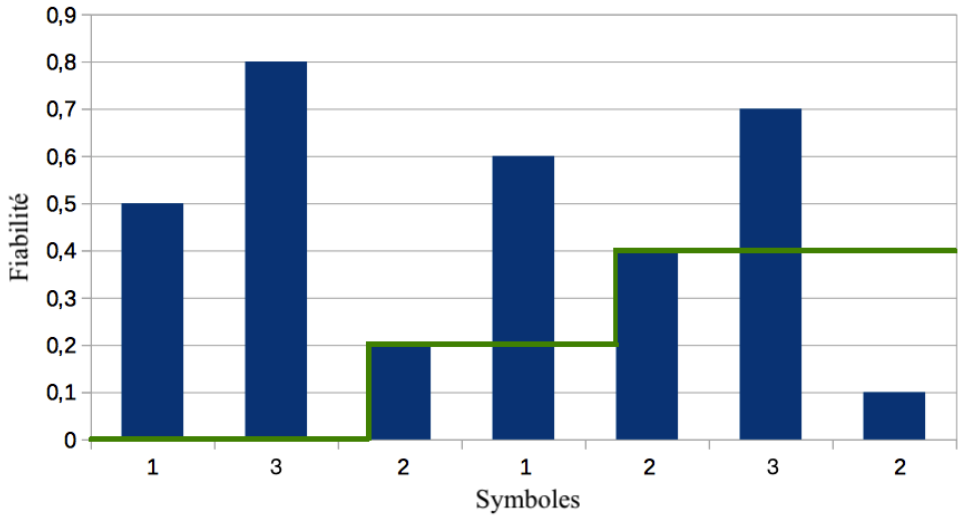


FIGURE 4.1 – Fiabilité du motif $1 \rightarrow 3 \rightarrow 2$ dans la séquence de symboles $1, 3, 2, 1, 2, 3, 2$

Définition 4.18 (Fiabilité du motif dans la base de séquences). À l'échelle de la base de séquences, la mesure de fiabilité ρ retenue pour un motif β est la moyenne de ses mesures

de fiabilité sur toutes les séquences couvertes par β :

$$\rho(\beta) = \frac{\sum_{seq(x,y) \in cover(\beta)} \rho_{pat}(seq(x,y), \beta)}{support(\beta)}$$

Définition 4.19 (Motif fiable, contrainte de fiabilité). Soit $\gamma \in \mathbb{R}_+$ un nombre réel positif qualifié de *seuil de fiabilité*, soit β un motif séquentiel, on dit que β est un *motif fiable* si sa mesure de fiabilité au niveau de la base de séquences est au moins égale à γ, i.e., $\rho(\beta) \geq \gamma$. La contrainte associée $C_\rho(\beta) \equiv \rho(\beta) \geq \gamma$ est appelée *contrainte de fiabilité*.

4.4.2 Recherche des occurrences les plus fiables

Afin de savoir si un motif β est fiable ou pas, nous avons besoin de calculer sa mesure de fiabilité dans chaque séquence $seq(x, y)$ qu'il couvre. À partir de la définition 4.17, cette tâche peut être accomplie en deux étapes :

1. lister toutes ses occurrences dans $seq(x, y)$ avec pour chacune des occurrences la mesure de fiabilité correspondante ;
2. prendre la valeur maximale parmi les mesures de fiabilité de toutes ces occurrences.

Cette approche est simple à comprendre puisqu'elle suit exactement les idées de la définition. Néanmoins, elle n'est pas efficace d'un point de vue algorithmique. En effet, elle nécessite d'isoler toutes les occurrences et pour chacune d'elles, un parcours sur tous les éléments la constituant doit être réalisé pour trouver la valeur minimale des indices de confiance (que l'on considère comme la fiabilité de l'occurrence).

Pour remédier à ce problème, nous proposons un algorithme basé sur la programmation dynamique. Le principe de cet algorithme consiste à diviser le problème en plusieurs sous problèmes, plus faciles à résoudre. Plus précisément, au lieu de calculer directement la mesure de fiabilité d'un motif β dans une séquence s, nous pouvons déduire cette mesure à partir de la fiabilité calculée pour des préfixes de β dans des préfixes de s.

Considérons une séquence s d'une longueur n et un motif β d'une longueur m, avec $m \leq n$. Une matrice C de taille $(m + 1) \times (n + 1)$ est construite. Dans la matrice C, chaque élément noté $c[i, j]$, dont les indices commencent à 0, possède la mesure de fiabilité du motif $\beta_{\rightarrow i}$ contenant les i premiers éléments de β dans la séquence $s_{\rightarrow j}$ contenant les j premiers éléments de s. Le but final consiste à trouver la valeur de l'élément $c[m, n]$, qui exprime la mesure de fiabilité du motif β dans la séquence s. Le pseudo-code de l'algorithme est présenté ci-dessous.

Input : Séquence $s = \langle (t_1, \alpha_1, \rho_1), (t_2, \alpha_2, \rho_2), \ldots, (t_j, \alpha_j, \rho_j), \ldots, (t_n, \alpha_n, \rho_n) \rangle$
Motif $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \ldots \beta_i \rightarrow \cdots \rightarrow \beta_m$
Output : Mesure de fiabilité $\rho_{pat}(s, \beta)$

1 Création des éléments $c[i, j]$ de la matrice C, avec $0 \leq i \leq m$ et $0 \leq j \leq n$;
2 $c[0, j] \leftarrow 1$, pout tout $j \in \mathbb{N}$ tel que $0 \leq j \leq n$;
3 $c[i, j] \leftarrow 1$, pour tout $(i, j) \in \mathbb{N}^2$ tel que $(0 \leq i \leq m)$ et $(0 \leq j \leq n)$ et $(j < i)$;
4 **for** $i \leftarrow 1$ **to** m **do**
5 **for** $j \leftarrow i$ **to** n **do**
6 $c[i, j] \leftarrow \begin{cases} \max\{\min\{c[i-1, j-1], \rho_j\}, c[i, j-1]\} & \textbf{Si } \alpha_j = \beta_i \\ c[i, j-1] & \textbf{Sinon} \end{cases}$
7 **end**
8 **end**
9 **return** $c[m, n]$

Algorithme 1 : Calcul de la fiabilité d'un motif dans une séquence

L'algorithme 1 commence par une étape d'initialisation de la matrice C, qui consiste à mettre en place les deux configurations suivantes :

1. À la ligne 2, pour un motif de longueur 0, la mesure de fiabilité est fixée à la valeur maximale que peut prendre l'indice de confiance, c'est-à-dire 1. Intuitivement, peu importe la séquence, la fiabilité d'un motif vide est toujours maximale.
2. À la ligne 3, pour un motif dont la longueur est supérieure à celle de la séquence, la mesure de fiabilité correspondante est fixée à la valeur minimale que peut prendre l'indice de confiance, c'est-à-dire 0. En effet, il n'existe aucune occurrence d'un motif dans une séquence dont la taille est inférieure à celle du motif.

Une fois la matrice initialisée, afin de trouver la valeur de l'élément $c[m, n]$, nous avons besoin de calculer pour chacune des positions non initialisées de la matrice la mesure de fiabilité associée. Le parcours s'effectue de façon à ce que la valeur des positions $[i-1, j-1]$ et $[i, j-1]$ soit calculée avant celle de la position $[i, j]$. Une telle condition peut être garantie en utilisant par exemple les boucles dans les lignes 4 et 5 de l'algorithme 1. Pour une position (i, j), il y a deux cas de figure possibles :

1. Si le j^{ieme} élément de la séquence s est égal au i^{ieme} élément du motif β, la fiabilité du motif $\beta_{\rightarrow i}$ dans la séquence $s_{\rightarrow j}$ est la valeur maximale entre la mesure de fiabilité des occurrences du motif qui contiennent le j^{ieme} élément de la séquence (i.e., $\min\{c[i-1, j-1], \rho_j\}$), avec ρ_j l'indice de confiance associé à α_j, et des occurrences ne contenant pas le j^{ieme} élément de la séquence (i.e., $c[i, j-1]$). Ce calcul permet d'évaluer la valeur maximale de fiabilité parmi toutes les occurrences de β dans la sous-séquence, respectant ainsi les définitions 4.16 et 4.17.
2. Si le j^{ieme} élément de la séquence s n'est pas égal au i^{ieme} élément du motif β, la fiabilité du motif $\beta_{\rightarrow i}$ dans la séquence $s_{\rightarrow j}$ est la valeur de fiabilité des occurrences ne contenant pas le j^{ieme} élément de la séquence (i.e., $c[i, j-1]$). En effet, dans ce cas, puisque l'ensemble d'occurrences de β dans $s_{\rightarrow j}$ reste inchangé par rapport à celui dans $s_{\rightarrow j-1}$, la mesure de fiabilité de β dans les deux sous-séquences est identique.

Exemple 4.4.2. Afin de montrer étape par étape la construction de la matrice C, nous reprenons l'exemple présenté dans la figure 4.1 pour la séquence $s = \langle (1, \mathbf{1}, 0.5), (2, \mathbf{3}, 0.8), (3, \mathbf{2}, 0.2), (4, \mathbf{1}, 0.6), (5, \mathbf{2}, 0.4), (6, \mathbf{3}, 0.7), (7, \mathbf{2}, 0.1) \rangle$ et le motif $\beta = 1 \rightarrow 3 \rightarrow 2$.

La matrice C après l'étape d'initialisation contient les valeurs suivantes :

$$C = \begin{array}{c} \\ \emptyset \\ 1 \\ 1 \to 3 \\ 1 \to 3 \to 2 \end{array} \begin{array}{c} \begin{array}{cccccccc} \emptyset & 1 & 3 & 2 & 1 & 2 & 3 & 2 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & - & - & - & - & - & - & - \\ 0 & 0 & - & - & - & - & - & - \\ 0 & 0 & 0 & - & - & - & - & - \end{pmatrix} \end{array}$$

Où "$-$" représente les valeurs qui ne sont pas encore calculées.

Ensuite, pour $i = 1$ (motif singleton 1), et j de 1 à 7, la matrice est complétée de la façon suivante :

$$C = \begin{array}{c} \\ \emptyset \\ 1 \\ 1 \to 3 \\ 1 \to 3 \to 2 \end{array} \begin{array}{c} \begin{array}{cccccccc} \emptyset & 1 & 3 & 2 & 1 & 2 & 3 & 2 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0.5 & 0.5 & 0.5 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0 & 0 & - & - & - & - & - & - \\ 0 & 0 & 0 & - & - & - & - & - \end{pmatrix} \end{array}$$

Dans la séquence s, il y a seulement 2 occurrences de ce motif singleton 1 qui sont $o_1 = \langle 1 \rangle$ et $o_2 = \langle 4 \rangle$. L'intuition correspondant au traitement réalisé est la suivante. En constatant que la mesure de fiabilité de o_1 vaut 0.5 et celle de o_2 vaut 0.6, nous pouvons déduire que pour les séquences $s_{\to j}$ avec $1 \leq j \leq 3$, la mesure de fiabilité vaut 0.5 ; et que pour les séquences $s_{\to j}$ avec $j \geq 4$, la mesure de fiabilité vaut 0.6.

Puis, pour $i = 2$ (motif $1 \to 3$) et j de 2 à 7, nous obtenons la matrice suivante :

$$C = \begin{array}{c} \\ \emptyset \\ 1 \\ 1 \to 3 \\ 1 \to 3 \to 2 \end{array} \begin{array}{c} \begin{array}{cccccccc} \emptyset & 1 & 3 & 2 & 1 & 2 & 3 & 2 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0.5 & 0.5 & 0.5 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0 & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0.6 & 0.6 \\ 0 & 0 & 0 & - & - & - & - & - \end{pmatrix} \end{array}$$

Pour ces valeurs de i et j, trois occurrences sont présentes : $o'_1 = \langle 1, 2 \rangle$, $o'_2 = \langle 1, 6 \rangle$, et $o'_3 = \langle 4, 6 \rangle$. Les mesures de fiabilité associées à ces occurrences sont respectivement 0.5, 0.5 et 0.6. Par conséquent, pour les séquences $s_{\to j}$ avec $1 \leq j \leq 5$, la mesure de fiabilité vaut 0.5 et pour les séquences $s_{\to j}$ avec $j \geq 6$, la mesure de fiabilité vaut 0.6.

Finalement, pour $i = 3$ et j de 3 à 7, l'algorithme conduit à la matrice suivante :

$$C = \begin{array}{c} \\ \emptyset \\ 1 \\ 1 \to 3 \\ 1 \to 3 \to 2 \end{array} \begin{array}{c} \begin{array}{cccccccc} \emptyset & 1 & 3 & 2 & 1 & 2 & 3 & 2 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0.5 & 0.5 & 0.5 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0 & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0.6 & 0.6 \\ 0 & 0 & 0 & 0.2 & 0.2 & 0.4 & 0.4 & 0.4 \end{pmatrix} \end{array}$$

La valeur de la mesure de fiabilité du motif p dans la séquence s est de 0.4, ce qui correspond à ce que nous avons obtenu dans l'exemple 4.4.1.

De façon plus formelle, la correction de l'algorithme 1 peut être énoncée et démontrée comme suit.

Théorème 4.4.1. *Pour toute séquence s et tout motif β, l'algorithme 1 termine et renvoie $\rho_{pat}(s, \beta)$.*

Démonstration. Soit l'ordre bien fondé $\prec$ défini sur les paires $(i, j) \in \mathbb{N}^2$ par $(i', j') \prec (i, j)$ si et seulement si $i' < i \vee (i' = i \wedge j' < j)$. La démonstration s'effectue par induction bien fondée sur (i, j) pour $\prec$.

Soit une valeur de fiabilité $c[i, j]$ obtenue par l'algorithme 1 pour une paire (i, j).

Hypothèse d'induction : Supposons que pour toute paire (i', j') telle que $(i', j') \prec (i, j)$, les valeurs $c[i', j']$ obtenues soient correctes, c'est-à-dire $c[i', j'] = \rho_{pat}(s_{\to j'}, \beta_{\to i'})$. Nous allons montrer que $c[i, j]$ est également correcte.

Les cas suivants sont présents :

Cas 1. Si $i = 0$, par initialisation, $c[i, j]$ est correcte (Algorithme 1, ligne 2).
Cas 2. Si $j < i$, par initialisation, $c[i, j]$ est correcte (Algorithme 1, ligne 3).
Cas 3. Si $i \geq 1$ et $j \geq i$, dans ce cas, $c[i, j]$ est calculée à la ligne 6. Elle représente la fiabilité du motif $\beta_{\to i} = \beta_1 \to \beta_2 \to \cdots \to \beta_i$ dans la séquence $s_{\to j} = \langle (t_1, \alpha_1, \rho_1), (t_2, \alpha_2, \rho_2), \ldots, (t_j, \alpha_j, \rho_j) \rangle$. En considérant l'ordre de parcours des boucles (lignes 4 et 5) et les initialisations réalisées (lignes 2 et 3), nous savons que les valeurs de $c[i', j']$ ont déjà été calculées pour toute paire (i', j') telle que $(i', j') \prec (i, j)$ et sont donc disponibles. Les deux sous-cas suivants sont alors possibles.

— Si $\alpha_j \neq \beta_i$, il n'existe aucune occurrence de $\beta_{\to i}$ dans $s_{\to j}$ qui comprend l'événement (t_j, α_j, ρ_j). Donc, $\mathcal{O}(s_{\to j}, \beta_{\to i}) = \mathcal{O}(s_{\to j-1}, \beta_{\to i})$, où (pour rappel) la notation $\mathcal{O}(s, \beta)$ représente l'ensemble des occurrences de β dans s.
Par conséquent, d'après la définition 4.17, nous obtenons :

$$\rho_{pat}(s_{\to j}, \beta_{\to i}) = \rho_{pat}(s_{\to j-1}, \beta_{\to i})$$

La valeur retenue par l'algorithme pour $\rho_{pat}(s_{\to j}, \beta_{\to i})$, et rangée dans $c[i, j]$, est alors correcte puisqu'il s'agit de $c[i, j-1]$ (alternative "Sinon" à la ligne 6), c'est-à-dire $\rho_{pat}(s_{\to j-1}, \beta_{\to i})$ par hypothèse d'induction.

— Si $\alpha_j = \beta_i$, l'ensemble des occurrences de $\beta_{\to i}$ dans $s_{\to j}$ est alors l'union de l'ensemble des occurrences de $\beta_{\to i}$ dans $s_{\to j-1}$, et de l'ensemble des occurrences de $\beta_{\to i-1}$ dans $s_{\to j-1}$, étendue chacune par la date d'occurrence de β_i, c'est-à-dire t_j. Nous noterons ceci de la façon suivante :

$$\mathcal{O}(s_{\to j}, \beta_{\to i}) = \mathcal{O}(s_{\to j-1}, \beta_{\to i}) \cup \text{extend}\left(\mathcal{O}\left(s_{\to j-1}, \beta_{\to i-1}\right), t_j\right)$$

avec $\text{extend}\left(\mathcal{O}\left(s_{\to j-1}, \beta_{\to i-1}\right), t_j\right) = \{\langle u_1, u_2, \ldots, u_{i-1}, t_j \rangle \mid \langle u_1, u_2, \ldots, u_{i-1} \rangle \in \mathcal{O}(s_{\to j-1}, \beta_{\to i-1})\}$

La mesure de fiabilité est alors exprimée comme suit :

$$\begin{aligned}\rho_{pat}(s_{\to j}, \beta_{\to i}) &= \max\{\rho_{occ}(s_{\to j}, o) \mid o \in \mathcal{O}(s_{\to j}, \beta_{\to i})\} \\ &= \max\{\rho_{occ}(s_{\to j}, o) \mid o \in \mathcal{O}(s_{\to j-1}, \beta_{\to i}) \\ &\cup \text{extend}\left(\mathcal{O}\left(s_{\to j-1}, \beta_{\to i-1}\right), t_j\right)\} \\ &= \max\{\max\{\rho_{occ}(s_{\to j}, o) \mid o \in \mathcal{O}(s_{\to j-1}, \beta_{\to i})\}, \\ &\max\{\rho_{occ}(s_{\to j}, o) \mid o \in \text{extend}\left(\mathcal{O}\left(s_{\to j-1}, \beta_{\to i-1}\right), t_j\right)\}\}\end{aligned}$$

Or :

$$\max\{\rho_{occ}(s_{\to j}, o) \mid o \in \mathcal{O}(s_{\to j-1}, \beta_{\to i})\} = \rho_{pat}(s_{\to j-1}, \beta_{\to i})$$

Et :

$$\max\{\rho_{occ}(s_{\to j}, o) \mid o \in \text{extend}\left(\mathcal{O}\left(s_{\to j-1}, \beta_{\to i-1}\right), t_j\right)\} = \min\left\{c[i-1, j-1], \rho_j\right\}$$

Donc :

$$\rho_{pat}(s_{\to j}, \beta_{\to i}) = \max\left\{\rho_{pat}(s_{\to j-1}, \beta_{\to i}), \min\left\{c[i-1, j-1], \rho_j\right\}\right\}$$

La valeur calculée pour $c[i, j]$ (ligne 6) est $\max\left(\min(c[i-1, j-1], \rho_j), c[i, j-1]\right)$. Par hypothèse d'induction, cette valeur est donc bien égale à $\rho_{pat}(s_{\to j}, \beta_{\to i})$. Ceci termine le dernier cas.

Par induction, pour toute paire (i, j), nous avons $c[i, j] = \rho_{pat}(s_{\to j}, \beta_{\to i})$. La terminaison de l'algorithme est triviale car toutes les répétitions sont des structures d'itérations bornées. □

L'algorithme 1 possède une complexité de $\mathcal{O}(n \times m)$ où n est la longueur de la séquence et m est la longueur du motif. Nous pouvons donc calculer la mesure de fiabilité d'un motif dans une séquence de façon plus efficace qu'une approche naïve qui localiserait chaque occurrence, et dont le coût serait proportionnel au nombre maximum d'occurrences dans la séquence.

En examinant la construction de la matrice C, nous pouvons observer que pour obtenir le résultat final $c[m, n]$, il n'est pas nécessaire de calculer les éléments $c[i, j]$ où $j > n+i-m$. Les lignes 2 et 5 de l'algorithme 1 peuvent donc être remplacées respectivement par les suivantes :

$c[0, j] \leftarrow 1,$ pout tout $j \in \mathbb{N}$ tel que $0 \leq j \leq n - m$;

for $j \leftarrow i$ **to** $n + i - m$ **do**

Avec cette optimisation, l'état final de la matrice C (pour $i = 3$) dans l'exemple 4.4.2 sera :

$$C = \begin{array}{c} \\ \emptyset \\ 1 \\ 1 \to 3 \\ 1 \to 3 \to 2 \end{array} \begin{array}{c} \begin{array}{cccccccc} \emptyset & 1 & 3 & 2 & 1 & 2 & 3 & 2 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & - & - & - \\ 0 & 0.5 & 0.5 & 0.5 & 0.6 & 0.6 & - & - \\ 0 & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0.6 & - \\ 0 & 0 & 0 & 0.2 & 0.2 & 0.4 & 0.4 & 0.4 \end{pmatrix} \end{array}$$

4.4.3 Prise en compte de la contrainte sur la mesure de fiabilité

Il est important de prendre en compte les propriétés de la contrainte utilisée dans le but de mieux l'intégrer dans le processus d'extraction. Par exemple, une contrainte anti-monotone permet de réduire de façon considérable l'espace de recherche, et amène donc à une extraction très efficace. Pour la contrainte de fiabilité, nous avons le lemme suivant :

Lemme 4.4.2. *La contrainte de fiabilité C_ρ ne fait pas partie des contraintes suivantes : anti-monotones, monotones, succinctes.*

Démonstration. Nous allons montrer par un contre-exemple que C_ρ n'est ni anti-monotone, ni monotone.

Soit les séquences s_1 et s_2 suivantes :

$s_1 = \langle(1, \mathbf{1}, 0.8), (2, \mathbf{2}, 0.8), (3, \mathbf{3}, 0.8), (4, \mathbf{1}, 0.8), (5, \mathbf{2}, 0.2)\rangle$

$s_2 = \langle(1, \mathbf{1}, 0.2), (2, \mathbf{2}, 0.2), (3, \mathbf{3}, 0.2)\rangle$

Soit la contrainte C_ρ pour le seuil de fiabilité de 0.7. Considérons le motif $1 \to 2 \to 3 \to 1$. Il apparaît seulement dans s_1 et nous avons $\rho_{pat}(1 \to 2 \to 3 \to 1, s_1) = 0.8$. D'où $\rho(1 \to 2 \to 3 \to 1) = \frac{0.8}{1} = 0.8$, et $C_\rho(1 \to 2 \to 3 \to 1)$ est donc satisfaite. Soit $1 \to 2 \to 3$ un sous-motif de $1 \to 2 \to 3 \to 1$. Il apparaît dans s_1 et s_2, avec $\rho_{pat}(1 \to 2 \to 3, s_1) = 0.8$ et $\rho_{pat}(1 \to 2 \to 3, s_2) = 0.2$. Nous avons alors $\rho(1 \to 2 \to 3) = \frac{0.8+0.2}{2} = 0.5$ et $C_\rho(1 \to 2 \to 3)$ n'est pas satisfaite. La contrainte C_ρ n'est donc pas anti-monotone.

Soit le motif $1 \to 2 \to 3 \to 1 \to 2$. Il apparaît seulement dans s_1 et $\rho_{pat}(1 \to 2 \to 3 \to 1 \to 2, s_1) = 0.2$. Donc $\rho(1 \to 2 \to 3 \to 1 \to 2) = \frac{0.2}{1} = 0.2$ et $C_\rho(1 \to 2 \to 3 \to 1 \to 2)$ n'est pas satisfaite. Comme $1 \to 2 \to 3 \to 1$ est un sous-motif de $1 \to 2 \to 3 \to 1 \to 2$ satisfaisant la contrainte, alors C_ρ n'est pas monotone.

Enfin, la spécification de cette contrainte C_ρ ne permet pas de générer directement les motifs la satisfaisant. En effet, pour vérifier sa satisfaction, il faut connaître les mesures de fiabilité des motifs dans la base de séquences. Cette contrainte n'est donc pas succincte. □

L'absence de ces propriétés empêche une intégration efficace de la contrainte puisqu'elle ne permet pas de réduire l'espace de recherche d'une manière directe. Par contre, en se basant sur la mesure de fiabilité, nous pouvons définir une contrainte "relaxée" qui soit, elle, anti-monotone, permettant d'élaguer de façon partielle l'espace de recherche.

Définition 4.20 (Mesure de fiabilité relative au support minimal). Pour chaque motif β, la mesure de fiabilité relative au support minimal est calculée comme suit :

$$\widetilde{\rho}(\beta) = \frac{\sum_{seq(x,y) \in cover(\beta)} \rho_{pat}(seq(x, y), \beta)}{\sigma}$$

où σ est le seuil support minimal.

Lemme 4.4.3. *Pour un motif séquentiel fréquent β, nous avons la relation suivante : $\rho(\beta) \leq \widetilde{\rho}(\beta)$*

Démonstration. β étant fréquent, nous avons donc : $support(\beta) \geq \sigma$. D'après les définitions 4.18 et 4.20, nous avons donc $\rho(\beta) \leq \widetilde{\rho}(\beta)$. □

Définition 4.21 (Contrainte sur la mesure de fiabilité relative au support minimal). La contrainte sur la mesure de fiabilité relative au support minimal est une contrainte qui précise que les motifs doivent avoir une valeur de $\widetilde{\rho}(\beta)$ supérieure ou égale au seuil de fiabilité :

$$C_{\widetilde{\rho}}(\beta) \equiv \widetilde{\rho}(\beta) \geq \gamma$$

Pour démontrer l'anti-monotonie de cette contrainte, nous allons tout d'abord établir le lemme suivant.

Lemme 4.4.4. *Soit β^* et β^{**} deux motifs séquentiels tels que β^* est un sous-motif de β^{**} ($\beta^* \prec \beta^{**}$) et $seq(x,y)$ une séquence d'événements, nous avons la propriété suivante :*

$$\forall (x,y), \forall (\beta^*, \beta^{**}), \beta^* \prec \beta^{**} \Rightarrow \rho_{pat}(seq(x,y), \beta^*) \geq \rho_{pat}(seq(x,y), \beta^{**})$$

Démonstration. Nous avons les deux cas de figure suivants :

— Si $seq(x,y)$ ne contient aucune occurrence de β^{**}, $\rho_{pat}(seq(x,y), \beta^{**})$ vaut 0, par conséquent, nous obtenons :

$$\rho_{pat}(seq(x,y), \beta^*) \geq 0 = \rho_{pat}(seq(x,y), \beta^{**})$$

— Si $seq(x,y)$ contient au moins une occurrence de β^{**}, comme β^* est un sous-motif de β^{**}, on a :

$$\begin{aligned}\mathcal{O}(seq(x,y), \beta^*) =& \{o^* \in \mathcal{O}(seq(x,y), \beta^*) \mid \exists o^{**} \in \mathcal{O}(seq(x,y), \beta^{**}) : o^* \prec o^{**}\} \\ & \cup \{o^* \in \mathcal{O}(seq(x,y), \beta^*) \mid \nexists o^{**} \in \mathcal{O}(seq(x,y), \beta^{**}) : o^* \prec o^{**}\}\end{aligned}$$

Le premier ensemble ($IN = \{o^* \in \mathcal{O}(seq(x,y), \beta^*) | \exists o^{**} \in \mathcal{O}(seq(x,y), \beta^{**}) : o^* \prec o^{**}\}$) contient les occurrences de β^* dans $seq(x,y)$ qui sont incluses dans au moins une occurrence de β^{**} dans la même séquence. Le deuxième ensemble ($OUT = \{o^* \in \mathcal{O}(seq(x,y), \beta^*) | \nexists o^{**} \in \mathcal{O}(seq(x,y), \beta^{**}) : o^* \prec o^{**}\}$) contient les occurrences de β^* qui ne sont incluses dans aucune occurrence de β^{**} dans cette séquence.
La fiabilité du motif β^* est exprimée comme suit :

$$\begin{aligned}\rho_{pat}(seq(x,y), \beta^*) &= \max_{o \in \mathcal{O}(seq(x,y), \beta^*)} (\rho_{occ}(seq(x,y), o)) \\ &= \max_{o \in IN \cup OUT} (\rho_{occ}(seq(x,y), o)) \\ &= \max \left(\max_{o \in IN} (\rho_{occ}(seq(x,y), o)), \max_{o \in OUT} (\rho_{occ}(seq(x,y), o)) \right) \\ &\geq \max_{o \in IN} (\rho_{occ}(seq(x,y), o))\end{aligned}$$

Il est à noter que pour chaque occurrence o^{**} de β^{**}, l'ensemble IN contient toutes les occurrences de β^* incluses dans o^{**}. Par conséquent :

$$\max_{o \in IN} (\rho_{occ}(seq(x,y), o)) \geq \max_{o \in \mathcal{O}(seq(x,y), \beta^{**})} (\rho_{occ}(seq(x,y), o))$$

Or :

$$\max_{o \in \mathcal{O}(seq(x,y), \beta^{**})} (\rho_{occ}(seq(x,y), o)) = \rho_{pat}(seq(x,y), \beta^{**})$$

Nous obtenons donc la relation suivante :

$$\rho_{pat}(seq(x,y), \beta^*) \geq \rho_{pat}(seq(x,y), \beta^{**})$$

□

Théorème 4.4.5. *La contrainte sur la mesure de fiabilité relative au support minimal $C_{\widetilde{\rho}}$ est anti-monotone.*

Démonstration. La propriété d'anti-monotonie de cette contrainte peut être démontrée comme suit :

Toute séquence couverte par un motif séquentiel est aussi couverte par ses sous-motifs, autrement dit :

$$\forall(\beta^*, \beta^{**}), \beta^* \prec \beta^{**} \Rightarrow cover(\beta^*) \supseteq cover(\beta^{**}) \tag{4.1}$$

En combinant cette inclusion avec le lemme 4.4.4, nous avons :

$$\forall(\beta^*, \beta^{**}), \beta^* \prec \beta^{**} \Rightarrow \sum_{s \in cover(\beta^*)} \rho_{pat}(seq(x,y), \beta^*) \geq \sum_{s \in cover(\beta^{**})} \rho_{pat}(seq(x,y), \beta^{**}) \tag{4.2}$$

D'où :

$$\forall(\beta^*, \beta^{**}), \beta^* \prec \beta^{**} \Rightarrow \widetilde{\rho}(\beta^*) \geq \widetilde{\rho}(\beta^{**}) \tag{4.3}$$

La contrainte $C_{\widetilde{\rho}}$ est donc anti-monotone. □

Grâce à la propriété d'anti-monotonie, cette contrainte peut être prise en compte de façon efficace en termes de l'espace de recherche pendant le processus d'extraction. En plus, le lemme 4.4.3 montre, pour les motifs fréquents, qu'avec le même seuil de fiabilité γ, la contrainte $C_{\widetilde{\rho}}$ est moins stricte que la contrainte C_ρ. Par conséquent, nous pouvons utiliser de façon conjointe ces deux contraintes pour avoir une extraction efficace : la contrainte $C_{\widetilde{\rho}}$ pour réduire l'espace de recherche et la contrainte C_ρ pour sélectionner les motifs fiables. En effet, si un motif fréquent β ne satisfait pas $C_{\widetilde{\rho}}$, tous ses sur-motifs fréquents ne la satisfont pas non plus. Comme $C_{\widetilde{\rho}}$ est moins stricte que C_ρ pour les motifs fréquents, β et ses sur-motifs ne satisfont alors pas non plus C_ρ et peuvent être élagués de l'espace de recherche. Par contre, dans le cas où β satisfait $C_{\widetilde{\rho}}$, l'espace de recherche reste inchangé et une vérification vis-à-vis de la contrainte C_ρ doit être effectuée pour savoir si β est fiable ou pas. Cette technique, qui consiste à se servir d'une contrainte relaxée pour avoir une extraction efficace en réduisant l'espace de recherche, est appelée communément technique de *push partiel*.

4.5 Expériences

La contrainte de fiabilité permet de sélectionner seulement les motifs qui apparaissent sur les données de déplacement qui disposent en moyenne d'un niveau de confiance suffisant. L'algorithme basé sur la programmation dynamique permet un calcul rapide de la fiabilité de chaque motif, et l'utilisation de la technique de *push* partiel basée sur la contrainte sur la mesure de fiabilité relative au support minimal peut amener à une extraction plus efficace en réduisant l'espace de recherche. Dans cette section, nous allons détailler les intérêts techniques et applicatifs d'utiliser les techniques proposées dans les sections précédentes via des expériences sur deux séries de champs de déplacements de glacier : une sur le Groenland et une sur le massif du Mont-Blanc. La première, provenant des images optiques des satellites

Landsat, contient des déplacements annuels tandis que la deuxième, provenant des images SAR du satellite TerraSAR-X, contient des déplacements de courtes durées.

4.5.1 STCD sur le Groenland provenant de données optiques

4.5.1.1 Préparation des données

Nous utilisons dans cette expérience la STCD préparée et utilisée dans Tedstone *et al.* (2015). Elle est construite à partir de séries d'images des satellites Landsat (5, 7, et 8), acquises entre avril et octobre [2], durant trois décennies : de 1985 jusqu'en 2014. Au total, 475 paires d'images avec des écarts temporels entre 352 et 400 jours [3] sont formées. Afin d'améliorer la qualité de l'estimation du déplacement, Tedstone *et al.* ont utilisé la première composante de l'Analyse en Composantes Principales (ACP) appliquée sur les bandes spectrales 2 et 3, qui sont identifiées comme étant des bandes spectrales optimales. Un filtre passe-haut (filtre de Sobel) est ensuite appliqué pour améliorer les caractéristiques de surface telles que les crevasses. Puis, la méthode d'*offset-tracking* est appliquée sur les mesures de gradient afin de produire les estimations de déplacement. Quant au paramétrage, la taille des fenêtres de corrélation est fixée à 44×44 pixels (1320×1320 m), et l'écart entre les fenêtres de corrélation à 8 pixels (240 m). Une fois les 475 champs de déplacements obtenus, les estimations de faible qualité sont écartées grâce à la mesure du Rapport Signal sur Bruit (en anglais *Signal to Noise Ratio*) (SNR) qui est disponible à l'issue de l'offset-tracking. Finalement, pour la période de 2000 à 2014, les champs de déplacements sont regroupés en périodes d'une année, et pour la période de 1985 à 2000, les champs sont regroupés en périodes de deux années puisqu'il y avait moins de paires d'images disponibles. Pour obtenir l'estimation de déplacement final pour chaque période, la valeur médiane de toutes les estimations disponibles pour chaque pixel est utilisée afin de réduire l'impact des valeurs aberrantes. À la sortie de ce processus, 20 champs de déplacements (dont 7 de 1985 à 2000 et 13 de 2000 à 2014), avec un pas de 240×240 m, sont disponibles. Les déplacements sont exprimés en mètre par an.

Concernant l'indice de confiance, c'est la mesure de cohérence des vecteurs de déplacements (en anglais *Velocity Vector Coherence*) (VVC), introduite dans Dehecq *et al.* (2015) qui est utilisée. Elle est définie comme suit :

$$\rho(x,y,t) = \begin{cases} \frac{\left\| \sum_{\vec{v} \in \mathcal{V}(x,y,t)} \vec{v} \right\|}{\sum_{\vec{v} \in \mathcal{V}(x,y,t)} \|\vec{v}\|}, & \text{si } \sum_{\vec{v} \in \mathcal{V}(x,y,t)} \|\vec{v}\| \neq 0 \\ 1, & \text{sinon} \end{cases} \tag{4.4}$$

où $\mathcal{V}(x,y,t)$ représente l'ensemble des vecteurs de déplacement utilisés pour obtenir le déplacement médian pour le pixel (x,y) à la période t, et $\rho(x,y,t)$ indique l'indice de confiance de l'événement à la date t dans la séquence $seq(x,y)$.

La valeur de cette mesure est comprise dans l'intervalle $[0,1]$. Elle est égale à 1 si toutes les estimations de déplacement possèdent la même direction, et elle tend vers 0 si les directions suivent une distribution uniforme [4].

2. Les déplacements glaciaires peuvent être plus facilement observés durant ces périodes de l'année par rapport aux autres, car il y a moins de neige fraîchement tombée.

3. Dans cette étude, l'objet principal de la recherche s'appuie sur les tendances à long terme (inter-annuelles) du déplacement des glaciers. Ce choix permet ainsi de réduire l'impact de la variabilité saisonnière des glaciers sur ces tendances.

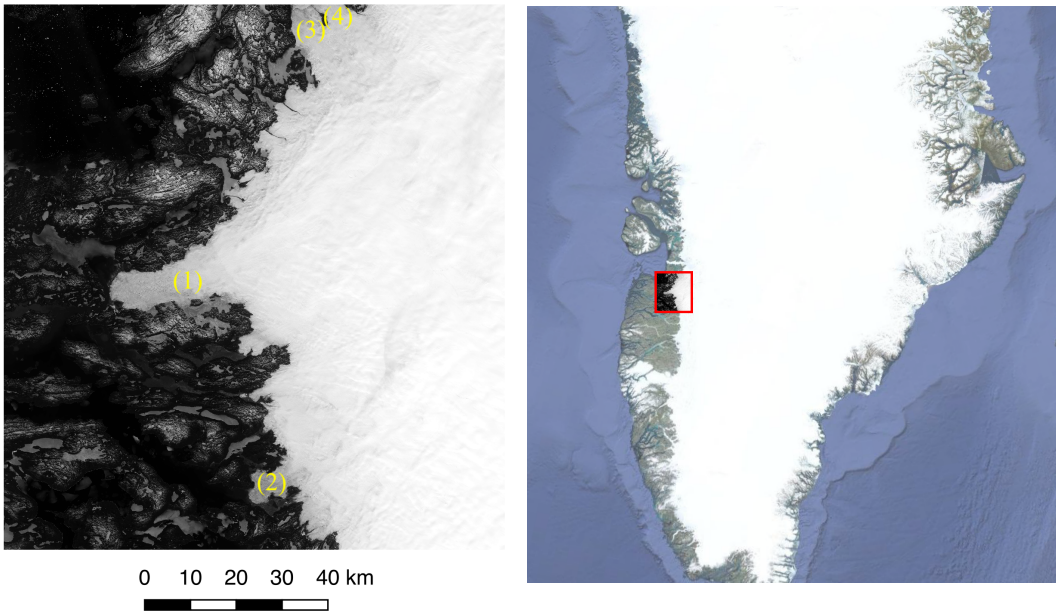
4. D'une façon plus précise, la mesure VVC dépend également de l'amplitude des estimations. Autrement dit, plus l'amplitude est grande, plus l'estimation aura de l'impact sur la valeur de VVC.

C'est à partir de cette étape que nous avons récupéré les données et dans cette étude, nous nous concentrons sur la variation de vitesse. Après avoir retenu des images de taille 458×500 pixels couvrant une région contenant les principaux déplacements d'intérêt (Figure 4.2), nous calculons la valeur de vitesse médiane différentielle (en anglais *Median Differential Speed*) (MDS) à partir des vecteurs de déplacement comme suit :

$$mds(x, y, t) = \frac{\|\vec{v}(x, y, t)\| - s_{med}(x, y)}{s_{med}(x, y)} \tag{4.5}$$

où $\vec{v}(x, y, t)$ est le vecteur de vitesse estimé pour le pixel (x, y) à la date t et $s_{med}(x, y)$ représente la médiane des valeurs $\|\vec{v}(x, y, t')\|$ pour t' décrivant l'ensemble des dates d'estimation.

Dans cette expression, les mesures de vitesse sont centrées et normalisées [5] par rapport à la médiane observée à la position associée. De cette manière, nous obtenons une mesure plus robuste aux valeurs aberrantes que si nous utilisions la moyenne au lieu de la médiane.



(a) Première composante de l'ACP d'une image Landsat 8 (27/04/2013), avec quatre des glaciers principaux sur la zone d'étude : (1) glacier Nordenskjöld, (2) glacier Polonia, (3) Sarqardliup Sermia et (4) Alangordliup Sermia (nommés d'après Rosenau *et al.* (2015) et NunaGIS (2018))

(b) Position de la zone d'étude (rectangle rouge), avec la carte de Google Earth en arrière-plan

FIGURE 4.2 – Images de la zone d'étude et position des glaciers

5. Il est à noter que le résultat de cette mesure n'est pas borné. Dans le cas particulier où $s_{med}(x, y) = 0$, i.e., toutes les vitesses sont nulles, pour une raison pratique, la valeur de MDS est fixée à 0.

4.5.1.2 Paramètres utilisés pour l'extraction des motifs fiables

La STCD symbolique est construite en encodant les mesures $mds(x, y, t)$ avec trois symboles (1, 2 et 3) et une stratégie de répartition en fréquence égale. Plus concrètement, nous parcourons tous les points de mesures pour obtenir une répartition détaillée des valeurs de données. C'est à partir de cette répartition que les seuils sont calculés de façon à ce qu'un même nombre de symboles (1 ou 2 ou 3) soit obtenu sur toute la base de données. Numériquement, pour ce jeu de données, le symbole 1 correspond aux valeurs allant de -0.999 (valeur minimale de MDS) à ≈ -0.055, le symbole 2 de ≈ -0.055 à ≈ 0.059, et le symbole 3 de ≈ 0.059 à 108.121 (valeur maximale de MDS). Nous pouvons observer que la répartition des valeurs est très déséquilibrée. L'histogramme illustré sur la figure 4.3 montre clairement que la plupart des valeurs de MDS se trouvent aux alentours de 0.

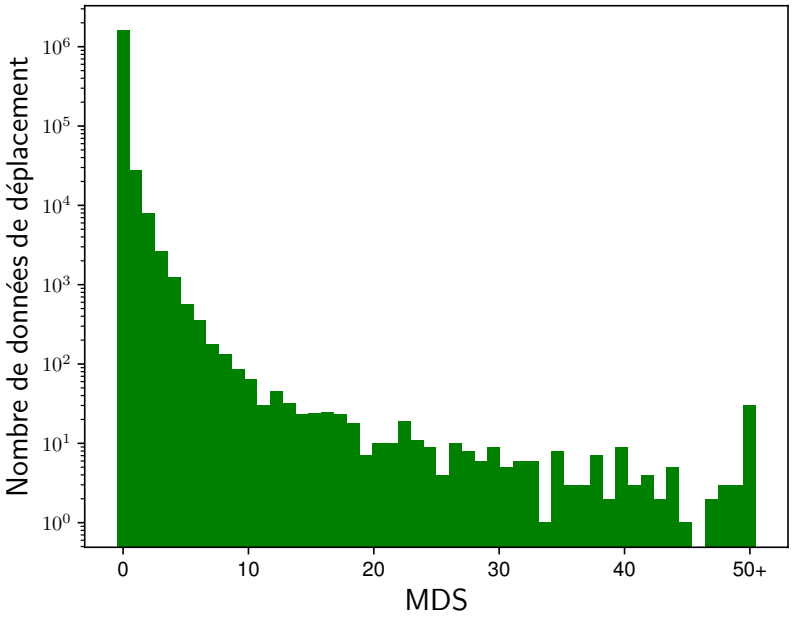


FIGURE 4.3 – Histogramme des mesures MDS (Groenland)

Pour chaque position (x, y), cet encodage produit une séquence $seq(x, y) = \langle (t_1, \alpha_1, \rho_1), \ldots, (t_i, \alpha_i, \rho_i), \ldots, (t_n, \alpha_n, \rho_n) \rangle$ avec t_i une date d'estimation et $\alpha_i \in \{1, 2, 3\}$, où les symboles $1, 2, 3$ représentent respectivement des valeurs de vitesse faible, proche de la valeur médiane, et élevée. Les valeurs de ρ_i sont simplement les valeurs de confiance associées qui sont définies par l'équation 4.4, $\rho_i = \rho(x, y, t_i)$.

La plus grande partie de la région choisie ne correspond pas qu'à de fortes confiances. Pour la plupart des positions, l'écart entre la plus forte et la plus faible confiance est très important. Ceci est visible sur les figures 4.4a, 4.4b et 4.4c, qui donnent, pour chaque position (x, y), la valeur minimale, maximale et moyenne des indices de confiance en (x, y) tout au long de la série.

À partir de la STCD symbolique, les motifs SFG sont extraits, et seulement les motifs maximaux sont retenus, afin de se concentrer uniquement sur les évolutions les plus spécifiques.

En ce qui concerne le réglage des paramètres, il est réalisé comme suit. D'abord, le seuil de

connexité est fixé à $\kappa = 5$. C'est un réglage standard permettant de révéler des phénomènes géophysiques (Julea *et al.*, 2011; Pericault *et al.*, 2015). Cela garantit qu'une évolution retenue satisfait la condition suivante : si l'évolution est présente à la position (x, y), elle est aussi présente, en moyenne, sur au moins 5 des 8 positions voisines de (x, y). Par conséquent, les évolutions apparaissant plutôt de façon isolée spatialement sont ainsi rejetées. Le deuxième paramètre est σ, le seuil de support. Il est réglé de façon à retenir le plus grand nombre de motifs SFG maximaux. Cette stratégie permet ainsi de garder la description [6] la plus riche de la base de séquences. La figure 4.5 présente le nombre de motifs maximaux en fonction du seuil de support, exprimé en pourcentage par rapport aux 458×500 pixels de la zone. D'après cette courbe, le seuil σ est ainsi fixé à 7.5%.

En ce qui concerne le seuil de fiabilité γ, plus il est élevé (i.e., proche de 1), plus les motifs extraits sont fiables, mais moins nombreux seront les motifs retenus. Par conséquent, la valeur de γ doit être choisie en vue d'obtenir un compromis entre le nombre de motifs extraits et leur qualité, exprimée ici par la mesure de fiabilité. Afin de trouver un tel compromis, une fonction objectif très simple est définie comme suit :

$$o(\gamma) = \gamma \times p \tag{4.6}$$

où p exprime le nombre de motifs maximaux obtenus pour le seuil de support γ.

Comme montré sur la figure 4.6, ce compromis est obtenu avec $\gamma = 0.85$.

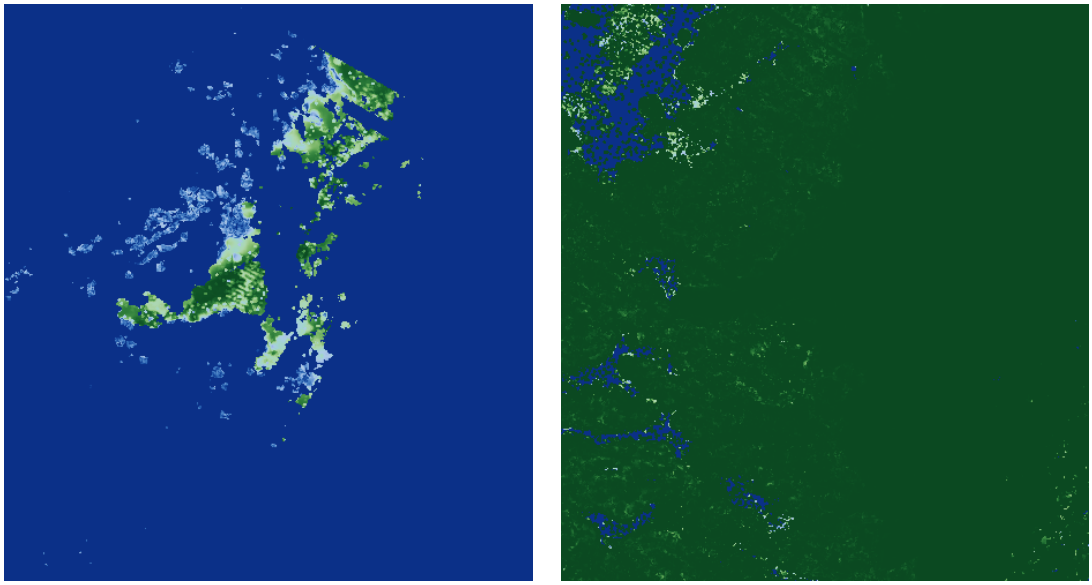
Enfin, bien que les contraintes de support minimal, de connexité et de fiabilité soient appliquées et que seulement les motifs maximaux soient sélectionnés pour se concentrer sur les motifs les plus intéressants, les motifs SFG fiables peuvent être encore nombreux. C'est pour cette raison que nous utilisons la procédure de classement basée sur une technique de *swap* randomisation, introduite dans (Méger *et al.*, 2015) pour identifier les motifs dont la carte de LST est soit difficilement, soit facilement détruite par la *swap* randomisation. Pour cela, les cartes LST obtenues pour la STCD d'origine sont comparées à celles obtenues pour une STCD randomisées en utilisant une mesure d'information mutuelle normalisée. Cette méthode (rappelée dans l'annexe A) produit un classement des motifs, allant des plus attendus aux plus inattendus, par rapport aux distributions des symboles dans la STCD symbolique initiale. Les K motifs les plus attendus et les K motifs les plus inattendus parmi l'ensemble des motifs SFG fiables sont finalement conservés, avec K un paramètre défini par l'utilisateur. Dans notre cas, nous fixons $K = 20$ et nous avons donc un total de 40 motifs obtenus.

4.5.1.3 Effet du push partiel

Comme nous pouvons constater sur la figure 4.7 qui présente le nombre de motifs à évaluer pendant l'extraction en fonction du seuil de fiabilité γ avec et sans l'utilisation de la contrainte relaxée, la réduction de l'espace de recherche intervient uniquement après un seuil élevé ($\gamma > 0.9$), et, elle est très faible. Nous verrons sur le second jeu de données, concernant le massif du Mont-Blanc, que cet impact peut être très différent.

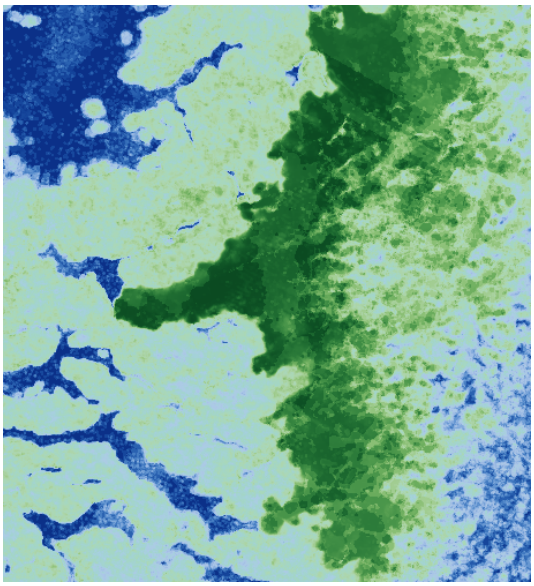
Le réglage du seuil de fiabilité à $\gamma = 0.85$, tel que déterminé dans la section 4.5.1.2, conduit à une collection de 375 motifs SFG fiables qui sont ensuite classés selon la méthode

6. L'ensemble des motifs sélectionnés constitue une description du jeu de données.



(a) Confiance minimale

(b) Confiance maximale



(c) Confiance moyenne

(d) Échelle de couleur

FIGURE 4.4 – Indice de confiance agrégé sur la série du Groenland : (a) minimale, (b) maximale, (c) moyenne. (d) échelle de couleur : de 0 (bleu foncé) à 1 (vert foncé)

de *swap* randomisation proposée par Méger *et al.* (2015). Le processus entier, implémenté en C et Python, nécessite 813 secondes en utilisant un seul cœur d'un Intel Xeon 3,5 GHz sous Linux (Ubuntu), et une consommation maximale de mémoire de 311 Mo.

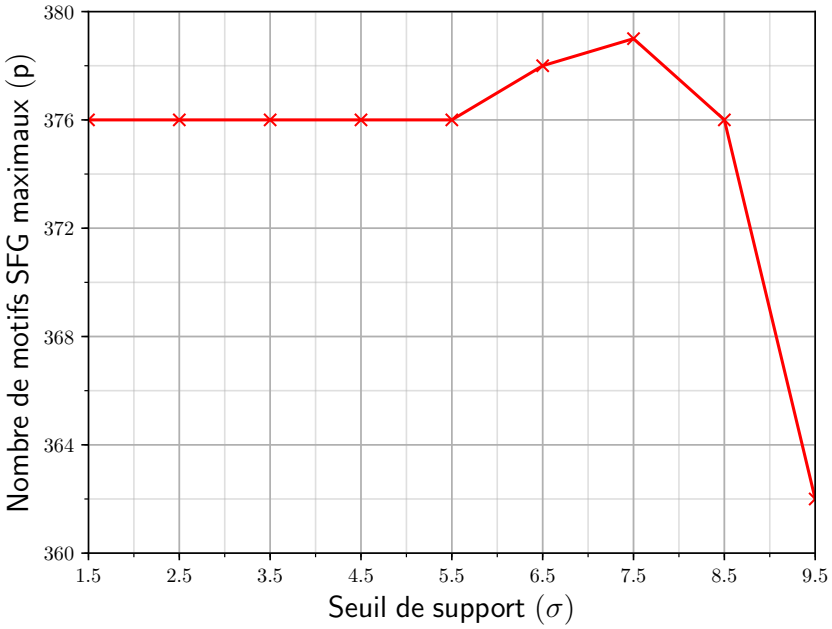


FIGURE 4.5 – Réglage du seuil de support minimal (Groenland)

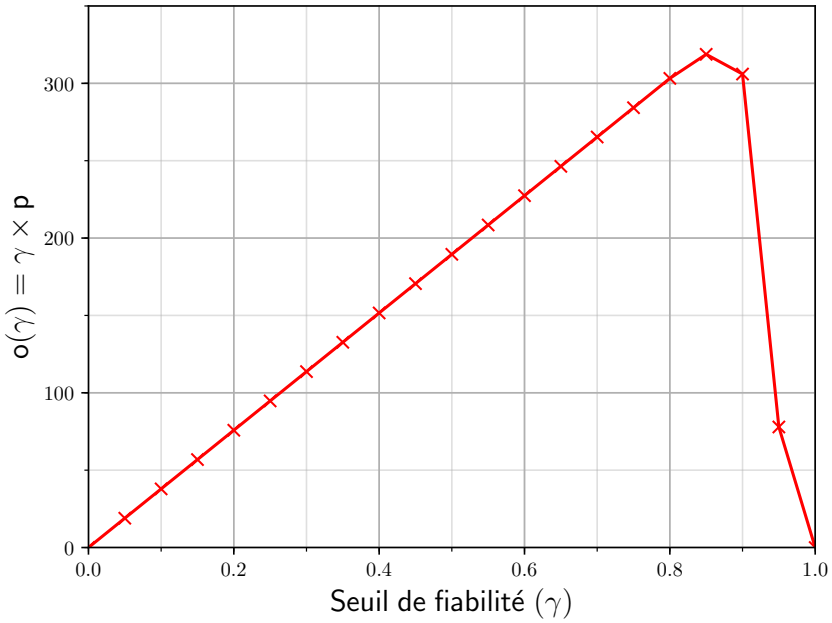


FIGURE 4.6 – Réglage du seuil de fiabilité (Groenland), avec p le nombre de motifs SFG maximaux obtenus pour le seuil γ

4.5.1.4 Résultats qualitatifs

Les 40 cartes LST obtenues représentent différentes évolutions sur diverses régions à différentes périodes dans la série temporelle. Trois cartes parmi les mieux classées sont présentées dans les figures 4.9, 4.11 et 4.13, tracées sur une image optique de la zone en niveau de gris.

Les positions des glaciers principaux sont indiquées avec les étiquettes (1), (2), (3) et (4)

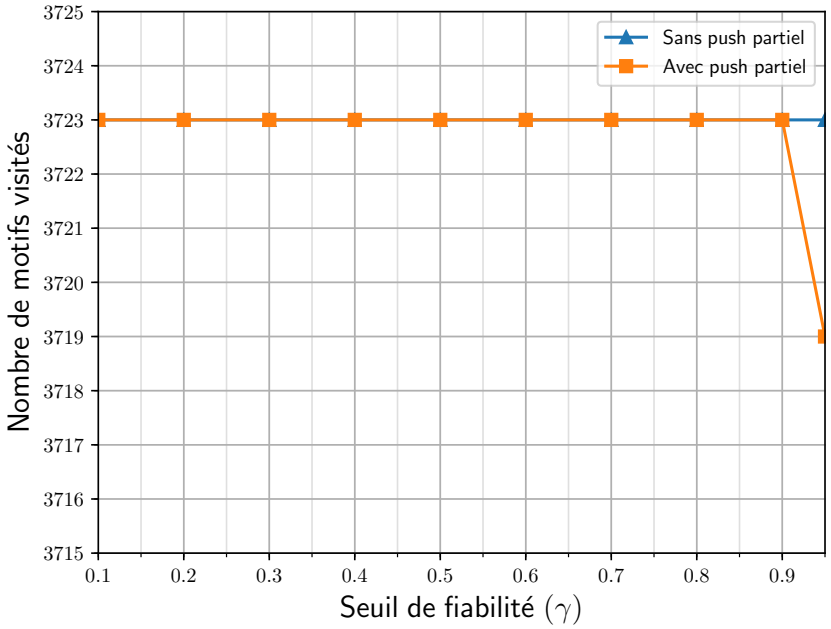


FIGURE 4.7 – Réduction de l'espace de recherche (Groenland)

sur la figure 4.2a.

Les deux premières cartes mettent en avant la décélération régionale signalée dans Tedstone *et al.* (2015). La carte de la figure 4.9 correspond au motif de décélération graduelle suivant : $3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1 \to 1$. Pour rappel, sur une carte LST, les pixels colorés correspondent aux positions spatiales où la séquence est couverte par le motif. Pour chacun de ces pixels, la couleur indique la date à laquelle le dernier symbole de la première occurrence du motif se trouve (suivant une échelle de couleur donnée Figure 4.8). La zone mise en évidence sur cette carte, au milieu du glacier Nordenskjöld (1), correspond à une décélération le long d'un transect longitudinal signalée dans Tedstone *et al.* (2015). De plus, la carte montre que ce motif de décélération est également présent sur la zone au nord, près des glaciers Sarqardliup Sermia (3) et Alangordliup Sermia (4), et au sud, dans la région du glacier Polonia (2). Il est également à noter que d'après l'échelle de couleur, la fin du motif est observée plus tôt au centre-gauche (essentiellement bleu) que sur les régions au sud et au nord (magenta). Un autre type de déplacement, différent de la décélération progressive précédente, est représenté sur des régions complémentaires dans la figure 4.11. Le motif associé, $3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 1$, souligne une longue période stable à une vitesse élevée avant une décélération soudaine. Une recherche détaillée dans les données montre que la séquence des symboles 2 dans le motif $3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1 \to 1$ commence au début des années 90 se termine vers les années 2006 - 2007 (Figure 4.10). Pour le motif $3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 1$, la fin de la séquence des symboles 3 se trouve plus tôt (vers les années 2003 - 2005) (Figure 4.12). Ceci indique qu'une vitesse élevée s'est maintenue dans les zones couvertes par le motif $3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 1$ alors qu'elle avait déjà nettement décru dans les zones couvertes par l'autre motif.

La troisième carte est montrée dans la figure 4.13, associée au motif $1 \to 1 \to 3 \to 1 \to 1 \to 1 \to 1$. Ce motif suggère une vitesse élevée singulière au milieu d'une séquence contenant des vitesses très basses. Il souligne en bleu le glacier Sarqardliup Sermia. Une vérification spécifique dans les données montre que le maximum local au niveau 3 se trouve vers 1997 -

1999 (en vert) dans la partie nord du glacier et vers 1990 - 1992 (en jaune) dans sa partie sud (Figure 4.14).

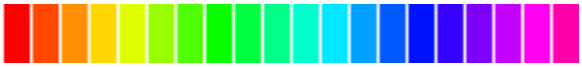


FIGURE 4.8 – Échelle de couleur des cartes LST : depuis 1985 en rouge jusqu'à 2013 en magenta (décomposée en 20 périodes)

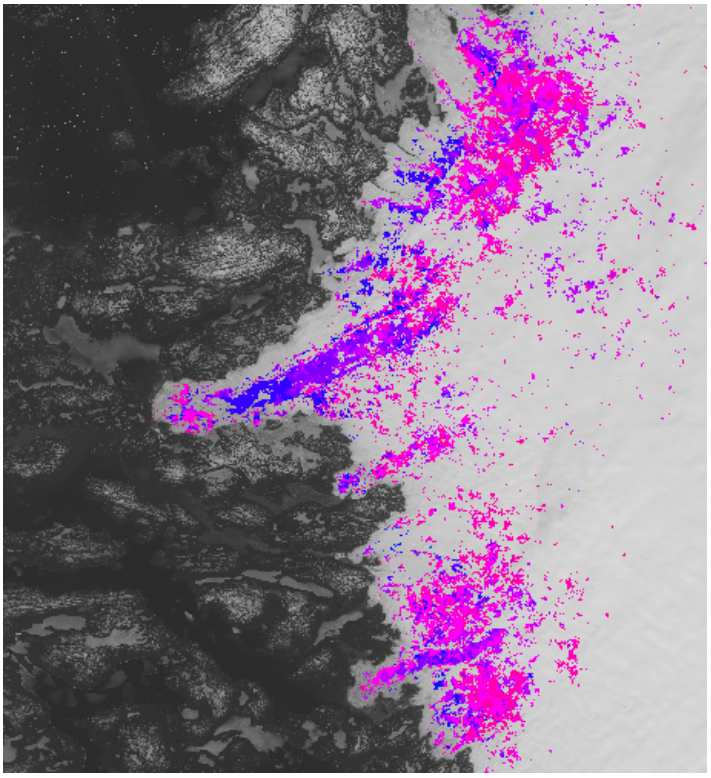


FIGURE 4.9 – Carte LST du motif $3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1 \to 1$

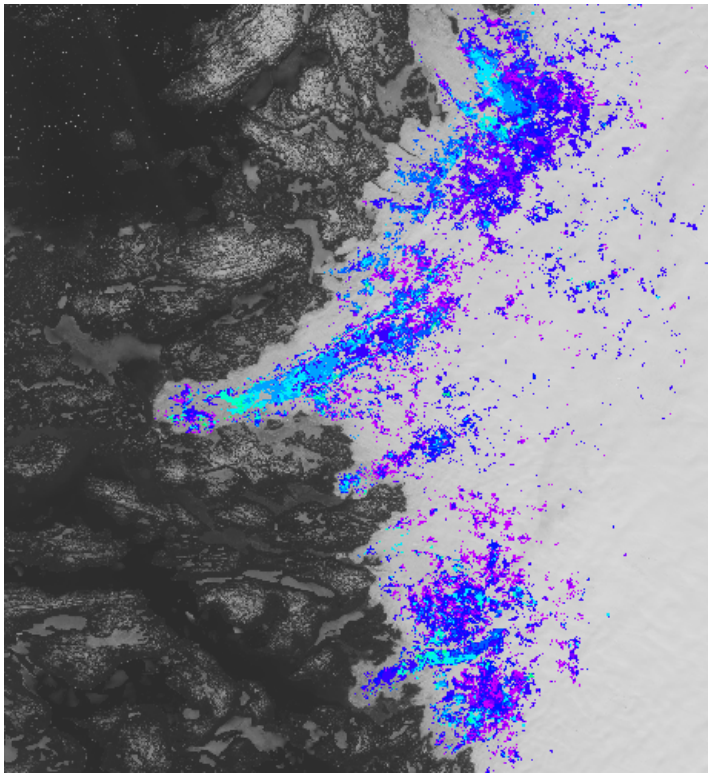


FIGURE 4.10 – Dates d'occurrences du dernier symbole 2 du motif $3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1 \to 1$

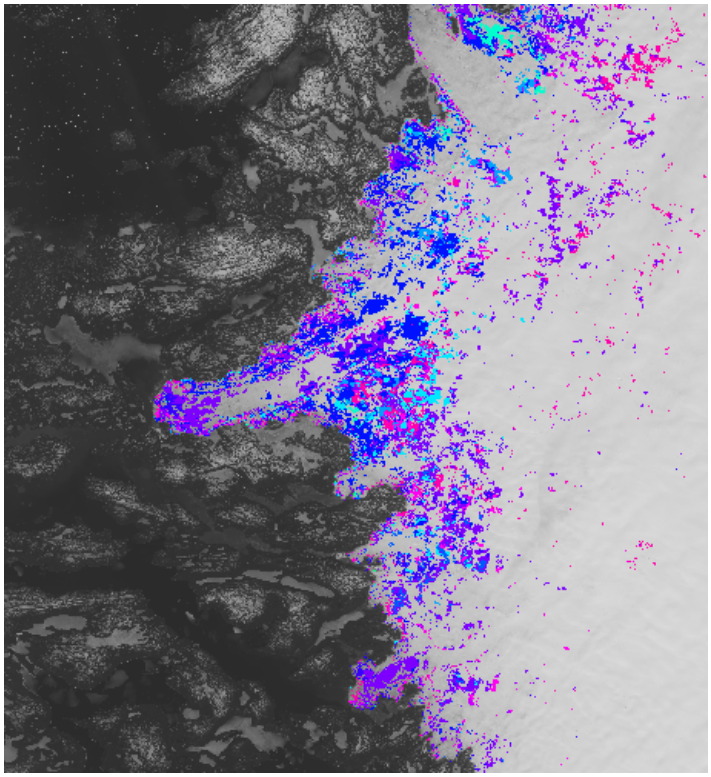


FIGURE 4.11 – Carte LST du motif $3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 1$

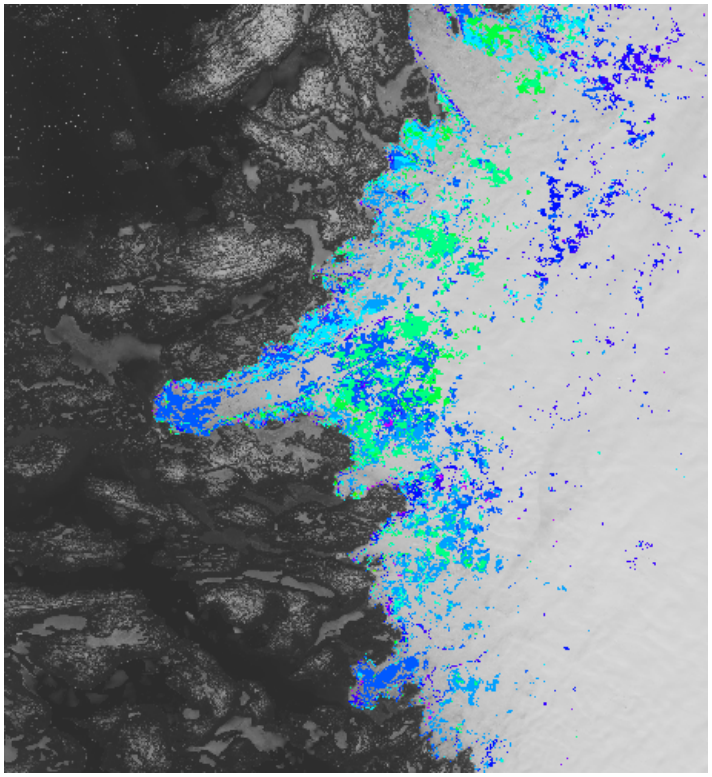


FIGURE 4.12 – Dates d'occurrences du dernier symbole 3 du motif $3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 1$

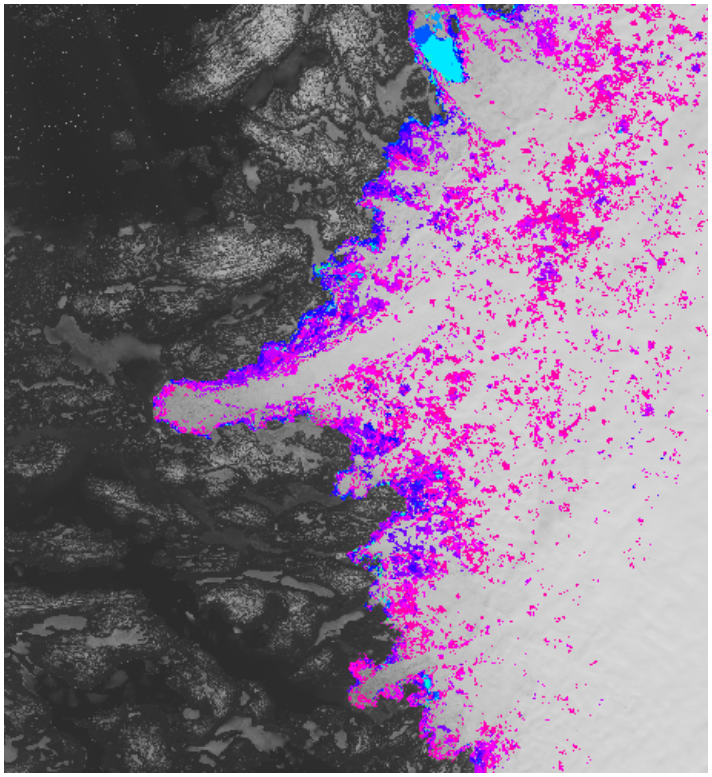


FIGURE 4.13 – Carte LST du motif $1 \to 1 \to 3 \to 1 \to 1 \to 1 \to 1$

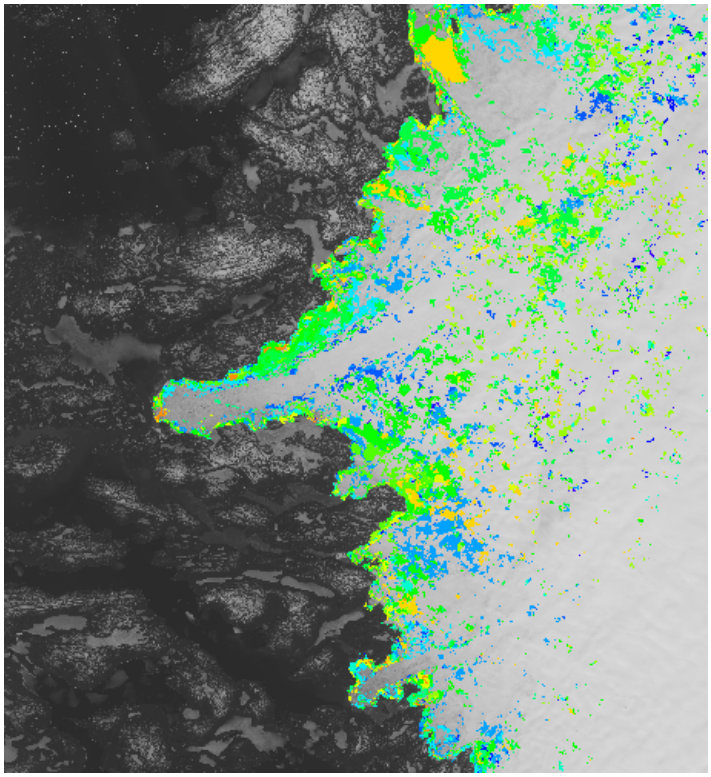


FIGURE 4.14 – Dates d'occurrences du symbole 3 du motif $1 \to 1 \to 3 \to 1 \to 1 \to 1 \to 1$

4.5.2 STCD sur le massif du Mont-Blanc provenant de données SAR

4.5.2.1 Préparation des données

L'approche proposée a également été appliquée pour explorer les évolutions de déplacements de courte durée sur les glaciers alpins dans le massif du Mont-Blanc. Pour cela, nous avons construit une STCD à partir de 26 images du satellite TerraSAR-X obtenues sur des orbites ascendantes, avec une résolution spatiale d'environ 2×2 m. La série a été acquise entre le 31/05/2009 et le 25/09/2011 pour faciliter l'estimation de déplacement, comme sur le Groenland. Elle couvre la partie française du massif du Mont-Blanc et contient ses quatre principaux glaciers : Argentière, Mer de Glace, Bossons et Taconnaz (Figure 4.15). Elle est composée de deux périodes distinctes, une en 2009 du 31 mai au 21 octobre et une autre en 2011 du 5 mai au 25 septembre. Ces deux périodes contiennent 13 images chacune avec une image tous les 11 jours.

Les images constituant la série ont une taille de 10484×9560 pixels. Tout d'abord, nous décrivons comment ces données ont été traitées pour calculer les champs de déplacements à pleine résolution, puis ont été réduites à une taille de 3494×3186 pixels pour obtenir des estimations robustes. Les images SAR contiennent des informations d'amplitude et de phase, qui peuvent toutes les deux être utilisées pour estimer différents types de déplacements. Pour les glaciers alpins, les changements de surface rapides sont susceptibles de réduire la cohérence de phase et les informations d'amplitude sont plus adaptées pour calculer les déplacements des glaciers (Strozzi *et al.*, 2008; Scherler *et al.*, 2008; Fallourd *et al.*, 2011). De plus, au lieu de n'avoir que des déplacements en LOS avec les méthodes basées sur la phase, les méthodes basées sur l'amplitude nous donnent des vecteurs de déplacement exprimés dans les directions

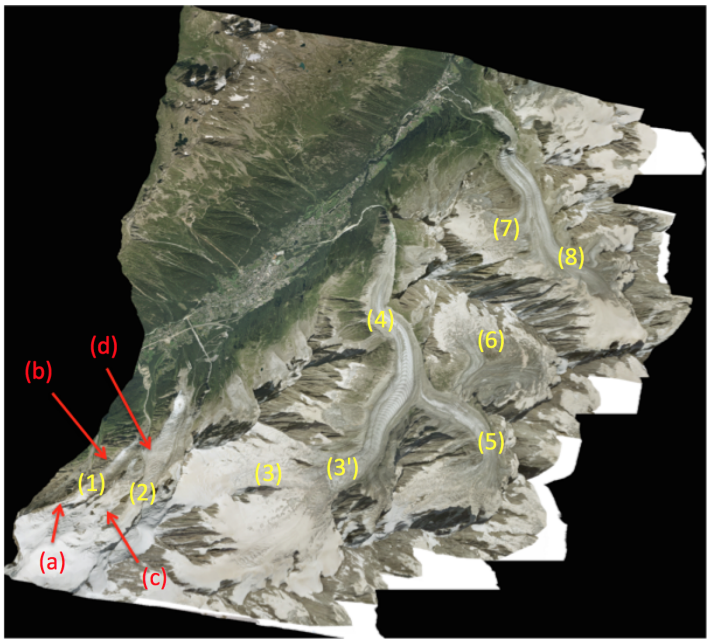


FIGURE 4.15 – Emplacement des glaciers en géométrie radar : (1) Taconnaz, (2) Bossons, (3) glacier du Géant, (3′) séracs du Géant, (4) Mer de Glace, (5) Leschaux, (6) Talèfre, (7) Rognons et (8) Argentière. En rouge : (a) haut du glacier de Taconnaz, (b) 2000 m depuis (a), (c) haut du glacier des Bossons et (d) 2000 m depuis (c).

en *range* et en azimut. C'est pourquoi la méthode de corrélation d'amplitude, mise en œuvre dans les EFIDIR Tools, a été utilisée pour calculer les champs de déplacements. Après avoir appliqué un masque de glacier à partir du *Randolph Glacier Inventory*[7], l'offset tracking est effectuée pour chaque paire d'images consécutives, avec une taille de fenêtre de corrélation de 65 pixels et une taille de fenêtre de recherche de 105 pixels. Un total de 25 champs de déplacements, dont chaque déplacement est un vecteur exprimé en m/jour, est obtenu.

Pour ce jeu de données, notre principal objectif consiste à étudier les variations de vitesse des glaciers. Comme dans Tedstone *et al.* (2015), où une étape de standardisation est appliquée sur les valeurs de vitesse en calculant les mesures MDS, nous proposons une standardisation exploitant la stationnarité locale et l'homogénéité de la STCD. Soit $\vec{v}(x, y, t)$ le vecteur de vitesse observé entre les dates t et $t + 1$ à la position (x, y). La taille de la STCD initiale est réduite en utilisant un filtre passe-bas et un sous-échantillonnage basé sur un pavage spatial. Soit $\Omega_{i,j,t}$ la liste des valeurs de $\|\vec{v}(x, y, t)\|$ contenues dans une fenêtre de pavage $w_{i,j,t}$, et $\Omega_{i,j}$ la concaténation des listes $\Omega_{i,j,1}, \Omega_{i,j,2}, \ldots, \Omega_{i,j,n}$ (i.e., toutes les valeurs dans le temps dans les fenêtres des indices (i, j)). Soit $MAD_{i,j}$ la Déviation Médiane Absolue à l'emplacement (i, j), définie simplement comme suit :

$$MAD_{i,j} = \underset{z \in \Omega_{i,j}}{\text{median}}\{|z - \text{median}(\Omega_{i,j})|\} \tag{4.7}$$

À chaque date t, les mesures décrivant le champ de déplacements sont les vitesses différentielles médianes mdv, et elles sont obtenues pour chaque fenêtre $w_{i,j,t}$ comme suit :

$$mdv(i, j, t) = \frac{\text{median}(\Omega_{i,j,t}) - \text{median}(\Omega_{i,j})}{MAD_{i,j}} \tag{4.8}$$

7. https ://www.glims.org/RGI/

Dans cette expression de mdv, les valeurs sont centrées et normalisées en utilisant la médiane et la MAD au lieu d'une moyenne et d'un écart-type afin de rendre cette normalisation robuste à la présence de valeurs aberrantes. La mesure résultante exprime la différence entre la médiane des valeurs de vitesse sur une fenêtre de pavage $w_{i,j,t}$ par rapport à la médiane sur toutes les fenêtres de la pile au même endroit.

En utilisant des fenêtres de pavage de 3×3 pixels, la taille initiale des champs de vitesse de 10484×9560 pixels est réduite à 3494×3186 pixels, ce qui permet d'avoir des estimations robustes.

Concernant l'indice de confiance, nous proposons d'utiliser la stationnarité de la direction de déplacement du glacier. Par exemple, dans (Scherler *et al.*, 2008; Harant *et al.*, 2011), l'hypothèse d'un déplacement dans la direction de la plus grande pente a été considérée. Cette hypothèse implique que, pour un emplacement donné, la direction du déplacement du glacier à des dates différentes doit être similaire.

Dans notre cas, pour chaque position (x, y) et pour chaque date t, l'indice de confiance ρ est défini comme suit :

$$\rho(x,y,t) = \begin{cases} \cos\left(\widehat{\vec{u}_{x,y,t}, \bar{u}_{x,y}}\right), & \text{si } \cos\left(\widehat{\vec{u}_{x,y,t}, \bar{u}_{x,y}}\right) \geq 0 \\ 0, & \text{sinon} \end{cases} \tag{4.9}$$

où $\vec{u}_{x,y,t}$ est le vecteur de vitesse unitaire à la position (x, y) à la date t et $\bar{u}_{x,y} = \sum_{t=1}^{n} \vec{u}_{x,y,t}$ représente la direction globale de la vitesse sur la séquence située en (x, y). L'opérateur cosinus est utilisé pour mesurer la différence angulaire entre la direction de la vitesse globale et le vecteur vitesse à la date t. Plus l'angle est grand, plus la confiance est petite. Elle est nulle au-delà de 90°.

Afin d'obtenir des estimations robustes, comme expliqué précédemment, la taille originale des acquisitions est réduite en utilisant des fenêtres de pavage $w_{i,j,t}$. Quant à l'indice de confiance, la médiane de la confiance ρ sur la fenêtre $w_{i,j,t}$ est retenue.

4.5.2.2 Paramètres utilisés

La STCD symbolique est ensuite construite à partir des valeurs de vitesses différentielles médianes, en utilisant une quantification avec trois symboles 1, 2 et 3 (indiquant à nouveau les valeurs basse, moyenne et haute) et une stratégie de répartition en fréquence égale. Pour ce jeu de données, le symbole 1 correspond aux valeurs allant de -96.1978 (valeur minimale de MDS) à ≈ -0.6028, le symbole 2 de ≈ -0.6028 à ≈ 0.7821, et le symbole 3 de ≈ 0.7821 à $1.30072e+06$ (valeur maximale de MDS obtenue) [8]. De la même manière que sur la STCD du Groenland, la répartition des valeurs de données est très déséquilibrée. L'histogramme Figure 4.16 montre clairement que la plupart des valeurs de MDS se retrouvent aux alentours de 0. La présence d'autres valeurs plus grandes de MDS reste minime.

Les autres paramètres sont réglés de la même façon que la STCD sur le Groenland :

— Le seuil de connexité κ est fixé à 5.
— Le seuil de support σ est fixé à 4%, représentant une couverture spatiale de 440500

8. Cette valeur maximale est un exemple de valeur aberrante que peuvent avoir les données de déplacement.

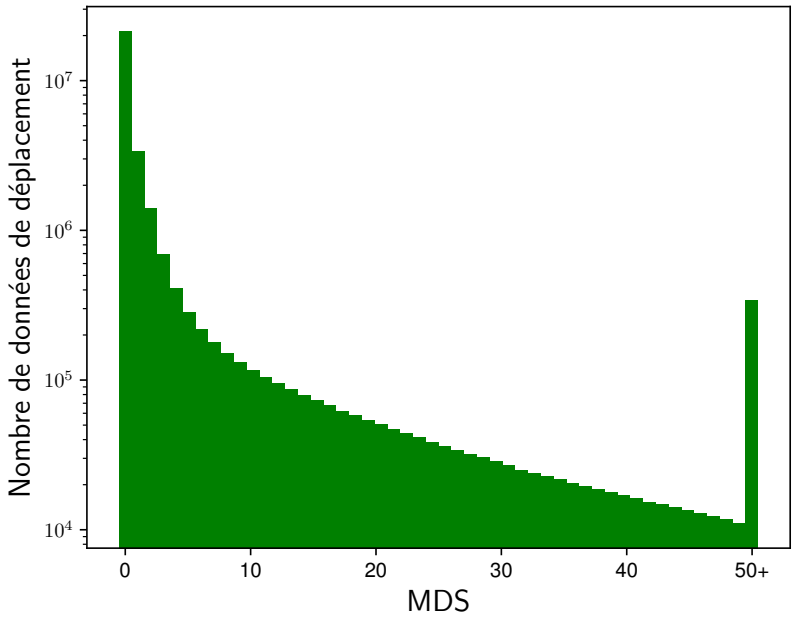


FIGURE 4.16 – Histogramme des mesures MDS (Mont-Blanc)

pixels. Pour rappel, ce seuil est réglé de façon à obtenir le plus grand nombre de motifs SFG maximaux (cf. Figure 4.18).

— Le seuil de fiabilité γ, représentant un compromis entre fiabilité et nombre de motifs, est fixé à 0.22 d'après la figure 4.19. Comparé à celui utilisé pour la STCD sur le Groenland, ce seuil est beaucoup plus faible. Cela est cohérent par rapport à ce que l'on observe sur la figure 4.17 qui montre que cette série dispose en général des indices de confiance plus faibles que la série du Groenland.

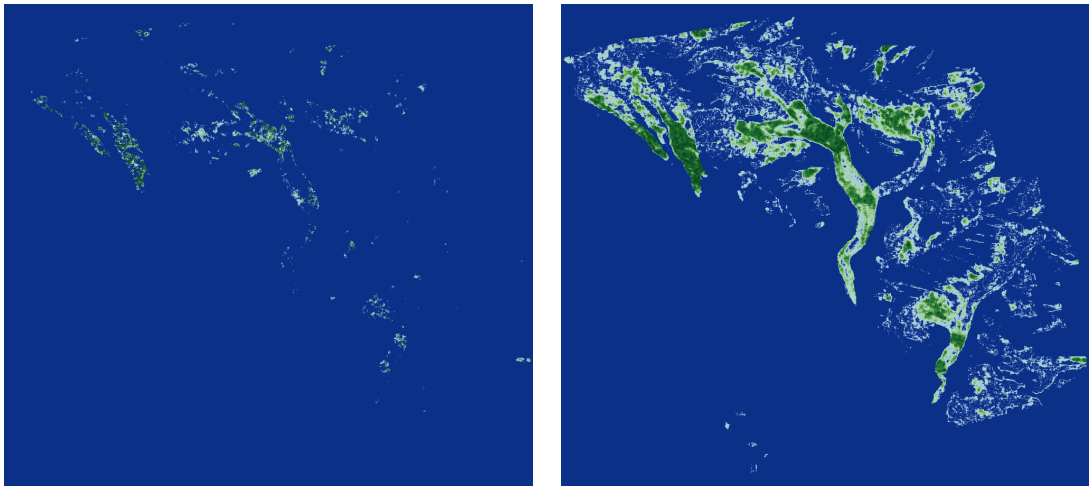
— Le nombre de motifs à garder à chaque extrémité du classement $K = 20$.

4.5.2.3 Effet du push partiel

Comme pour la STCD sur le Groenland, nous regardons la réduction de l'espace de recherche grâce à l'utilisation de la contrainte $C_{\widetilde{\rho}}$. Pour ce jeu de données, les courbes présentées sur la figure 4.20 montrent que le *push* partiel peut amener à une réduction importante du nombre de motifs à évaluer. Le temps d'exécution de l'extraction est alors lui aussi réduit dans les mêmes proportions (cf. Figure 4.21).

Pour un seuil γ, choisi dans la section précédente, de 0.22, nous parcourons au total 22992 motifs en utilisant la contrainte relaxée au lieu de 23174 motifs sans cette contrainte. Cela fait une réduction de 182 motifs. Pour d'autres choix de seuil γ plus élevés, i.e., si l'utilisateur est plus exigeant sur la fiabilité des motifs, le gain serait beaucoup plus important (cf. Figures 4.20 et 4.21)

Pour le seuil $\gamma = 0.22$, la méthode proposée extrait 5625 motifs SFG fiables maximaux. Le processus complet (extraction et classement des motifs par randomisation) prend au total 119894 secondes (environ 33 heures et 18 minutes) avec une consommation maximale de mémoire de 7470 Mo (même configuration logicielle et matérielle que pour la STCD sur le Groenland).



(a) Confiance minimale

(b) Confiance maximale

(c) Confiance moyenne

(d) Échelle de couleur

FIGURE 4.17 – Indice de confiance agrégé sur la série du Mont-Blanc : (a) minimale, (b) maximale, (c) moyenne. (d) échelle de couleur : de 0 (bleu foncé) à 1 (vert foncé)

4.5.2.4 Résultats qualitatifs

Les motifs SFG fiables extraits et leurs cartes LST reflètent différentes évolutions sur la série de déplacements, et parmi les cartes les mieux classées, l'une est montrée dans la figure 4.23. Les positions des glaciers concernés sont localisées sur la figure 4.15. La figure 4.23 représente le motif $3 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1 \to 3 \to 3 \to 2 \to 2$. Ce motif présente deux ralentissements : du niveau 3 au niveau 1 puis du niveau 3 au niveau 2. Les couleurs de la figure 4.23 indiquent les dates d'occurrence du dernier élément du motif (le dernier symbole 2) et, d'après l'échelle de couleurs de la figure 4.22, ces dates sont à la fin de la série (fin 2011).

Considérons maintenant le premier ralentissement capturé par ce motif (i.e., la partie

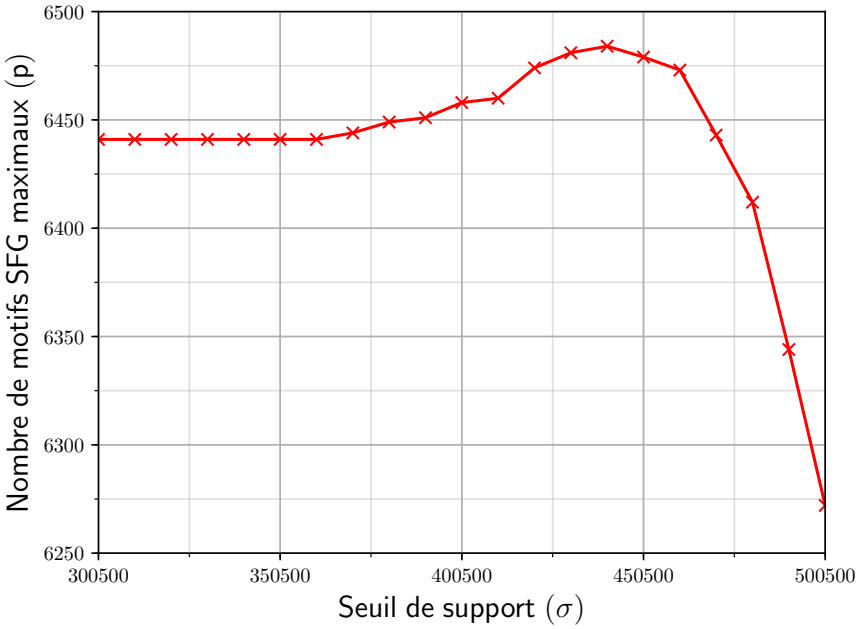


FIGURE 4.18 – Réglage du seuil de support minimal (Mont-Blanc)

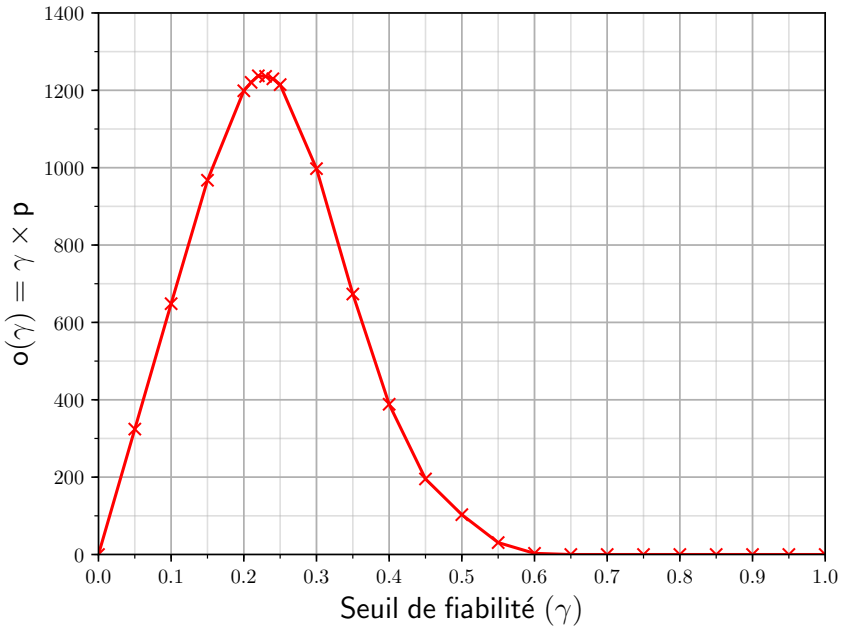


FIGURE 4.19 – Réglage du seuil de fiabilité (Mont-Blanc), avec p le nombre de motifs SFG maximaux obtenus pour le seuil γ

$3 \rightarrow 2 \rightarrow 2 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1$). Le début du phénomène se situe au début de l'été 2009 comme montré sur la figure 4.24a qui donne les dates d'apparition du premier symbole du motif (i.e., le symbole 3). En plus, nous pouvons observer sur la figure 4.24b, qui montre les dates d'apparition du dernier symbole de ce sous-motif (symbole 1), que ces dates correspondent à la fin de la première moitié de la série (i.e., été et automne de l'année 2009). La deuxième partie du motif ($3 \rightarrow 3 \rightarrow 2 \rightarrow 2$) révèle que le ralentissement s'est répété au cours de la deuxième moitié de la série (année 2011). À partir de ces observations, nous pouvons confirmer que pour

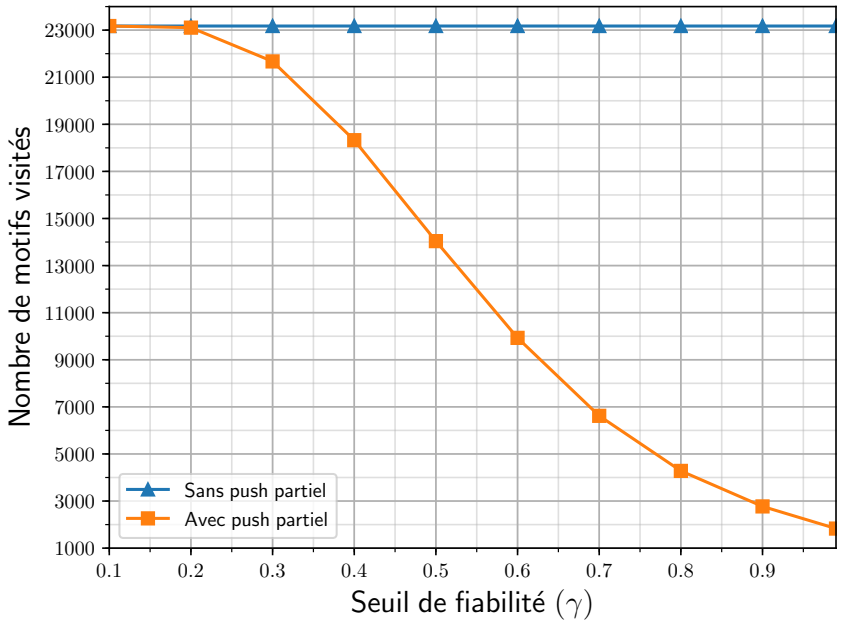


FIGURE 4.20 – Courbe de réduction de l'espace de recherche sur le Mont-Blanc

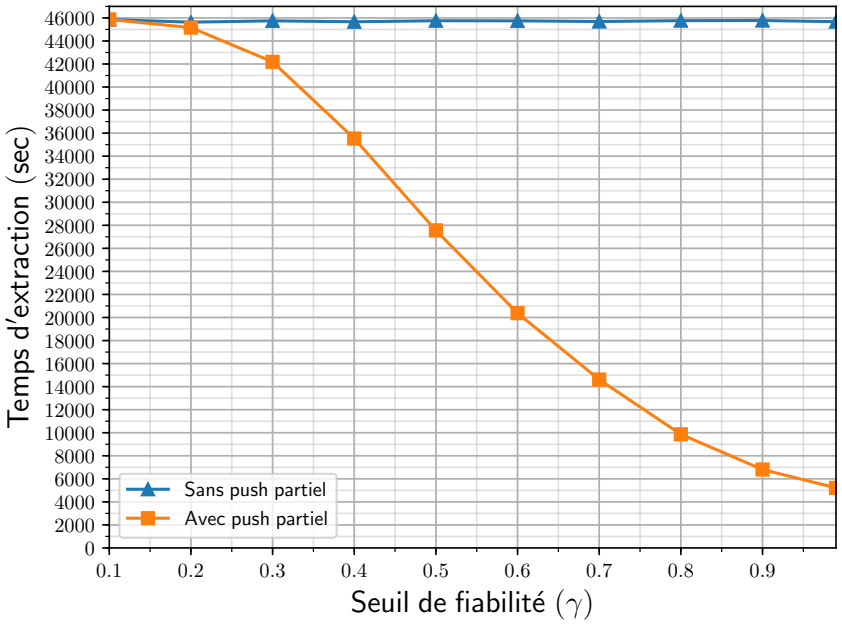


FIGURE 4.21 – Temps d'extraction

cette série, notre méthode basée sur les motifs séquentiels est capable d'extraire les motifs qui représentent le cycle annuel de la variation de vitesse. Ceci est un processus glaciologique connu pour les glaciers tempérés (Vincent et Moreau, 2016). Ce phénomène de cycle annuel a déjà été signalé sur trois transects [9] dans Fallourd (2012); Ponton *et al.* (2014), qui rapporte un ralentissement sur les glaciers de Taconnaz (1) et des Bossons (2) du début jusqu'au terminus, et du glacier du Géant(3) jusqu'au terminus du glacier de la Mer de Glace (4).

9. Le long de trois segments de droites.

En plus de ces études basées sur des transects, la carte LST fournit comme information l'étendue spatiale du phénomène, en particulier pour le ralentissement signalé dans (Fallourd, 2012, page 168) à ≈ 2000 m depuis le sommet des glaciers de Taconnaz et des Bossons, et dans la zone des séracs du glacier du Géant (emplacements donnés sur la figure 4.15 comme étiquettes b, d et 3′). Pour le glacier des Bossons, cette étendue spatiale indique que de telles fluctuations sont observées jusqu'à ≈ 3000 m a.s.l. [10] et suggèrent que la zone glaciaire froide [11] est aujourd'hui limitée aux altitudes plus élevées.

Dans cette série de plus grande taille et de confiance plus faible que celle du Groenland, notre méthode reste apte à effectuer une analyse globale, en permettant de rechercher des régularités parmi les estimations de déplacement de tous les principaux glaciers de la partie française du massif du Mont-Blanc. Trouver de tels motifs communs dans l'espace et le temps est une caractéristique intéressante de la méthode, et par exemple, la figure 4.23 montre que le même schéma de ralentissement que celui des glaciers de Taconnaz et des Bossons est également présent sur le glacier des Rognons (étiquette 7 sur la figure 4.15), dont la vitesse n'a jamais été étudiée jusqu'à présent.

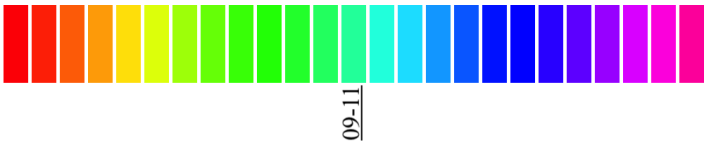


FIGURE 4.22 – Échelle de couleur des cartes LST : de rouge (Mai 2009) à vert (Octobre 2009) et de bleu (Mai 2011) à magenta (Septembre 2011)

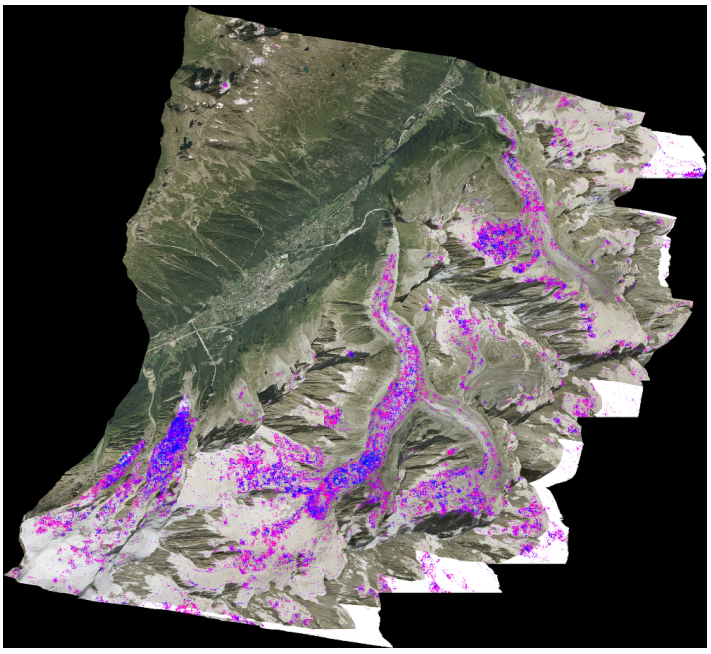
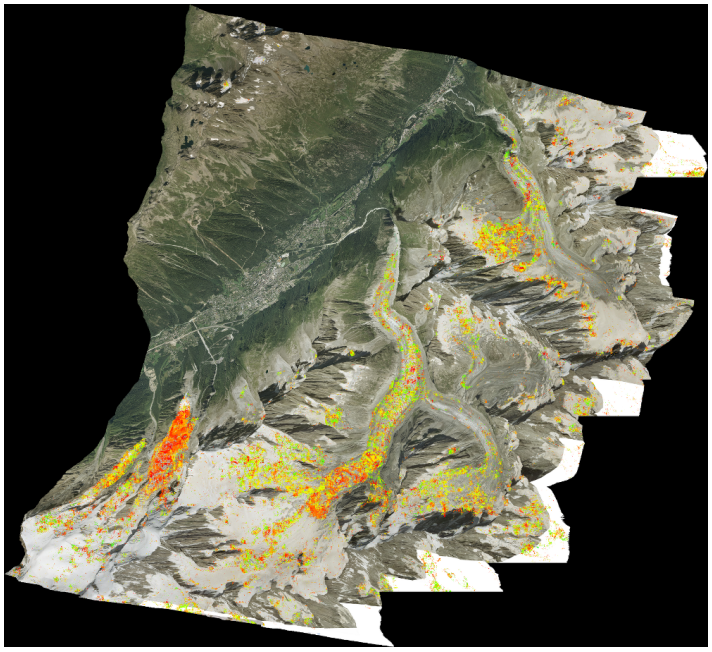
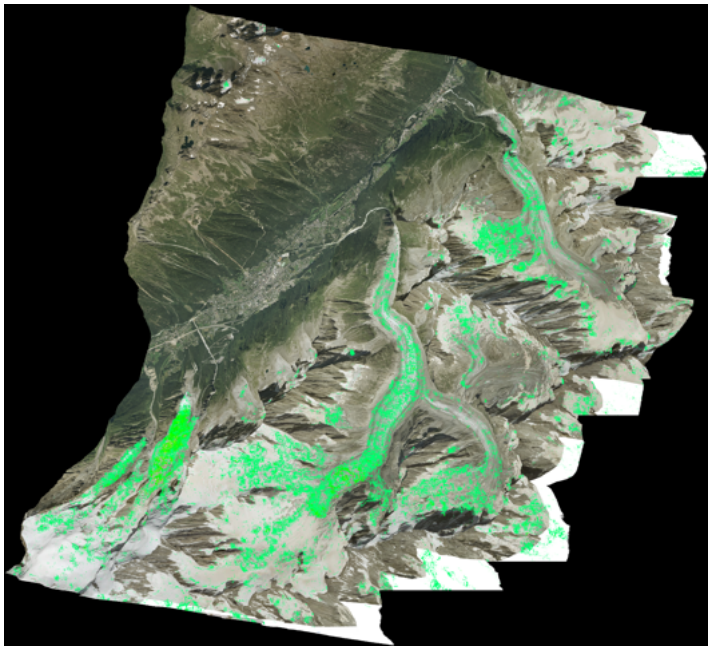


FIGURE 4.23 – Carte LST du motif $3 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1 \to 3 \to 3 \to 2 \to 2$ sur fond composite RGB en géométrie radar

10. Mètres au-dessus du niveau de la mer (*metres above sea level*).
11. La température de la glace des glaciers froids est en dessous du point de fusion. Le glacier colle alors au lit rocheux et les vitesses sont moins sensibles aux fluctuations saisonnières.



(a) Dates d'occurrence du premier symbole 3



(b) Dates d'occurrence du dernier symbole 1

FIGURE 4.24 – Dates d'occurrence des éléments du motif $3 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1 \to 3 \to 3 \to 2 \to 2$

4.5.3 Comparaison avec l'extraction des motifs SFG sur des STCD symboliques seuillées

Une approche alternative pour prendre en compte les indices de confiance, basée sur une combinaison des techniques disponibles dans l'état de l'art, consisterait à filtrer les données de déplacement qui possèdent un indice de confiance trop faible (Tedstone *et al.*, 2015; Dehecq *et al.*, 2015; Quincey *et al.*, 2015), et à extraire les motifs SFG comme dans Pericault *et al.*

(2015) pour analyser les évolutions de déplacements dans les données restantes.

Afin de comparer notre proposition à cette approche alternative, nous définissons une mesure qui indique, en moyenne, dans quelle mesure les motifs extraits décrivent la base de séquences en termes de nombre de données de déplacement couvertes. Soit $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \cdots \rightarrow \beta_m$ un motif SFG. Si une séquence $seq(x, y)$ est couverte par β, l'occurrence la plus fiable de β dans $seq(x, y)$ possède m éléments. Cumulé sur toutes les séquences, ceci représente un nombre d'éléments que nous appellerons *Data Point cover* (DP) de β, défini par :

$$DP_{cover}(\beta) \ = \ cover(\beta) * m$$

Considérons la moyenne de ce score sur un ensemble R de motifs retenus, le *Mean Data Point cover* (MDP) est simplement défini comme suit :

$$MDP_{cover}(R) = \frac{\sum_{\beta \in R} DP_{cover}(\beta)}{|R|} \tag{4.10}$$

Cette mesure va nous permettre de comparer notre méthode à la méthode alternative que nous appellerons *extraction de motifs SFG par seuillage*. Cette dernière ne nécessite pas de paramètre γ, mais nécessite un autre paramètre, $\rho_{threshold}$, qui est le seuil de confiance utilisé pour filtrer les données de déplacement de faible confiance. Pour les comparaisons, les paramètres de l'extraction de motifs SFG par seuillage sont réglés de façon à avoir des configurations d'exécution les plus proches possible, $\rho_{threshold}$ est fixé à la même valeur que γ et les autres paramètres (K et σ) restent identiques.

4.5.3.1 Comparaison sur le Groenland

Soit R l'ensemble des 40 motifs retenus par notre approche pour la STCD sur le Groenland, la valeur de MDP obtenue est $MDP_{cover}(R) = 201981.6$, où R représente l'ensemble de 40 motifs sélectionnés. Avec la méthode alternative, nous obtenons aussi un ensemble R' contenant 40 motifs mais avec une mesure de $MDP_{cover}(R')$ valant 188418.6.

Cette variation de la mesure MDP_{cover} montre que la méthode proposée extrait des motifs qui, en moyenne, représentent 7.2% plus de données de déplacement que ceux obtenus avec l'extraction de motifs SFG par seuillage. La distribution des couvertures est donnée par la figure 4.25, qui représente les histogrammes du nombre de motifs en fonction de la valeur DP_{cover}. Elle montre que les fortes valeurs de DP_{cover} ont tendance à correspondre plutôt à des motifs SFG fiables (en bleu) qu'à des motifs SFG obtenus avec l'approche alternative (en orange).

4.5.3.2 Comparaison sur le Mont-Blanc

Pour ce jeu de données, nous avons, pour l'ensemble R des 40 motifs obtenus par notre approche, une valeur de $MDP_{cover}(R)$ égale à 5231810.6. L'approche alternative sélectionne un ensemble R' de 40 motifs pour lequel $MDP_{cover}(R')$ vaut 3410626.7. Comparée à cette approche, la méthode proposée extrait des motifs qui, en moyenne, représentent 53.4% plus

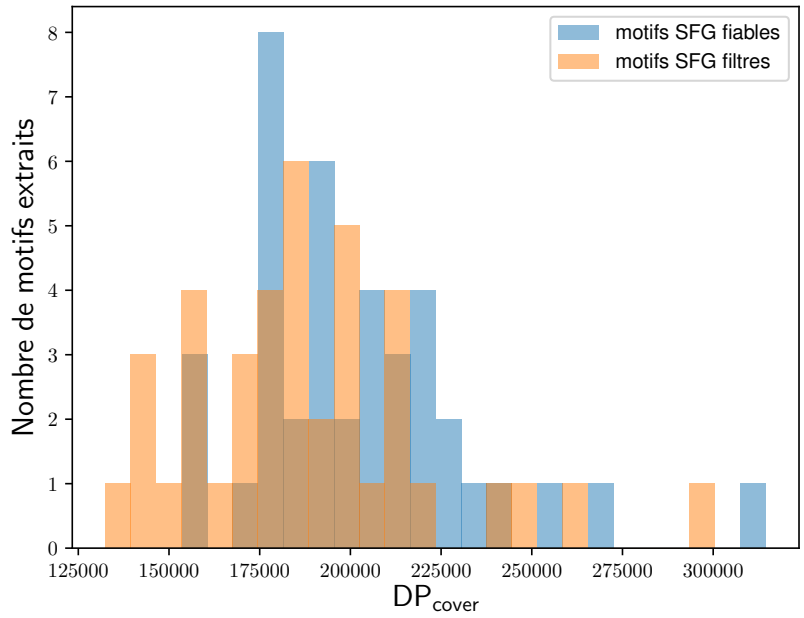


FIGURE 4.25 – Nombre de motifs extraits vs. les mesures individuelles DP_{cover} (Groenland)

de données de déplacement. Cette amélioration est de loin supérieure à celle observée pour la série du Groenland et provient directement du fait que les données dans la série du massif du Mont-Blanc sont en général moins fiables. En d'autres termes, plus les indices de confiance sont faibles, plus le traitement de l'indice de confiance au niveau des occurrences est utile et permet d'explorer des parties de données qui auraient été supprimées par une extraction de motifs SFG par seuillage. La figure 4.26 confirme ce comportement : tous les motifs SFG fiables (en bleu) couvrent plus de données de déplacement que les motifs SFG extraits par l'approche par seuillage (en orange).

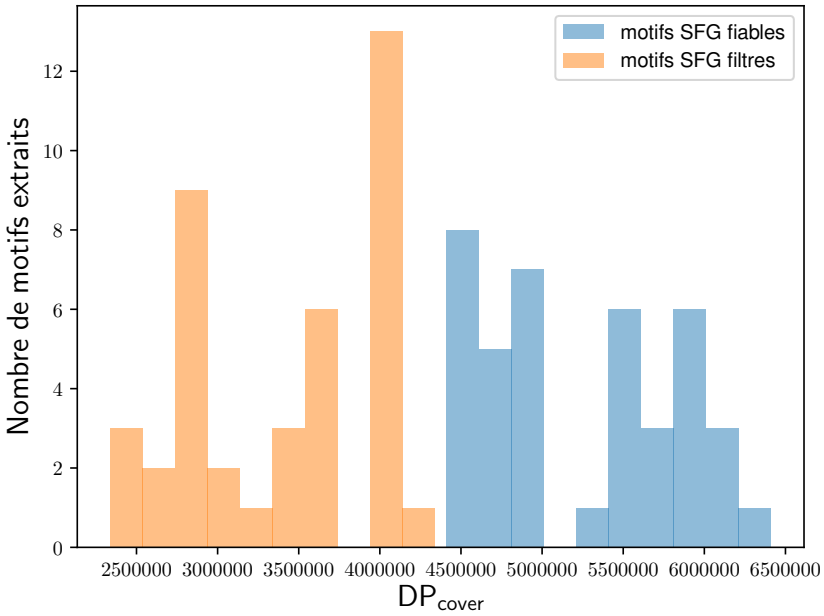


FIGURE 4.26 – Nombre de motifs extraits vs. les mesures individuelles DP_{cover} (Mont-Blanc)

4.6 Conclusions

Dans ce chapitre, la première proposition du travail de la thèse est présentée. Elle concerne la prise en compte des indices de confiance associés à nos données. Cette tâche s'effectue grâce aux apports suivants :

1. l'introduction des mesures de fiabilité à différents niveaux (occurrence, séquence, base de séquences) pour sélectionner seulement les motifs dont la mesure de fiabilité dépasse un seuil défini par l'utilisateur,
2. le calcul efficace par programmation dynamique de la mesure de fiabilité de chaque motif au niveau de la séquence,
3. l'introduction d'une contrainte relaxée qui est anti-monotone, permettant une stratégie de *push* partiel qui s'avère efficace lorsque l'utilisateur souhaite des motifs très fiables dans des données dont les indices de confiance ne sont pas élevés,
4. une méthode de réglage des paramètres basée sur un compromis entre une bonne fiabilité des motifs et une bonne description de la base de séquences.

Notre proposition a été appliquée à deux STCD de nature très différente et couvrant des glaciers de différentes tailles : une à partir des images optiques couvrant les glaciers du Groenland, et une à partir des images SAR couvrant les glaciers du massif du Mont-Blanc. Des motifs intéressants ont été extraits des deux séries temporelles. Différentes cartes LST ont été produites pour localiser dans l'espace et dans le temps les évolutions de déplacements associées à ces motifs. Ces évolutions et les cartes LST obtenues confirment les connaissances glaciologiques concernant ces zones. Elles complètent ces connaissances en fournissant également des informations supplémentaires au niveau spatial et en pointant des phénomènes locaux qui ne semblent pas encore répertoriés. Finalement, comparée à une autre méthode basée sur les techniques de l'état de l'art, notre proposition extrait des motifs qui représentent une quantité plus importante de données de déplacement.

Chapitre 5

Sélection de motifs séquentiels complémentaires sur critère informationnel

Sommaire

5.1 Introduction

Dans ce chapitre, nous envisageons les Séries Temporelles de Champs de Déplacements (STCD) sous un angle probabiliste (Suciu et Dalvi, 2005). Ce dernier permet en effet de traiter des données incertaines (Aggarwal, 2009; Fournier-Viger *et al.*, 2017). C'est ainsi que l'incertitude des données symboliques peut être représentée par exemple en donnant, pour chaque transaction/événement, la probabilité de chacun des symboles qui peuvent y apparaître (Leung, 2011). Ces probabilités peuvent être utilisées dans le cadre des *mondes possibles*, défini par Abiteboul *et al.* (1987), pour extraire des *itemsets* dont l'espérance de la mesure du support dépasse un seuil utilisateur. Cette contrainte, appelée contrainte de *support attendu*, est également utilisée pour réduire l'espace de recherche (Leung *et al.*, 2014). Des stratégies complémentaires peuvent être adoptées pour réduire la consommation de ressources de calcul, en utilisant par exemple des structures de données efficaces (e.g., Leung *et al.* (2008)), de nouveaux algorithmes (e.g., Sun *et al.* (2010)), des techniques d'échantillonnage (e.g., Calders *et al.* (2010)), ou encore des contraintes supplémentaires (Leung, 2011). Les techniques d'extraction des *itemsets* fréquents dans les bases de données probabilistes ont été également étendues aux flux de données incertaines (e.g., Leung et Hao (2009); Leung et Jiang (2011)).

Afin de se concentrer sur les *itemsets* fréquents les plus intéressants, des approches top-k (e.g., Zhang *et al.* (2008); Cormode *et al.* (2009)) et la recherche des motifs optimaux de compression basée sur le principe Longueur de Description Minimale (en anglais *Minimal Description Length*) (MDL) (e.g., Bonchi *et al.* (2011)) ont été proposées.

Muzammal et Raman (2010) ont adapté l'approche probabiliste aux bases de séquences. Dans ce cas, une base de séquences probabiliste est vue comme un ensemble de sources (i.e., les séquences) générant des événements. En ce qui concerne les STCD, une source correspond à une localisation (x, y) et un événement correspond à un symbole représentant une mesure de déplacement et sa date d'occurrence. L'incertitude peut ainsi apparaître à plusieurs niveaux : les sources, les événements, ou les dates d'occurrence des événements. Muzammal et Raman ont modélisé les incertitudes liées aux sources et aux événements en considérant une base de séquences probabiliste dans laquelle chaque événement est décrit par sa source, sa date d'apparition, son symbole/type d'événement et une probabilité. Cette dernière représente soit la possibilité que l'événement se soit effectivement produit (incertitude de l'événement), soit la possibilité que l'événement soit associé à la source (incertitude de la source). Pour un tel événement, les probabilités de tous les autres symboles ou sources possibles sont considérées comme égales à 0. Un tel cadre permet d'extraire des motifs séquentiels fréquents en s'appuyant sur la mesure de *support attendu* ou de *fréquence probabiliste* (Muzammal et Raman, 2010). Le problème d'extraction de motifs séquentiels fréquents en utilisant la fréquence probabiliste est #P-complet lorsqu'on considère les incertitudes de source. Dans le cas du support attendu, il peut être calculé avec des algorithmes efficaces (Muzammal et Raman, 2011, 2015). Comme expliqués dans Muzammal et Raman (2011), ces algorithmes peuvent également être utilisés dans le cas des incertitudes d'événements. Dans ce cadre, un autre algorithme basé sur le support attendu a également été proposé dans Hooshsadat *et al.* (2012). Zhao *et al.* (2012); Zhou Zhao *et al.* (2014) ont proposé, quant à eux, des algorithmes basés sur la projection de la base de séquences pour extraire les motifs séquentiels sous une sémantique de fréquence probabiliste. La même sémantique a été utilisée pour extraire des épisodes sériels fréquents à partir d'une seule séquence d'événements avec des incertitudes d'événements (Wan *et al.*, 2013). Quant aux incertitudes liées aux dates d'apparition, elles peuvent être prises en compte en modélisant le temps par des variables aléatoires (Ge *et al.*,

2015, 2017).

Dans notre cas, nous postulons que les indices de confiance associés à des estimations de déplacement peuvent être interprétés comme étant des probabilités. Toutefois, l'indice de confiance en un point nous donne seulement la probabilité de l'un des symboles. Cela correspond à une base probabiliste dont les distributions ne sont que partiellement connues, tandis que les méthodes présentées précédemment considèrent seulement des bases probabilistes dont les distributions sont connues de façon complète.

Nous nous intéressons dans ce chapitre à des motifs complémentaires sur le plan informationnel, avec une vision probabiliste. Nous proposons ainsi une sélection des motifs dont les occurrences ont tendance à décrire l'ensemble de données de manière complémentaire, tout en prenant en compte des indices de confiance. Plus précisément, les événements sont considérés comme des variables aléatoires discrètes dont les distributions correspondent à celles que l'on connaît. Ces distributions sont modifiées suite à la découverte des occurrences d'un motif. Le gain informationnel obtenu via la connaissance des occurrences peut être ensuite mesuré en quantifiant ces modifications. De cette manière, un ensemble de motifs conduisant à un gain d'information élevé peut être construit de façon itérative, en sélectionnant, à chaque itération, le motif associé au plus haut gain informationnel, compte tenu des motifs sélectionnés lors des itérations précédentes.

Après avoir défini les STCD comme étant des bases de séquences probabilistes partielles et redéfini la notion de motif séquentiel fréquent dans ce contexte en section 5.2, le principe général de la méthode proposée est présenté dans la section 5.3. Le calcul du gain informationnel associé à un motif séquentiel est présenté dans la section 5.4. Enfin, la section 5.5 fournit des expériences sur des STCD, montrant qu'avec de bonnes mesures de complémentarité spatiotemporelle, les motifs sélectionnés aident les utilisateurs finaux à compléter leur connaissance des zones étudiées.

5.2 Représentation d'une STCD en tant que base probabiliste partielle

Dans cette section, nous donnons la définition d'une base de séquences probabiliste partielle, laquelle peut représenter une STCD dont les indices de confiance sont interprétés comme des probabilités. Pour ce faire, nous définissons un événement probabiliste partiel comme suit :

Définition 5.1 (Item). Un *item* est un élément dans l'ensemble des symboles distincts $\mathbb{E} = \{e_1, e_2, \ldots, e_n\}$. Dans ce chapitre, l'*item* est également appelé *type d'événement*, ou encore *symbole*.

Dans le cadre des STCD, un type d'événement représente la grandeur discrétisée d'une mesure de déplacement.

Définition 5.2 (Probabilité d'un type d'événement). Soit une date $t \in \mathbb{N}$ à laquelle un type d'événement est observé. La *probabilité d'un type d'événement*, notée $\rho^{t,e}$, avec $t \in \mathbb{N}$, $e \in \mathbb{E}$ et $\rho^{t,e} \in [0, 1]$, est la probabilité d'observer le type d'événement e à la date t. Par définition, nous avons : $\sum_{e \in \mathbb{E}} \rho^{t,e} = 1$.

Définition 5.3 (Événement probabiliste partiel). Un *événement probabiliste partiel* est représenté sous forme d'un triplet $\langle t, e, \rho^{t,e} \rangle$ avec $t \in \mathbb{N}$, et $e \in \mathbb{E}$, e étant le symbole le plus

probable à la date t, i.e., $\rho^{t,e} > \rho^{t,e'}$ pour tout $e' \in \mathbb{E} \setminus \{e\}$. Cet événement probabiliste est dit partiel parce qu'il ne spécifie que la probabilité du symbole le plus probable, on notera que ceci implique $\rho^{t,e} \in]\frac{1}{|\mathbb{E}|}, 1]$. Par souci de clarté, lorsque le contexte le permet, un événement probabiliste partiel $\langle t, e, \rho^{t,e}\rangle$ est noté $\langle t, e, \rho\rangle$.

Contrairement à un événement probabiliste complet où chacun des types d'événement est associé à une probabilité, un événement probabiliste partiel dispose seulement de la probabilité d'un seul type d'événement e, la probabilité des autres types d'événement $\mathbb{E} \setminus \{e\}$ étant inconnues. Ainsi, $\langle t, e, \rho\rangle$ n'est pas une *probabilité existentielle* [1] comme dans Muzammal et Raman (2010) puisque l'existence de l'événement est certaine dans notre cas. Ce qui est incertain dans notre approche, c'est le type d'événement associé à chaque événement probabiliste partiel. Dans le cadre des STCD, cette probabilité peut être déduite à partir des indices de confiance ou d'incertitude obtenus à l'issue de l'estimation du déplacement.

Par exemple, un événement $\langle 5, A, 50\%\rangle$ nous indique qu'à la date $t = 5$, il y a eu un événement, et que pour cet événement, le symbole A a une probabilité d'apparition de 50%. Il ne précise pourtant pas la probabilité d'apparition des autres symboles dans l'ensemble $\mathbb{E}$. Nous pouvons néanmoins savoir que les probabilités des autres symboles sont réelles positives, inférieures à 50%, et que la somme des probabilités vaut 1.

Une base de séquences probabiliste partielle peut alors être définie comme suit :

Définition 5.4 (Séquence d'événements probabilistes partiels). Pour une source s, une *séquence d'événements probabilistes partiels* de s est une paire (sid, seq), où sid est l'identifiant de la source s et seq est une suite d'événements probabilistes partiels, $seq = \langle\langle t_1, e_1, \rho_1\rangle, \langle t_2, e_2, \rho_2\rangle, \ldots, \langle t_n, e_n, \rho_n\rangle\rangle$, où les dates d'événement sont dans l'ordre chronologique, i.e., $(t_1 < t_2 < \cdots < t_n)$.

Dans le cadre des STCD, une séquence d'événements représente l'évolution de déplacements date par date sur un pixel précis de la STCD. Elle est alors notée $((x, y), seq)$ où (x, y), représente l'identifiant de la source, c'est-à-dire de la séquence.

Définition 5.5 (Base de séquences probabiliste partielle). Une Base de Séquences Probabiliste Partielle (BSPP) est un ensemble de séquences d'événements probabilistes partiels. Un exemple d'une Base de Séquences Probabiliste Partielle (BSPP) est donné dans l'exemple 5.2.1.

Exemple 5.2.1. Soit $\mathbb{E} = \{A, B, C\}$ l'ensemble des *items*. Une BSPP, notée $\mathcal{B}$, est donnée ci-après. Elle contient quatre séquences d'événements probabilistes partiels, dont chacune contient quatre événements probabilistes partiels qui se produisent aux dates t_1, t_2, t_3 et t_4. $\mathcal{B}$ pourrait par exemple donner les niveaux de dioxyde de carbone mesurés par quatre capteurs de qualité de l'air qui fournissent également les probabilités d'avoir ces niveaux en fonction des conditions de fonctionnement (e.g., température, humidité). Ces probabilités expriment à quel niveau nous pouvons faire confiance aux mesures.

$$\begin{aligned}\mathcal{B} = \{&(1, \langle\langle t_1, B, 0.4\rangle, \langle t_2, C, 0.9\rangle, \langle t_3, A, 0.7\rangle, \langle t_4, C, 0.5\rangle\rangle),\\ &(2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_3, C, 0.4\rangle, \langle t_4, A, 0.5\rangle\rangle),\\ &(3, \langle\langle t_1, B, 0.9\rangle, \langle t_2, A, 0.9\rangle, \langle t_3, A, 0.7\rangle, \langle t_4, A, 0.8\rangle\rangle),\\ &(4, \langle\langle t_1, C, 0.8\rangle, \langle t_2, A, 0.7\rangle, \langle t_3, B, 0.8\rangle, \langle t_4, A, 0.7\rangle\rangle)\}\end{aligned}$$

1. Sous cette sémantique, un événement $(5, A, 40\%)$ indiquerait qu'à la date $t = 5$, l'événement lui-même peut apparaître avec une probabilité de 40%, et le type d'événement sera dans ce cas le symbole A.

Dans notre cas, $\mathcal{B}$ peut également représenter une STCD contenant quatre séquences représentant l'évolution des zones pour quatre périodes distinctes.

À partir des BSPP, nous proposons d'extraire des *motifs séquentiels fréquents* comme définis à l'origine dans Agrawal et Srikant (1995). De façon formelle, dans notre contexte où un événement est décrit par un type d'événement unique, un motif séquentiel est défini comme suit :

Définition 5.6 (Motif séquentiel). Un *motif séquentiel* β est un tuple $\langle \beta_1, \beta_2, \ldots, \beta_m \rangle$ où $\beta_1, \beta_2, \ldots, \beta_m$ sont des types d'événement dans $\mathbb{E}$ et m est la longueur de β. Un tel motif est également noté $\beta = \beta_1 \to \beta_2 \to \cdots \to \beta_m$.

Définition 5.7 (Occurrence, support). Soit $(sid, seq) = (sid, \langle \langle t_1, e_1, \rho_1 \rangle, \langle t_2, e_2, \rho_2 \rangle, \ldots, \langle t_n, e_n, \rho_n \rangle \rangle)$ une séquence d'événements probabilistes partiels et $\beta = \beta_1 \to \beta_2 \to \cdots \to \beta_m$ un motif séquentiel. S'il existe $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ tels que $\beta_1 = e_{i_1}, \beta_2 = e_{i_2}, \ldots, \beta_m = e_{i_m}$, alors $o = (sid, \langle \langle t_{i_1}, e_{i_1}, \rho_{i_1} \rangle, \langle t_{i_2}, e_{i_2}, \rho_{i_2} \rangle, \ldots, \langle t_{i_m}, e_{i_m}, \rho_{i_m} \rangle \rangle)$ est une *occurrence* de β dans la séquence (sid, seq). Dans ce cas, nous appelons $t_{i_1}, t_{i_2}, \ldots, t_{i_m}$ les *dates d'occurrence*, et la séquence (sid, seq) est dite *couverte* par β. L'ensemble des séquences couvertes par β représente sa *couverture*, notée $cover(\beta)$. Le *support* de β dans une BSPP $\mathcal{B}$, noté $support(\beta)$, est le nombre de séquences dans $\mathcal{B}$ couvertes par β, i.e., $|cover(\beta)|$.

En comptant le nombre de séquences d'événements probabilistes partiels dans lesquelles un motif β se produit au moins une fois, il est possible de se concentrer sur les *motifs séquentiels fréquents* :

Définition 5.8 (Motif séquentiel fréquent). Soit $\sigma \in [0, 1]$ un seuil de support relatif, β un motif séquentiel, et $\mathcal{B}$ une BSPP. Un motif séquentiel β est fréquent dans $\mathcal{B}$ si $support(\beta)/|\mathcal{B}| \geq \sigma$, où $|\mathcal{B}|$ est le nombre de séquences d'événements probabilistes partiels dans $\mathcal{B}$.

Exemple 5.2.2. Considérons la BSPP $\mathcal{B}$ donnée dans l'exemple 5.2.1. Avec $\sigma = \frac{1}{2}$, c'est-à-dire que les motifs séquentiels doivent apparaître dans au moins deux séquences d'événements probabilistes partiels pour être considérés comme fréquents, les motifs séquentiels fréquents sont indiqués dans la table 5.1.

motif fréquent	sid séquences couvertes	support
A	$1, 2, 3, 4$	4
B	$1, 2, 3, 4$	4
C	$1, 2, 4$	3
$A \to A$	$3, 4$	2
$B \to A$	$1, 2, 3, 4$	4
$B \to C$	$1, 2$	2
$C \to A$	$1, 2, 4$	3
$C \to C$	$1, 2$	2
$B \to C \to A$	$1, 2$	2
$B \to C \to C$	$1, 2$	2

TABLE 5.1 – Les motifs séquentiels fréquents dans $\mathcal{B}$

Par exemple, considérons le motif $B \to C$:

— Dans la première séquence $(1, \langle\langle t_1, B, 0.4\rangle, \langle t_2, C, 0.9\rangle, \langle t_3, A, 0.7\rangle, \langle t_4, C, 0.5\rangle\rangle)$, nous avons les occurrences suivantes : $o_1 = (1, \langle\langle t_1, B, 0.4\rangle, \langle t_2, C, 0.9\rangle\rangle)$ et $o_2 = (1, \langle\langle t_1, B, 0.4\rangle, \langle t_4, C, 0.5\rangle\rangle)$.
— Dans la deuxième séquence $(2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_3, C, 0.4\rangle, \langle t_4, A, 0.5\rangle\rangle)$, nous avons : $o_3 = (2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle\rangle)$ et $o_4 = (2, \langle\langle t_1, B, 0.5\rangle, \langle t_3, C, 0.4\rangle\rangle)$.
— Dans les deux autres séquences, il n'existe aucune occurrence de $B \to C$.

Les séquences couvertes par ce motif sont donc les séquences 1 et 2, et son support est par conséquent 2 séquences.

En plus de la contrainte de fréquence minimale, d'autres contraintes telles que celle de *maximalité* (e.g., Luo et Chung (2004)) ou celle de *motifs clos* (e.g., Yan *et al.* (2003)) peuvent être utilisées pour se concentrer sur une collection plus petite de motifs. Pour rappel, un motif dans une collection de motifs $\mathcal{C}$ est maximal lorsqu'il n'est sous-motif d'aucun motif dans $\mathcal{C}$. Un motif β est clos lorsqu'il n'est sous-motif d'aucun motif qui a le même support que β. Par exemple, l'ensemble des motifs séquentiels fréquents maximaux dans l'exemple précédent est $\{A \to A, B \to C \to A, B \to C \to C\}$, tandis que l'ensemble des motifs séquentiels fréquents clos est $\{A \to A, B \to A, C \to A, B \to C \to A, B \to C \to C\}$. Nous pouvons remarquer que l'ensemble des motifs séquentiels fréquents maximaux est un sous-ensemble de celui des motifs séquentiels fréquents clos qui est, à son tour, un sous-ensemble des motifs séquentiels fréquents.

Dans le cas des STCD où l'identifiant des séquences contient la localisation, il est également possible de considérer la contrainte de connexité moyenne, définie dans la section 4.3, afin de se concentrer sur des motifs apparaissant sur les séquences qui ne sont pas trop isolées en moyenne, i.e., des motifs Séquentiels Fréquents Groupés (SFG).

Même si de telles contraintes sont utiles, les utilisateurs finaux sont généralement confrontés à un grand nombre de motifs sélectionnés. Par conséquent, des méthodes automatiques pour sélectionner un nombre limité de motifs les plus prometteurs sont indispensables. Par exemple, Lam *et al.* (2014) sélectionnent les motifs qui compressent le mieux la base de séquences ou Méger *et al.* (2015) évaluent les motifs avec des techniques de randomisation. Néanmoins, à notre connaissance, aucune méthode de sélection exploitant les probabilités disponibles dans des bases de séquences probabilistes partielles n'est disponible. Une telle méthode est donc proposée dans ce travail. Son principe général est présenté dans la section suivante.

5.3 Principe général

Considérons une BSPP $\mathcal{B}$ qui contient au total N événements probabilistes partiels [2]. La quantité d'information de $\mathcal{B}$ peut être exprimée par l'entropie de l'ensemble de N variables aléatoires (Rényi et al., 1961), notées $\mathcal{V} = \{X_1, X_2, \ldots, X_N\}$, où chaque variable aléatoire est associée de manière bijective à un événement probabiliste partiel. Au début, lorsqu'aucune occurrence n'est découverte, nous n'avons aucune information sur la distribution des variables aléatoires. Dans ce cas de figure, la loi de probabilité de chaque variable aléatoire dans $\mathcal{V}$ est considérée comme étant uniforme sur l'ensemble $\mathbb{E}$. En effet, cette distribution correspond à un maximum d'incertitude sur la réalisation de la variable aléatoire, à un minimum de connaissance sur celle-ci, ou encore un maximum d'entropie de la variable (Rényi et al.,

2. N = nombre de séquences × longueur de chaque séquence.

1961). Supposons maintenant que notre connaissance sur $\mathcal{B}$ soit seulement partielle, i.e., notre connaissance de la base de séquences probabiliste partielle est elle même partielle. L'intuition clé de notre méthode est que si on nous donne les occurrences d'un motif β, alors cette connaissance fournit des informations supplémentaires sur les distributions des variables dans $\mathcal{V}$ et réduit ainsi l'entropie de $\mathcal{V}$. Plus cette réduction est importante, plus le motif est informatif. Dans ce qui suit, ce gain d'information par rapport à la connaissance partielle actuelle que nous avons sur $\mathcal{B}$, apporté par les occurrences d'un motif β, est noté $\Delta(\beta)$. Le problème consiste donc à chercher un ensemble de motifs qui réduit au maximum, du point de vue de l'utilisateur, l'entropie de $\mathcal{B}$, autrement dit qui apporte un maximum de connaissance sur $\mathcal{B}$. La recherche d'un ensemble optimal de motifs par rapport à un critère d'entropie est en général NP-difficile (Lam *et al.*, 2014). Comme l'algorithme SeqKrimp (Lam *et al.*, 2014) qui a été conçu pour trouver un ensemble de motifs séquentiels qui compressent une base de séquences, nous adoptons une stratégie heuristique gloutonne pour sélectionner les motifs séquentiels les plus informatifs et complémentaires de manière itérative (Algorithme 2).

Input : k, le nombre de motifs à sélectionner, P, l'ensemble de motifs séquentiels dans la base de séquences $\mathcal{B}$, $\mathcal{V}$, l'ensemble de variables aléatoires représentant $\mathcal{B}$
Output : Φ, ensemble de motifs séquentiels informatifs et complémentaires
1 $\Phi \leftarrow \emptyset$
2 Pour tout $X \in \mathcal{V}$, définir la distribution de X comme étant uniforme
3 **while** ($|\Phi| < k$ *et* $|P| > 0$) **do**
4 $\quad \beta^* \leftarrow \text{argmax}_{\beta \in P}(\Delta(\beta))$, i.e., le motif qui maximise le gain d'information par rapport à la connaissance actuelle des distributions des variables de $\mathcal{V}$
5 $\quad \Phi \leftarrow \Phi \cup \{\beta^*\}$
6 $\quad P \leftarrow P \setminus \{\beta^*\}$
7 $\quad$ Pour tout $X \in \mathcal{V}$, mettre à jour la distribution de X
8 **end**
9 **return** Φ

Algorithme 2 : Sélection d'un ensemble de k motifs séquentiels informatifs et complémentaires

L'algorithme 2 prend les paramètres suivants en entrée : k, le nombre de motifs à sélectionner ; $\mathcal{V}$, l'ensemble des variables aléatoires représentant le contenu informationnel de $\mathcal{B}$; et P, une collection de motifs séquentiels dans $\mathcal{B}$. L'ensemble P peut contenir n'importe quel type de motifs séquentiels tels que les fréquents, les clos, les maximaux, ou encore les motifs SFG.

L'initialisation consiste à créer un ensemble vide pour contenir les motifs sélectionnés (ligne 1) et à considérer une distribution uniforme sur le domaine $\mathbb{E}$ pour toutes les variables de $\mathcal{V}$ (ligne 2). Cette distribution correspond à un maximum d'entropie, c'est-à-dire que l'utilisateur ne connaît rien sur la réalisation de la variable aléatoire (à l'exception du domaine $\mathbb{E}$). Tant que l'ensemble P n'est pas vide, l'algorithme itère jusqu'à ce que les k motifs séquentiels informatifs et complémentaires soient trouvés (ligne 3), qui est finalement retourné à la sortie de l'algorithme. À chaque itération, le motif séquentiel β^* dont les occurrences conduisent au gain d'information le plus élevé est sélectionné (ligne 4). Ce motif β^* sera ensuite ajouté à l'ensemble des motifs sélectionnés Φ (ligne 5). Il sera également enlevé de l'ensemble initial P (ligne 6). Enfin, à la ligne 7, les distributions des variables aléatoires

concernées par les occurrences ayant servi pour le calcul du gain d'information du motif β^* sont mises à jour.

Soit β un motif séquentiel et $\Delta_{occ}(o)$ le gain d'information obtenu en révélant l'occurrence o de β. Pour chaque séquence d'événements probabilistes partiels (sid, s) couverte par β, nous considérons que seule la *meilleure* occurrence de β, i.e., celle menant au gain d'information le plus élevé, est révélée à l'utilisateur. Ainsi, la quantité d'information fournie par β dans (sid, s) est $\Delta((sid, s), \beta) = \max_{o \in \mathcal{O}}\{\Delta_{occ}(o)\}$, où $\mathcal{O}$ représente l'ensemble des occurrences de β dans (sid, s). S'il y a plusieurs occurrences qui apportent la même quantité d'information, alors une occurrence va être sélectionnée de façon aléatoire. Enfin, la mesure du gain $\Delta(\beta)$, qui évalue l'intérêt du motif β sur toutes les séquences qu'il couvre, est définie comme suit :

$$\Delta(\beta) = \frac{\sum_{(sid,s) \in cover(\beta)} \Delta((sid, s), \beta)}{support(\beta)} \tag{5.1}$$

L'intuition de la normalisation dans ce calcul (i.e., la division par $support(\beta)$) est qu'un motif couvrant un grand nombre de séquences d'événements probabilistes partiels avec des occurrences apportant peu d'informations est considéré comme étant moins intéressant qu'un motif couvrant moins de séquences, mais reposant sur des occurrences informatives.

Calculer le gain $\Delta_{occ}(o)$ est une étape cruciale, qui peut être réalisée en exploitant les types d'événements et les probabilités fournis par les éléments formant o. Dévoiler un événement probabiliste partiel $\langle t, e, \rho \rangle$ permet d'affiner la distribution actuellement connue p de la variable aléatoire associée X par une nouvelle distribution q en imposant que $\Pr(X = e) = \rho$. Les types d'occurrences, comme les occurrences *minimales* définies dans Mannila *et al.* (1997), peuvent également fournir des contraintes supplémentaires. Par exemple, pour une occurrence minimale d'un motif $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \cdots \rightarrow \beta_m$, les événements situés entre la date d'occurrence de β_1 et la date d'occurrence de β_2 ne peuvent pas avoir le type d'événement β_1, car sinon l'occurrence ne serait pas minimale. Ces contraintes peuvent être utilisées pour affiner la distribution des variables aléatoires associées aux événements probabilistes partiels ayant lieu entre les événements qui forment chaque occurrence du motif β. Les différentes contraintes possibles et leurs combinaisons sont définies et étudiées dans la section suivante.

Plus généralement, considérons une variable aléatoire X et sa distribution actuellement connue p. Les informations sur X apportées par une occurrence o sont exprimées avec une contrainte ξ. Le problème consiste alors à trouver une distribution q pour X qui est la plus proche de p, d'un point de vue informationnel (i.e., celle qui exprime la plus petite quantité de gain d'information), et qui satisfait ξ. Pour ce faire, nous utilisons la divergence *Kullback-Leibler*, définie comme suit :

$$D(q||p) = \sum_{e \in \mathbb{E}} q(e) \log_2 \left(\frac{q(e)}{p(e)} \right)$$

où $p(e)$ (resp. $q(e)$) correspond à $\Pr(X = e)$ dans la distribution p (resp. q). $D(q||p)$ quantifie *le gain d'information* obtenu si la distribution p est remplacée par la distribution q (Rényi et al., 1961). Cette mesure a été utilisée par exemple dans Fernando *et al.* (2014) pour estimer le degré de redondance entre des *itemsets* de mots visuels (e.g., descripteurs SIFT(Lowe, 1999)) caractérisants des propriétés locales dans des images. La distribution q que nous recherchons est alors une distribution qui satisfait la contrainte ξ et qui minimise $D(q||p)$. Elle est obtenue par une optimisation sous contrainte décrite dans la section 5.4.2. La

valeur de $\Delta_{occ}(o)$ est alors la somme des valeurs minimales des divergences Kullback-Leibler, sur toutes les paires (p, q) associées aux variables aléatoires dans $\mathcal{V}$ qui sont impliquées dans les contraintes dérivées de l'occurrence o. Cette procédure de minimisation est également appliquée par l'algorithme 2 à la ligne 7 pour mettre à jour les distributions affectées par les contraintes impliquées par les occurrences de β^*, les distributions actuelles étant les distributions p, et les nouvelles distributions étant les distributions q.

5.4 Gain informationnel d'un motif séquentiel

5.4.1 Contraintes introduites par la connaissance des occurrences

La révélation des occurrences formées par des événements probabilistes partiels permet de contraindre les distributions des variables aléatoires représentant le contenu informationnel des événements concernés par ces occurrences. La section 5.4.1.1 répertorie les types de contraintes qui peuvent être imposées et la section 5.4.1.2 présente une étude sur la combinaison de ces contraintes.

5.4.1.1 Types de contraintes

Considérons un événement probabiliste partiel $\langle t, e, \rho \rangle$ dont le contenu informationnel peut être exprimé par l'entropie de la variable aléatoire X associée à l'événement. Soit ξ une contrainte qui est révélée et appliquée sur la variable aléatoire X. Cette contrainte peut être formalisée :

— en fournissant $\mathbb{E}_C$, l'ensemble des types d'événement candidats appartenant à $\mathbb{E}$ tels que chacun de ces candidats puisse être le type d'événement e de l'événement concerné,
— et en définissant une contrainte sur la probabilité d'observer le type d'événement e (i.e., $\Pr(X = e)$), sachant que e appartient à l'ensemble $\mathbb{E}_C$.

La contrainte ξ appartient alors à un des trois types de contraintes suivants :

Contraintes de type #1 : contraintes directes

Ce sont les contraintes les plus fortes qui se produisent lorsqu'un événement probabiliste partiel $\langle t, e, \rho \rangle$ formant une occurrence d'un motif est révélé et utilisé pour raffiner la distribution de la variable aléatoire X associée. Dans ce cas, une telle contrainte est écrite comme suit :

$$\xi = \begin{cases} \mathbb{E}_C = \{e\} \\ \Pr(X = e) = \rho \end{cases}$$

Par exemple, à la découverte d'une occurrence $o = (sid, \langle \langle t_{i_1}, e_{i_1}, \rho_{i_1} \rangle, \langle t_{i_2}, e_{i_2}, \rho_{i_2} \rangle, \ldots, \langle t_{i_m}, e_{i_m}, \rho_{i_m} \rangle \rangle)$, nous obtenons des connaissances sur le type d'événement ainsi que sa probabilité d'apparition sur chacun des événements $\langle t_{i_k}, e_{i_k}, \rho_{i_k} \rangle$, avec $1 \leq k \leq m$, dans la séquence sid. Ces connaissances sont les plus informatives que l'on puisse obtenir sur les événements probabilistes partiels concernés. C'est pour cette raison que pour chaque événement, une contrainte de type #1 est la contrainte la plus forte.

Contrainte de type #2 : contraintes fortes par propagation

Il est également possible de propager les informations fournies par des occurrences révélées pour raffiner la distribution des variables aléatoires associées aux événements qui se produisent entre les événements formant chaque occurrence. Cette propagation s'effectue en tenant compte des caractéristiques spécifiques des occurrences. Les occurrences standard, telles que définies selon la définition 5.7, n'ont pas de caractéristiques particulières permettant de propager l'information. Plus précisément, ces occurrences n'appliquent aucune contrainte sur le type d'événement ou sa probabilité d'apparition dans les événements autres que ceux qui les forment. Nous proposons donc un autre type d'occurrences, les Occurrences Minimales avec Dates Intermédiaires au plus Tôt (OMDIT) dont la définition s'appuie sur les *occurrences minimales*. Nous adaptons la définition des *occurrences minimales*, qui sont initialement proposées dans Mannila *et al.* (1997), dans le contexte des BSPP comme suit :

Définition 5.9 (Occurrence minimale). Soit $\mathcal{B}$ une base de séquences probabiliste partielle et $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \ldots \rightarrow \beta_m$ un motif séquentiel. L'occurrence $(sid, \langle\langle t_{i_1}, e_{i_1}, \rho_{i_1}\rangle, \langle t_{i_2}, e_{i_2}, \rho_{i_2}\rangle, \ldots, \langle t_{i_m}, e_{i_m}, \rho_{i_m}\rangle\rangle)$ de β est une *occurrence minimale* de β s'il n'existe aucune autre occurrence $(sid, \langle\langle t_{j_1}, e_{j_1}, \rho_{j_1}\rangle, \langle t_{j_2}, e_{j_2}, \rho_{j_2}\rangle, \ldots, \langle t_{j_m}, e_{j_m}, \rho_{j_m}\rangle\rangle)$ de β telle que $t_{j_1} > t_{i_1} \wedge t_{j_m} < t_{i_m}$ ou $t_{j_1} > t_{i_1} \wedge t_{j_m} = t_{i_m}$ ou $t_{j_1} = t_{i_1} \wedge t_{j_m} < t_{i_m}$.

Autrement dit, une occurrence d'un motif β est minimale si elle ne recouvre pas une partie de séquence qui contient une autre occurrence de β. En réutilisant l'exemple 5.2.1, les occurrences minimales du motif séquentiel $A \rightarrow A$ dans la séquence $(3, \langle\langle t_1, B, 0.9\rangle, \langle t_2, A, 0.9\rangle, \langle t_3, A, 0.7\rangle, \langle t_4, A, 0.8\rangle\rangle)$ sont $(3, \langle\langle t_2, A, 0.9\rangle, \langle t_3, A, 0.7\rangle\rangle)$ et $(3, \langle\langle t_3, A, 0.7\rangle, \langle t_4, A, 0.8\rangle\rangle)$. L'occurrence $(3, \langle\langle t_2, A, 0.9\rangle, \langle t_4, A, 0.8\rangle\rangle)$ n'est pas une occurrence minimale puisque sa fenêtre temporelle $[t_2, t_4]$ contient celle des deux premières occurrences ($[t_2, t_3]$ et $[t_3, t_4]$). Comme dans le cas des occurrences standard, les événements probabilistes partiels formant une occurrence minimale n'ont pas besoin d'être contigus. Par exemple, dans la séquence $(4, \langle\langle t_1, C, 0.8\rangle, \langle t_2, A, 0.7\rangle, \langle t_3, B, 0.8\rangle, \langle t_4, A, 0.8\rangle\rangle)$, l'occurrence $(4, \langle\langle t_2, A, 0.7\rangle, \langle t_4, A, 0.8\rangle\rangle)$ est une occurrence minimale pour le même motif.

Les OMDIT sont, quant à elles, définies comme suit :

Définition 5.10 (Occurrence minimale avec dates intermédiaires au plus tôt). Soit $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \ldots \rightarrow \beta_m$ un motif séquentiel qui couvre une séquence d'événements probabilistes s d'identifiant sid. Le terme $\langle sid, \langle\langle t_{i_1}, \beta_{i_1}, \rho_{i_1}\rangle, \langle t_{i_2}, \beta_{i_2}, \rho_{i_2}\rangle, \ldots, \langle t_{i_m}, \beta_{i_m}, \rho_{i_m}\rangle\rangle, \rho_{min}\rangle$ est une Occurrence Minimale avec dates Intermédiaires au plus Tôt (OMDIT) lorsque $(sid, \langle\langle t_{i_1}, \beta_{i_1}, \rho_{i_1}\rangle, \langle t_{i_2}, \beta_{i_2}, \rho_{i_2}\rangle, \ldots, \langle t_{i_m}, \beta_{i_m}, \rho_{i_m}\rangle\rangle)$ est une occurrence minimale et qu'il n'y a aucune autre occurrence minimale $(sid, \langle\langle t_{j_1}, \beta_{j_1}, \rho_{j_1}\rangle, \langle t_{j_2}, \beta_{j_2}, \rho_{j_2}\rangle, \ldots, \langle t_{j_m}, \beta_{j_m}, \rho_{j_m}\rangle\rangle)$ de β, avec $j_1 = i_1$ et $j_m = i_m$, pour laquelle il existe $k \in \{2, \ldots, m-1\}$ tel que $j_k < i_k$. Le dernier élément du triplet de l'OMDIT, i.e., ρ_{min}, indique la probabilité minimale observée pour les événements partiels apparaissant dans s dans la fenêtre temporelle de l'occurrence, c'est-à-dire :

$$\rho_{min} = \min \{\rho \mid \langle t, e, \rho\rangle \in s \wedge t \in [t_{i_1}, t_{i_m}]\}$$

Dans l'exemple 5.2.1, deux occurrences minimales du motif séquentiel $B \rightarrow C \rightarrow A$ peuvent être trouvées dans la séquence $(2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_3, C, 0.4\rangle, \langle t_4, A, 0.5\rangle\rangle)$: $(2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_4, A, 0.5\rangle\rangle)$ et $(2, \langle\langle t_1, B, 0.5\rangle, \langle t_3, C, 0.4\rangle, \langle t_4, A, 0.5\rangle\rangle)$. Seule la première peut être utilisée pour former une OMDIT $\langle 2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_4, A, 0.5\rangle\rangle, 0.4\rangle$, en ajoutant la probabilité minimale observée dans l'intervalle temporelle $[t_1, t_4]$. Il est important de noter qu'en considérant les OMDIT au lieu des occurrences standard (cf. Définition 5.7), les mêmes motifs séquentiels fréquents sont pris en compte puisque la mesure de

support est établie en comptant le nombre de séquences d'événements probabilistes partiels dans lesquelles ces motifs apparaissent au moins une fois.

Considérons un motif $\beta = \beta_1 \rightarrow \beta_2 \rightarrow \ldots \rightarrow \beta_m$ et une OMDIT de β : $\langle sid, \langle\langle t_{i_1}, \beta_{i_1}, \rho_{i_1}\rangle, \langle t_{i_2}, \beta_{i_2}, \rho_{i_2}\rangle, \ldots, \langle t_{i_m}, \beta_{i_m}, \rho_{i_m}\rangle\rangle, \rho_{min}\rangle$. Pour tout $j \in \{2, \ldots, m\}$, et pour chaque événement $\langle t_u, e_u, \rho_u\rangle$ de la séquence *sid* tel que $t_{i_{j-1}} < t_u < t_{i_j}$, la variable aléatoire associée à l'événement est soumise à la contrainte $\mathbb{E}_C = \mathbb{E} \setminus \{\beta_{i_j}\}$ puisque β_{i_j} ne peut pas apparaître à cet événement (d'après la définition des OMDIT). De la même manière, le type d'événement β_1 ne peut pas apparaître pour les événements aux dates t_u, avec $t_{i_1} < t_u < t_{i_2}$, et doit être retiré de l'ensemble $\mathbb{E}_C$ des variables aléatoires correspondantes, sinon l'OMDIT ne serait pas minimale. De plus, toujours par définition des OMDIT, pour chacune des variables aléatoires correspondant à un événement dont la date est dans l'intervalle $[t_{t_1}, t_{t_m}]$, il existe un type d'événement appartenant à $\mathbb{E}_C$ tel que sa probabilité est au moins égale à ρ_{min}. La réduction de l'ensemble $\mathbb{E}_C$ ainsi que cette contrainte sur les probabilités permettent de contraindre les distributions des variables aléatoires associées aux événements qui ont lieu entre les événements formant les occurrences. Cette technique sera appelée technique de raffinement de distributions par propagation.

Dans l'exemple 5.2.1, l'OMDIT $\langle 2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_4, A, 0.5\rangle\rangle, 0.4\rangle$ du motif $B \rightarrow C \rightarrow A$ impose que le type de l'événement à la date t_3 de la séquence 2 ne puisse pas être A et que sa probabilité d'apparition soit supérieure ou égale à 0.4.

La propagation sur les événements concernés par une occurrence est *forte* lorsqu'il y a seulement 2 types d'événement, i.e., $|\mathbb{E}| = 2$. En effet, dans ce cas, la réduction de l'ensemble $\mathbb{E}_C$ conduit toujours à un singleton, i.e., $\mathbb{E}_C = \{e_{restant}\}$. De plus, nous avons également la contrainte sur la probabilité d'apparition du symbole restant $\Pr(X = e_{restant}) \geq \rho_{min}$. La propagation peut aussi être *forte* lorsque $|\mathbb{E}| = 3$ et $\beta_1 \neq \beta_2$. Dans ce cas, pour chaque événement $\langle t_u, e_u, \rho_u\rangle$ de la séquence *sid* tel que $t_{i_1} < t_u < t_{i_2}$, la contrainte appliquée sur la variable associée est exprimée par $\mathbb{E}_C = \mathbb{E} \setminus \{\beta_1, \beta_2\} = \{e_{restant}\}$. Nous retrouvons par conséquent le même type de contrainte que dans le cas précédent avec $|\mathbb{E}| = 2$. Ces contraintes, obtenues à partir des propagations fortes, sont appelées contraintes de type # 2. Elles sont moins strictes que les contraintes de type # 1 et sont notées comme suit :

$$\xi = \begin{cases} \mathbb{E}_C = \{e_{restant}\} \\ \Pr(X = e_{restant}) \geq \rho_{min} \end{cases}$$

Par exemple, dans la séquence 1 de l'exemple 5.2.1, l'OMDIT $\langle 1, \langle\langle t_1, B, 0.4\rangle, \langle t_3, A, 0.7\rangle\rangle, 0.4\rangle$ du motif $B \rightarrow A$ impose que le type d'événement de l'événement probabiliste partiel à la date t_2 soit C. De plus, la probabilité d'apparition minimale de C à cet événement est de 0.4.

Enfin, nous avons un troisième type de contraintes défini comme suit.

Contrainte de type #3 : contraintes faibles par propagation

Si la propagation est effectuée sans pouvoir réduire l'ensemble $\mathbb{E}_C$ à un singleton, la contrainte résultante est de type #3 : *contrainte faible par propagation*. Une telle contrainte est plus faible qu'une contrainte de type #1 ou #2 et elle est formalisée de la façon suivante :

$$\xi = \begin{cases} \mathbb{E}_C \text{ avec } |\mathbb{E}_C| \geq 2 \\ \exists e \in \mathbb{E}_C \text{ t.q. } \Pr(X = e) \geq \rho_{min} \end{cases}$$

Par exemple, l'OMDIT $\langle 2, \langle \langle t_1, B, 0.5 \rangle, \langle t_2, C, 0.5 \rangle, \langle t_4, A, 0.5 \rangle \rangle, 0.4 \rangle$ implique la contrainte suivante sur l'événement à la date t_3 de la séquence 2 :

$$\xi = \begin{cases} \mathbb{E}_C = \{B, C\} \\ \exists e \in \mathbb{E}_C \text{ t.q. } \Pr(X = e) \geq 0.4 \end{cases}$$

5.4.1.2 Combinaisons des contraintes

L'algorithme 2 considère qu'initialement, toutes les distributions des variables aléatoires sont uniformes. À chaque itération, la sélection du motif le plus informatif β^* (ligne 4) est effectuée grâce au raffinement de ces distributions en utilisant les contraintes introduites par les OMDIT des motifs dans l'ensemble P. Comme chaque événement probabiliste partiel peut être concerné par des OMDIT de différents motifs, la variable aléatoire X associée peut être affectée par plusieurs contraintes. Une combinaison s'effectue alors entre la contrainte $\xi^{courant}$ sur une variable aléatoire X introduite par une OMDIT, et la contrainte $\xi^{précédent}$ sur X qui a été établie au cours des itérations précédentes. Les contraintes résultantes concernées par les OMDIT du motif optimal β^* seront gardées pour les itérations suivantes (où elles joueront le rôle de $\xi^{précédent}$). Finalement, la distribution des variables aléatoires concernées sera également mise à jour en utilisant ces contraintes résultantes (ligne 7). Les autres contraintes (non concernées par β^*) resteront les mêmes pour l'itération suivante.

Puisque la distribution de chaque variable aléatoire est initialement supposée uniforme [3], une contrainte imposant une distribution uniforme est construite pour chaque variable aléatoire avant la première itération. Contrairement aux autres types de contraintes, cette contrainte n'est imposée par aucune OMDIT. Elle est appelée contrainte *uniforme* ξ_U et définie par :

$$\xi_U = \begin{cases} \mathbb{E}_C = \mathbb{E} \\ \forall e_c \in \mathbb{E}_C \Rightarrow \Pr(X = e_c) = \frac{1}{|\mathbb{E}|} \end{cases}$$

Ensuite, dès la découverte de la première contrainte $\xi^{courant}$ sur une variable aléatoire X, celle-ci remplace la contrainte initiale ξ_U. En effet, toute contrainte différente d'une contrainte uniforme est préférée, car elle diffère des connaissances initiales, permettant ainsi d'avoir plus d'information sur la réalisation de X.

Dans le cas général, lorsque la contrainte $\xi^{précédent}$ n'est pas simplement la contrainte uniforme ξ_U, nous avons les deux possibilités suivantes :

1. Si une des deux contraintes, $\xi^{courant}$ ou $\xi^{précédent}$, est une contrainte de type #1, la contrainte résultante ξ est la même que la contrainte de type #1 puisque celle-ci est toujours la contrainte la plus forte que l'on peut avoir à chaque événement. D'ailleurs, lorsque deux contraintes de type #1 affectent la même variable aléatoire,

3. D'un point de vue informationnel, cette distribution est celle qui ne fournit aucune information sur la réalisation de la variable aléatoire.

elles sont identiques puisqu'elles correspondent à la découverte d'un même événement probabiliste partiel formant une OMDIT.

2. Pour les contraintes de type #2 ou #3, $\xi^{précédent}$ et $\xi^{courant}$ sont combinées de manière à obtenir la contrainte la plus forte possible (pour obtenir autant d'informations que possible). Dans ce but, l'intersection entre deux ensembles de types d'événement candidats est effectuée et la borne inférieure la plus élevée est conservée pour la probabilité. Plus formellement, puisque $\xi^{précédent}$ peut être de type # 2 ou # 3, $\xi^{précédent}$ est exprimée comme suit :

$$\xi^{précédent} = \begin{cases} \mathbb{E}_C^{précédent} \\ \exists e \in \mathbb{E}_C^{précédent} \text{ t.q. } \Pr(X = e) \geq \rho_{min}^{précédent} \end{cases}$$

La même chose s'applique pour $\xi_{courant}$:

$$\xi^{courant} = \begin{cases} \mathbb{E}_C^{courant} \\ \exists e \in \mathbb{E}_C^{courant} \text{ t.q. } \Pr(X = e) \geq \rho_{min}^{courant} \end{cases}$$

Finalement, ξ est obtenue en combinant $\xi^{précédent}$ et $\xi^{courant}$ comme suit :

$$\xi = \begin{cases} \mathbb{E}_C = \mathbb{E}_C^{précédent} \cap \mathbb{E}_C^{courant} \\ \exists e \in \mathbb{E} \text{ t.q. } \Pr(X = e) \geq \rho_{min}, \text{ avec } \rho_{min} = \max\{\rho_{min}^{précédent}, \rho_{min}^{courant}\} \end{cases}$$

# $\xi_{précédent}$	# $\xi_{courant}$	# ξ
uniforme	1	1
uniforme	2	2
uniforme	3	3
1	1, 2, 3	1
2	1	1
2	2, 3	2
3	1	1
3	2	2
3	3	2, 3

TABLE 5.2 – Combinaisons des contraintes et types de contraintes résultantes

La table 5.2 récapitule les combinaisons possibles ainsi les types de contraintes qui peuvent être obtenus. Deux observations suivantes sont à noter :

1. Combiner deux contraintes de type #3 peut produire des contraintes de type #2 grâce à l'intersection des ensembles de symboles possibles.
2. Combiner une contrainte de type #2 et une contrainte de type #3 conduit toujours à une contrainte de type #2.

5.4.2 Minimisation sous contraintes de la divergence de Kullback-Leibler

Une fois que la nouvelle contrainte ξ sur une variable aléatoire X est déduite à partir de la connaissance d'une OMDIT, ξ est utilisée pour remplacer la distribution p de X par une nouvelle distribution q. Ce raffinement de distribution apporte une quantité d'information qui

peut être mesurée par la divergence de Kullback-Leibler (KL) (Kullback et Leibler, 1951). Nous allons nous intéresser à la quantité d'information garantie suite à la connaissance de la contrainte ξ. Autrement dit, l'objectif consiste à déterminer le minimum de la divergence KL entre p et toute distribution q satisfaisant ξ. Soit I un ensemble de symboles. La divergence KL, également appelée *entropie relative*, est définie comme suit :

$$D(q||p) = \sum_{\lambda \in I} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right)$$

Par convention, $0 \log_2 \frac{0}{0} = 0$, $0 \log_2 \frac{0}{p} = 0$, et $p \log_2 \frac{q}{0} = \infty$. Cette mesure est toujours positive, et égale à 0 si et seulement si les deux distributions p et q sont identiques. Ce n'est pas une mesure de distance au sens strict puisqu'elle n'est pas symétrique et qu'elle ne vérifie pas l'inégalité triangulaire. Néanmoins, d'après Rényi et al. (1961), $D(q||p)$ quantifie *l'information obtenue* si la distribution p est remplacée par la distribution q.

La recherche de la nouvelle distribution q remplaçant la distribution p et minimisant la divergence KL sous la contrainte ξ peut être effectuée en utilisant le théorème 5.4.1.

Théorème 5.4.1. *Minimisation de la divergence KL sous des contraintes de sommes partielles des probabilités*

Soit $I = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ un ensemble de n symboles. Soit I' un sous ensemble de I. Soit X et Y dans l'intervalle $]0, 1[$. Soit p et q deux distributions sur I telles que $\forall i \in \{1, \ldots, n\}$, $p(\lambda_i) > 0 \wedge q(\lambda_i) > 0$. Si p et q satisfont les contraintes $\sum_{\lambda \in I'} p(\lambda) = X$ et $\sum_{\lambda \in I'} q(\lambda) = Y$, alors la divergence KL $D(q||p)$ est minimale si et seulement si p et q vérifient :

$$\frac{p(\lambda)}{q(\lambda)} = \begin{cases} \frac{X}{Y}, & si\ \lambda \in I' \\ \frac{1-X}{1-Y}, & si\ \lambda \in I \setminus I' \end{cases} \tag{5.2}$$

Et $D(q||p)$ vaut alors :

$$D(q||p)_{min} = \log_2 \left[\left(\frac{1-Y}{1-X} \right)^{1-Y} \times \left(\frac{Y}{X} \right)^{Y} \right] \tag{5.3}$$

Démonstration. Supposons que p et q satisfont les deux contraintes $\sum_{\lambda \in I'} p(\lambda) = X$ et $\sum_{\lambda \in I'} q(\lambda) = Y$. Reprenons la formulation de la divergence KL :

$$\begin{aligned} D(q||p) &= \sum_{\lambda \in I} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right) \\ &= \sum_{\lambda \in I'} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right) + \sum_{\lambda \in I \setminus I'} q(\lambda) \log_2 \left(\frac{q(\lambda)}{p(\lambda)} \right) \end{aligned} \tag{5.4}$$

Donc,

$$\begin{aligned} -D(q||p) &= -\sum_{\lambda\in I'} q(\lambda)\log_2\left(\frac{q(\lambda)}{p(\lambda)}\right) - \sum_{\lambda\in I\setminus I'} q(\lambda)\log_2\left(\frac{q(\lambda)}{p(\lambda)}\right) \\ &= \sum_{\lambda\in I'} q(\lambda)\log_2\left(\frac{p(\lambda)}{q(\lambda)}\right) + \sum_{\lambda\in I\setminus I'} q(\lambda)\log_2\left(\frac{p(\lambda)}{q(\lambda)}\right) \end{aligned} \tag{5.5}$$

Ensuite, si les deux côtés sont exponentiés en utilisant la base 2, nous obtenons :

$$2^{-D(q||p)} = \left(\prod_{\lambda\in I'}\left(\frac{p(\lambda)}{q(\lambda)}\right)^{q(\lambda)}\right) \times \left(\prod_{\lambda\in I\setminus I'}\left(\frac{p(\lambda)}{q(\lambda)}\right)^{q(\lambda)}\right) \tag{5.6}$$

Comme Y est dans l'intervalle $]0,1[$, cette équation peut être transformée comme suit :

$$2^{-D(q||p)} = \left(\sqrt[Y]{\prod_{\lambda\in I'}\left(\frac{p(\lambda)}{q(\lambda)}\right)^{q(\lambda)}}\right)^{Y} \times \left(\sqrt[1-Y]{\prod_{\lambda\in I\setminus I'}\left(\frac{p(\lambda)}{q(\lambda)}\right)^{q(\lambda)}}\right)^{1-Y} \tag{5.7}$$

D'après l'inégalité arithmético-géométrique pondérée (Kazarinoff, 1961), considérons les nombres non négatifs $x_1, x_2, \cdots, x_n$ et les poids non négatifs $w_1, w_2, \cdots, w_n$ tels que $w = w_1 + w_2 + \cdots + w_n > 0$, nous avons :

$$\frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w} \geq \sqrt[w]{x_1^{w_1}x_2^{w_2}\cdots x_n^{w_n}} \tag{5.8}$$

avec égalité si et seulement si $x_1 = x_2 = \cdots = x_n$.

En utilisant l'inéquation 5.8, nous obtenons, en identifiant les termes $\frac{p(\lambda)}{q(\lambda)}$ aux x_i et $q(\lambda)$ aux w_i :

$$\sqrt[Y]{\prod_{\lambda\in I'}\left(\frac{p(\lambda)}{q(\lambda)}\right)^{q(\lambda)}} \leq \frac{\sum_{\lambda\in I'} q(\lambda)\frac{p(\lambda)}{q(\lambda)}}{Y} \tag{5.9}$$

et

$$\sqrt[1-Y]{\prod_{\lambda\in I\setminus I'}\left(\frac{p(\lambda)}{q(\lambda)}\right)^{q(\lambda)}} \leq \frac{\sum_{\lambda\in I\setminus I'} q(\lambda)\frac{p(\lambda)}{q(\lambda)}}{1-Y} \tag{5.10}$$

Comme la contrainte $\sum_{\lambda\in I'} q(\lambda) = Y$ est satisfaite, et donc aussi $\sum_{\lambda\in I\setminus I'} q(\lambda) = 1 - Y$, nous avons alors :

$$\begin{aligned} 2^{-D(q||p)} &\leq \left(\frac{\sum_{\lambda\in I'} q(\lambda)\frac{p(\lambda)}{q(\lambda)}}{Y}\right)^{Y} \times \left(\frac{\sum_{\lambda\in I\setminus I'} q(\lambda)\frac{p(\lambda)}{q(\lambda)}}{1-Y}\right)^{1-Y} \\ \Leftrightarrow 2^{-D(q||p)} &\leq \left(\frac{\sum_{\lambda\in I'} p(\lambda)}{Y}\right)^{Y} \times \left(\frac{\sum_{\lambda\in I\setminus I'} p(\lambda)}{1-Y}\right)^{1-Y} \end{aligned} \tag{5.11}$$

Comme la contrainte $\sum_{\lambda \in I'} p(\lambda) = X$ est satisfaite, et alors aussi $\sum_{\lambda \in I \setminus I'} p(\lambda) = 1 - X$, cette inégalité est équivalente à :

$$\Leftrightarrow 2^{-D(q||p)} \leq \left(\frac{X}{Y}\right)^{Y} \times \left(\frac{1-X}{1-Y}\right)^{1-Y} \tag{5.12}$$

avec égalité si et seulement si tous les $\frac{p(\lambda)}{q(\lambda)}$, avec $\lambda \in I'$, sont égaux, et tous les $\frac{p(\lambda)}{q(\lambda)}$, avec $\lambda \in I \setminus I'$, sont égaux. C'est-à-dire :

$$\frac{q(\lambda)}{p(\lambda)} = \begin{cases} \frac{Y}{X}, \text{ si } \lambda \in I' \\ \frac{1-Y}{1-X}, \text{ si } \lambda \in I \setminus I' \end{cases} \tag{5.13}$$

D'après l'inégalité 5.12, nous avons :

$$D(q||p) \geq \log_2 \left[\left(\frac{Y}{X}\right)^{Y} \times \left(\frac{1-Y}{1-X}\right)^{1-Y} \right] \tag{5.14}$$

D'où la valeur minimale de $D(q||p)$:

$$D(q||p)_{min} = \log_2 \left[\left(\frac{Y}{X}\right)^{Y} \times \left(\frac{1-Y}{1-X}\right)^{1-Y} \right] \tag{5.15}$$

Cette valeur étant obtenue si et seulement si les distributions p et q satisfont l'équation 5.13. □

Dans les sections suivantes, nous allons présenter pour chaque type de contraintes la méthodologie permettant lors de l'ajout d'une contrainte de raffiner la distribution courante p en une distribution q satisfaisant la contrainte toute en minimisant la divergence KL $D(q||p)$.

5.4.2.1 Raffinement de la distribution avec une contrainte de type #1

Soit ξ une contrainte de type #1 associée à un événement probabiliste partiel $\langle t, e, \rho \rangle$. Dans ce cas, l'ensemble $\mathbb{E}_C$ contient un seul symbole e et la probabilité d'apparition de ce symbole est connue, i.e., $\mathbb{E}_C = \{e\}$, $q(e) = \Pr(e) = \rho$. En considérant $I' = \{e\}$ et $I = \mathbb{E}$ et en posant simplement $X = p(e)$ et $Y = \rho$, l'équation 5.2 permet d'obtenir les probabilités des types d'événement restant et minimisant la divergence KL entre l'ancienne distribution p et la nouvelle q.

Dans le cas où $q(e) = 1$, le théorème 5.4.1 ne s'applique pas. La probabilité des autres types d'événement est alors simplement fixée à 0 et la divergence KL est calculée en se servant directement de la définition de $D(q||p)$ et des conventions associées.

5.4.2.2 Raffinement de la distribution avec une contrainte de type #2

Soit ξ une contrainte de type #2 obtenue par la propagation telle que $\mathbb{E}_C = \{e_{restant}\}$ et $\Pr(e_{restant}) \geq \rho_{min}$. Pour obtenir la distribution q qui minimise la divergence KL, nous fixons

$q(e_{restant}) = \rho_{min}$. La probabilité des autres symboles est ensuite déterminée en utilisant la stratégie employée pour les contraintes de type #1 (Section 5.4.2.1).

Fixer $q(e_{restant})$ à ρ_{min} se justifie de la façon suivante. Considérons $I' = \{e_{restant}\}$ tel que $I' \subset I$ (avec $I = \mathbb{E}$). Posons $X = \sum_{\lambda \in I'} p(\lambda)) = p(e_{restant})$ et $Y = \sum_{\lambda \in I'} q(\lambda) = q(e_{restant})$ avec X et Y compris dans l'intervalle $]0, 1[$ [4]. Puisqu'avec une contrainte de type #2 combinant deux contraintes, nous avons toujours une augmentation (ou au moins la même valeur) de la borne inférieure ρ_{min} pour la probabilité du symbole $e_{restant}$. Ceci implique que $Y \geq X$. Dans ces conditions, calculons $\Theta(Y)$, la valeur minimale de la divergence KL exprimée en fonction de Y :

$$\begin{aligned}\Theta(Y) &= \log_2 \left[\left(\frac{1-Y}{1-X} \right)^{1-Y} \times \left(\frac{Y}{X} \right)^{Y} \right] \\ &= (1-Y) \log_2 \left(\frac{1-Y}{1-X} \right) + Y \log_2 \frac{Y}{X} \end{aligned} \tag{5.16}$$

Sa dérivée est exprimée comme suit :

$$\begin{aligned}\frac{d\Theta(Y)}{dY} &= \log_2 \left(\frac{1-X}{1-Y} \right) - (1-Y) \left(\frac{1}{(1-Y) ln 2} \right) \\ &\quad + \log_2 \frac{Y}{X} + Y \frac{1}{Y \ln 2} \\ &= \log_2 \left(\frac{1-X}{1-Y} \right) - \frac{1}{ln 2} + \log_2 \frac{Y}{X} + \frac{1}{\ln 2} \\ &= \log_2 \left(\frac{1-X}{1-Y} \times \frac{Y}{X} \right) \end{aligned} \tag{5.17}$$

Nous avons donc :

$$\frac{d\Theta(Y)}{dY} \begin{cases} < 0, & \text{lorsque } Y \in]0, X[\\ = 0, & \text{lorsque } Y = X \\ > 0, & \text{lorsque } Y \in]X, 1[\end{cases} \tag{5.18}$$

Comme $Y \geq X$ et $Y = \sum_{\lambda \in I'} q(\lambda) = q(e_{restant})$, la valeur minimale de la divergence KL est obtenue lorsque $q(e_{restant})$ est fixé à sa valeur la plus petite possible, tout en satisfaisant la contrainte ξ, i.e., $q(e_{restant}) \geq \rho_{min}$. Autrement dit, $q(e_{restant})$ est fixé à ρ_{min}.

5.4.2.3 Raffinement de la distribution avec une contrainte de type #3

Soit ξ une contrainte de type #3 telle que :

$$\xi = \begin{cases} \mathbb{E}_C \subset \mathbb{E} \text{ avec } |\mathbb{E}_C| \geq 2 \\ \exists e \in \mathbb{E}_C \text{ t.q. } \Pr(e) \geq \rho_{min} \end{cases}$$

En utilisant les mêmes approches que pour les contraintes de types #1 et #2, nous ne pouvons pas obtenir de façon déterministe la nouvelle distribution q qui minimise la divergence

4. Les valeurs extrêmes (0 et 1) peuvent être traitées séparément de façon similaire au cas particulier mentionné dans la section 5.4.2.1.

KL à partir de cette contrainte, car nous ne savons pas sur quel élément de $\mathbb{E}_C$ s'applique $\Pr(e) \geq \rho_{min}$. Par conséquent, comme nous cherchons à établir le gain d'information minimal, une version *relaxée* de cette contrainte est considérée. Elle est notée ξ' et définie comme suit :

$$\xi' = \begin{cases} \mathbb{E}_C \subset \mathbb{E}, \text{ avec } |\mathbb{E}_C| \geq 2 \\ \sum_{e \in \mathbb{E} \setminus \mathbb{E}_C} \Pr(e) \leq 1 - \rho_{min} \end{cases}$$

Comme la probabilité est non négative, nous pouvons constater qu'une distribution qui satisfait ξ satisfait également ξ'. L'inverse n'est pas vrai. Cette contrainte est donc moins forte que l'originale et est appelée contrainte de type #3'.

Par exemple, considérons $\mathbb{E} = \{A, B, C, D\}$ et la contrainte de type #3 suivante :

$$\xi = \begin{cases} \mathbb{E}_C = \{A, B\} \\ \exists e \in \mathbb{E}_C \text{ t.q. } \Pr(e) \geq 0.4 \end{cases}$$

Cette contrainte ξ précise que l'ensemble de candidats contient deux symboles A et B, et que la probabilité d'apparition du symbole associé à l'événement probabiliste partiel correspondant [5] est au moins égale à 0.4. La contrainte relaxée de ξ sera :

$$\xi' = \begin{cases} \mathbb{E}_C = \{A, B\} \\ \sum_{e \in \{C,D\}} \Pr(e) \leq 0.6 \end{cases}$$

Toute distribution satisfaisant ξ, e.g., $\Pr(A) = 0.5, \Pr(B) = 0.2, \Pr(C) = 0.1, \Pr(D) = 0.2$, satisfait également à ξ'. L'inverse n'est pas vrai. Par exemple, la distribution avec les probabilités suivantes $\Pr(A) = 0.2, \Pr(B) = 0.3, \Pr(C) = 0.2, \Pr(D) = 0.3$ satisfait ξ' mais ne satisfait pas ξ.

Une fois que la nouvelle contrainte relaxée ξ' est définie, la nouvelle distribution qui satisfait ξ' et qui minimise la divergence KL peut être ensuite calculée. Plus précisément, considérons $I = \mathbb{E}$ l'ensemble des types d'événement et $I' = \mathbb{E}_{rejeté} = \mathbb{E} \setminus \mathbb{E}_C$ l'ensemble des types d'événement *rejetés*. Posons $X = \sum_{e \in I'} p(e)$ et $Y = \sum_{e \in I'} q(e)$. La contrainte de type #3′ implique alors $Y \leq 1 - \rho_{min}$. D'après la dérivée de la fonction exprimant la valeur minimale de la divergence KL par rapport à Y (cf. Section 5.4.2.2), la quantité minimale d'information est obtenue lorsque la valeur de Y est fixée selon les deux cas suivants :

— Si $1 - \rho_{min} \geq X$, nous pouvons choisir $Y = X$ pour atteindre la valeur minimale de $\Theta(Y)$ tout en satisfaisant $Y \leq 1 - \rho_{min}$.
— Si au contraire $1 - \rho_{min} < X$, la plus petite valeur de $\Theta(Y)$ pouvont être atteinte tout en satisfaisant $Y \leq 1 - \rho_{min}$ est celle obtenue en prenant le plus grand Y possible, c nous fixons $Y = 1 - \rho_{min}$, et la divergence KL est obtenue avec l'équation 5.16.

La probabilité de chacun des types d'événement dans la nouvelle distribution peut être ensuite calculée en utilisant le théorème 5.4.1.

5. Ce symbole est parmi les deux symboles candidats.

5.4.3 Exemple

Cette section vise à donner un exemple de déroulé de la méthode proposée. Il reprend les données dans l'exemple 5.2.1. Supposons que nous voulons sélectionner les deux motifs les plus informatifs de l'ensemble des motifs séquentiels clos, $\{A \to A, B \to A, C \to A, B \to C \to A, B \to C \to C\}$. Considérons l'ensemble des variables aléatoires X_j^i tel que $i \in \{1,2,3,4\}$ représente l'identifiant de séquence et que $j \in \{1,2,3,4\}$ fait référence à l'identifiant dans cette séquence du $j^{ième}$ événement probabiliste partiel. Initialement, toutes les variables aléatoires sont supposées uniformes. Le gain d'information de chaque motif est établi en affinant la distribution de ces variables, en minimisant la divergence KL sous les contraintes imposées par ses OMDIT. Ce calcul du gain d'information est détaillé ci-après pour chacun des motifs clos.

Les OMDIT du motif $A \to A$ sont les suivantes : $\langle 3, \langle\langle t_2, A, 0.9\rangle, \langle t_3, A, 0.7\rangle\rangle, 0.7\rangle$, $\langle 3, \langle\langle t_3, A, 0.7\rangle, \langle t_4, A, 0.8\rangle\rangle, 0.7\rangle$, $\langle 4, \langle\langle t_2, A, 0.7\rangle, \langle t_4, A, 0.7\rangle\rangle, 0.7\rangle$. À partir de la connaissance de ces occurrences, nous pouvons commencer à raffiner les distributions concernées pour obtenir le gain d'information. Il est à noter que pour les contraintes de type #3, seule leur version relaxée est affichée.

$\Delta(A \to A) = 2.230653311/2 = 1.1153266555$					
sid	X_j^i	contrainte (type : détails)	p	q	$D(q\|\|p)$
3	X_2^3	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.9, 0.05, 0.05$	1.015966907
	X_3^3	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.419638508
	X_3^3	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
	X_4^3	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.8$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.8, 0.1, 0.1$	0.663034406
				$\Delta_{occ} = \sum D(q\|\|p)$	1.066706007
				non comptée	non comptée
4	X_2^4	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
	X_3^4	$\#3' : \mathbb{E}_C = \{B, C\}, Pr(A) \leq 0.3$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.35, 0.35$	0.003671601
	X_4^4	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	0.811014803
				$\sum \Delta_{occ}$	2.230653311

Puisque le motif $A \to A$ apparaît deux fois dans la séquence 3, seule l'occurrence la plus informative est choisie, c'est-à-dire celle représentée par X_2^3 et X_3^3. La quantité d'information apportée par l'autre occurrence n'est donc pas comptabilisée et les contraintes associées ne sont pas sauvegardées.

De la même manière, les OMDIT du motif $B \to A$ sont : $\langle 1, \langle\langle t_1, B, 0.4\rangle, \langle t_3, A, 0.7\rangle\rangle, 0.4\rangle$, $\langle 2, \langle\langle t_1, B, 0.5\rangle, \langle t_4, A, 0.5\rangle\rangle, 0.4\rangle$, $\langle 3, \langle\langle t_1, B, 0.9\rangle, \langle t_2, A, 0.9\rangle\rangle, 0.9\rangle$, $\langle 4, \langle\langle t_3, B, 0.8\rangle, \langle t_4, A, 0.7\rangle\rangle, 0.7\rangle$. Nous avons donc :

$\Delta(B \to A) = 3.728284049/4 = 0.93207101225$					
sid	X_j^i	contrainte (type : détails)	p	q	$D(q\|\|p)$
1	X_1^1	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.4, 0.3$	0.014011906
	X_2^1	$\#2 : \mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_3^1	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	0.431695414
2	X_1^2	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.5, 0.25$	0.084962501
	X_2^2	$\#2 : \mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_3^2	$\#2 : \mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_4^2	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.197948814
3	X_1^3	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.9, 0.05$	1.015966907
	X_2^3	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.9, 0.05, 0.05$	1.015966907
				$\Delta_{occ} = \sum D(q\|\|p)$	2.031933814
4	X_3^4	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.8$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.1, 0.8, 0.2$	0.663034406
	X_4^4	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.066706007
				$\sum \Delta_{occ}$	3.728284049

Ensuite, les OMDIT du motif $C \to A$ sont : $\langle 1, \langle\langle t_2, C, 0.9\rangle, \langle t_3, A, 0.7\rangle\rangle, 0.7\rangle, \langle 2, \langle\langle t_3, C, 0.4\rangle, \langle t_4, A, 0.5\rangle\rangle, 0.4\rangle, \langle 4, \langle\langle t_1, C, 0.8\rangle, \langle t_2, A, 0.7\rangle\rangle, 0.7\rangle$.

$\Delta(C \to A) = 2.585318922/3 = 0.861772974$					
sid	X_j^i	contrainte (type : détails)	p	q	$D(q\|\|p)$
1	X_2^1	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.05, 0.9$	$1, 015966907$
	X_3^1	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.419638508
2	X_3^2	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_4^2	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.098974407
4	X_1^4	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.8$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.1, 0.1, 0.8$	0.663034406
	X_2^4	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.066706007
				$\sum \Delta_{occ}$	2.585318922

Les OMDIT du motif $B \to C \to A$ sont : $\langle 1, \langle\langle t_1, B, 0.4\rangle, \langle t_2, C, 0.9\rangle, \langle t_3, A, 0.7\rangle\rangle, 0.4\rangle, \langle 2, \langle\langle t_1, B, 0.5\rangle, \langle t_2, C, 0.5\rangle, \langle t_4, A, 0.5\rangle\rangle, 0.4\rangle$.

$\Delta(B \to C \to A) = 1.688537918/2 = 0.844268959$					
sid	X_j^i	contrainte (type : détails)	p	q	$D(q\|\|p)$
1	X_1^1	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.4, 0.3$	0.014011906
	X_2^1	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.05, 0.9$	1.015966907
	X_3^1	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.433650415
2	X_1^2	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.5, 0.25$	0.084962501
	X_2^2	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.25, 0.5$	0.084962501
	X_3^2	$\#3' : \mathbb{E}_C = \{B, C\}, Pr(A) \leq 0.6$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	0
	X_4^2	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.254887503
				$\sum \Delta_{occ}$	1.688537918

Et finalement, les OMDIT du motif $B \to C \to C$ sont : $\langle 1, \langle \langle t_1, B, 0.4 \rangle, \langle t_2, C, 0.9 \rangle, \langle t_4, C, 0.5 \rangle \rangle, 0.4 \rangle, \langle 2, \langle \langle t_1, B, 0.5 \rangle, \langle t_2, C, 0.5 \rangle, \langle t_3, C, 0.4 \rangle \ \rangle, 0.4 \rangle$.

$\Delta(B \to C \to C) = 1.298878222/2 = 0.649439111$					
sid	X_j^i	contrainte (type : détails)	p	q	$D(q\|\|p)$
1	X_1^1	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.4, 0.3$	0.014011906
	X_2^1	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.05, 0.9$	1.015966907
	X_3^1	$\#3' : \mathbb{E}_C = \{A, B\}, Pr(C) \leq 0.6$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	0
	X_4^1	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.25, 0.5$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	1.114941314
2	X_1^2	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.5, 0.25$	0.084962501
	X_2^2	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.25, 0.5$	0.084962501
	X_3^2	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
				$\Delta_{occ} = \sum D(q\|\|p)$	0.183936908
				$\sum \Delta_{occ}$	1.298878222

Par conséquent, le premier motif à sélectionner est $A \to A$ puisque le gain d'information apporté par les occurrences de ce motif est le plus grand parmi l'ensemble des motifs clos. Les distributions des variables aléatoires qui sont affectées par les contraintes imposées par ses occurrences sont remplacées par les nouvelles distributions calculées et sélectionnées lors de l'estimation du gain d'information. De plus, les contraintes associées à ces distributions sont également sauvegardées pour être combinées avec les contraintes des motifs restants lors des itérations suivantes.

À la deuxième itération, les motifs restants sont réévalués à nouveau pour prendre en compte les modifications au niveau des distributions et des contraintes apportées par les occurrences sélectionnées du motif $A \to A$. Les tableaux suivants contiennent, pour les événements probabilistes concernés, la contrainte nouvellement découverte ainsi que la contrainte obtenue par $A \to A$ (marquée par *). Les combinaisons de contraintes sont effectuées avec la méthode proposée dans la section 5.4.1.2, en fonction de leur type. Pour ce jeu de données, nous allons observer que les contraintes résultantes seront identiques à celles qui viennent d'être découvertes par les nouveaux motifs.

Concrètement, pour le motif $B \to A$, nous avons les résultats suivants :

$\Delta(B \to A) = 2.259639254/4 = 0.5649098135$					
sid	X_j^i	contrainte (type : détails)	p	q	$D(q\|\|p)$
1	X_1^1	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.4, 0.3$	0.014011906
	X_2^1	$\#2 : \mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_3^1	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	$0, 403671601$
				$\Delta_{occ} = \sum D(q\|\|p)$	0.431695414
2	X_1^2	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.25, 0.5, 0.25$	0.084962501
	X_2^2	$\#2 : \mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_3^2	$\#2 : \mathbb{E}_C = \{C\}, Pr(C) \geq 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_4^2	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.197948814
3	X_1^3	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.9, 0.05$	1.015966907
	X_2^3	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.9$	$0.9, 0.05, 0.05$	$0.9, 0.05, 0.05$	0
		$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.9^*$			
				$\Delta_{occ} = \sum D(q\|\|p)$	1.015966907
4	X_3^4	$\#1 : \mathbb{E}_C = \{B\}, Pr(B) = 0.8$	$0.3, 0.35, 0.35$	$\frac{6}{65}, 0.8, \frac{7}{65}$	0.614028119
		$\#3' : \mathbb{E}_C = \{B, C\}, Pr(A) \leq 0.3^*$			
	X_4^4	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$0.7, 0.15, 0.15$	$0.7, 0.15, 0.15$	0
		$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7^*$			
				$\Delta_{occ} = \sum D(q\|\|p)$	0.614028119
				$\sum \Delta_{occ}$	2.259639254

Et pour le motif $C \to A$, nous avons :

$\Delta(C \to A) = 2.181647322/3 = 0.727215774$					
sid	X_j^i	contrainte (type : détails)	p	q	$D(q\|\|p)$
1	X_2^1	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.9$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.05, 0.05, 0.9$	$1, 015966907$
	X_3^1	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.7, 0.15, 0.15$	0.403671601
				$\Delta_{occ} = \sum D(q\|\|p)$	1.419638508
2	X_3^2	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.4$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.3, 0.3, 0.4$	0.014011906
	X_4^2	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.5$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.5, 0.25, 0.25$	0.084962501
				$\Delta_{occ} = \sum D(q\|\|p)$	0.098974407
4	X_1^4	$\#1 : \mathbb{E}_C = \{C\}, Pr(C) = 0.8$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	$0.1, 0.1, 0.8$	0.663034406
	X_2^4	$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7$	$0.7, 0.15, 0.15$	$0.7, 0.15, 0.15$	0
		$\#1 : \mathbb{E}_C = \{A\}, Pr(A) = 0.7^*$			
				$\Delta_{occ} = \sum D(q\|\|p)$	0,663034406
				$\sum \Delta_{occ}$	2.181647322

Le gain d'information du motif $B \to C \to A$ reste identique ($\Delta(B \to C \to A) = 0.844268959$) puisqu'il se produit dans des séquences d'événements probabilistes partiels où $A \to A$ n'apparaît pas. C'est le même cas pour le motif $B \to C \to C$, avec $\Delta(B \to C \to C) = 0.649439111$.

En comparant les gains d'information à l'issue de cette itération, le motif $B \to C \to A$ est sélectionné. Ainsi, l'ensemble des motifs sélectionnés à l'issue de 2 itérations est : $\{A \to A, B \to C \to A\}$, qui est également l'ensemble final puisque 2 motifs sont demandés. Nous pouvons observer que la méthode sélectionne des motifs qui sont complémentaires puisqu'ils ont tendance à apporter des informations sur des événements probabilistes partiels distincts.

5.5 Expériences

Nous allons maintenant appliquer la méthode proposée sur des données réelles. Deux bases de séquences d'événements probabilistes partiels, construites à partir de deux STCD utilisées dans le chapitre précédent : une sur le Groenland provenant de données optiques et une sur le massif du Mont-Blanc provenant de données radar, sont examinées. Pour chacun de ces jeux de données, nous allons présenter comment les indices de confiance sont transformés en probabilités, les paramètres utilisés ainsi que les résultats qualitatifs et quantitatifs.

5.5.1 STCD sur le Groenland provenant de données optiques

5.5.1.1 Préparation des données

Pour cette expérience, la base de séquences probabiliste partielle est construite à partir de la STCD utilisée dans la section 4.5.1. Pour rappel, ce jeu de données contient 20 champs de déplacements de taille de 458×500 pixels, couvrant notamment 4 glaciers (Nordenskjöld, Polonia, Sarqardliup Sermia, et Alangordliup Sermia). Chaque déplacement est exprimé par la vitesse médiane différentielle (en anglais *Median Differential Speed*) (MDS) qui représente, en chaque point et chaque date d'observation, la variation de vitesse par rapport à la médiane des vitesses observées au même point tout au long de la période d'observation. L'indice de confiance est, quant à lui, exprimé par la cohérence des vecteurs de déplacements (en anglais *Velocity Vector Coherence*) (VVC), qui favorise les vecteurs de déplacements dont le sens suit celui du vecteur somme des vitesses observées au même point pour les différentes observations. Au niveau de la base de séquences, nous avons au total 1649435 mesures de déplacement, avec une valeur de confiance minimale, maximale et moyenne qui est respectivement de 0.004, 1.0, et 0.965. Ces valeurs montrent qu'en moyenne, les valeurs de confiance sont très élevées. En effet, l'histogramme illustré sur la figure 5.1 montre clairement que la plupart des valeurs de confiance se trouvent près de 1.

Cette mesure VVC, qui exprime le niveau de confiance que l'on peut avoir sur les vecteurs de déplacement estimés, peut être utilisée dans le contexte des BSPP pour construire la probabilité des événements probabilistes partiels. Comme la valeur de VVC est comprise dans l'intervalle $]0, 1]$, et que la probabilité ρ associée à chaque événement $\langle t, e, \rho \rangle$ d'une BSPP doit être dans l'intervalle $]\frac{1}{|\mathbb{E}|}, 1]$, une transformation linaire est effectuée pour obtenir la probabilité des événements à partir de la VVC :

$$\rho = VVC \times \frac{|\mathbb{E}| - 1}{|\mathbb{E}|} + \frac{1}{|\mathbb{E}|}$$

où $|\mathbb{E}|$ représente le nombre de symboles dans l'ensemble des types d'événement $\mathbb{E}$. Nous postulons ainsi que les symboles présents dans la STCD sont les symboles ayant le plus de chances d'être observés.

Comme dans l'expérience de la section 4.5.1, nous utilisons 3 types d'événement, $\{1, 2, 3\}$, pour discrétiser les mesures de déplacement, où les symboles 1, 2, 3 représentent respectivement des valeurs de déplacement faibles, proches de la valeur médiane, et élevées. Ainsi, nous avons :

$$\rho = VVC \times \frac{2}{3} + \frac{1}{3}$$

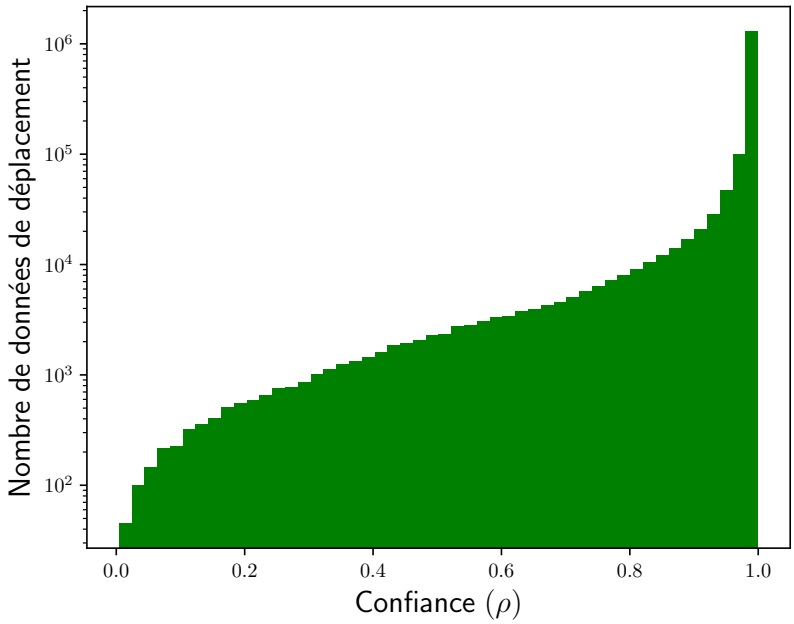


FIGURE 5.1 – Histogramme des indices de confiance (Groenland)

5.5.1.2 Paramètres utilisés

Dans cette expérience, nous allons construire l'ensemble des motifs les plus informatifs en nous basant sur les motifs SFG maximaux, qui représentent, pour rappel, les évolutions de déplacements les plus spécifiques parmi l'ensemble des motifs extraits. Ces motifs maximaux sont obtenus avec les réglages standard de support et de connexité présentés dans la section 4.5.1.2. Ainsi, le seuil de connexité κ est fixé à 5, et le seuil de support minimal à 17175 (7.5% du nombre de pixels de chaque champ de déplacement). Après avoir construit la BSPP et extrait l'ensemble des motifs maximaux, nous souhaitons sélectionner un ensemble contenant 20 motifs qui sont complémentairement informatifs.

5.5.1.3 Résultats quantitatifs

Cette expérience est faite sur un prototype dédié, écrit en C++. L'algorithme implémenté est basé sur les OMDIT. Dans ce cas, une borne supérieure de la complexité dans le pire des cas est $\mathcal{O}(|P| \times k \times |\mathcal{B}| \times N^2)$, avec $|P|$ le nombre de motifs dans l'ensemble des motifs à classer, k le nombre de motifs à sélectionner, $|\mathcal{B}|$ le nombre de séquences de $\mathcal{B}$ et N le nombre d'événements d'une séquence, chaque séquence étant supposée contenir le même nombre d'événements. L'étude de cette borne est disponible dans l'annexe B.

Le calcul du gain informationnel des motifs à chaque itération est également parallélisé sur tous les cœurs disponibles du processeur. Ainsi, la sélection les 20 motifs informatifs a pris environ 206 secondes, avec l'utilisation d'un processeur Intel Xeon 3.5 GHz (16 cœurs) sous Linux (Ubuntu).

La liste des motifs sélectionnés ainsi que le gain d'information et le support associés est présentée dans la table 5.3.

#	Motif séquentiel informatif	Gain d'information	Support
1	3→3→3→3→2→2→2→1→1→1→1	17.6718	17645
2	2→3→2→2→2→2→2→2→1→1	13.5495	18370
3	3→1→3→3→1→3→1→1→1	11.8934	18252
4	3→1→3→3→2→2→2→2→1	8.60465	25050
5	3→3→3→3→1→1→1→1→1→1	6.87313	19802
6	2→2→3→3→2→1→2→1	5.76651	28184
7	1→1→3→3→3→1	4.26914	22208
8	3→3→1→2→2→1→2→1	3.6746	32716
9	3→3→3→3→3→3→3→2→1	2.89112	19592
10	2→2→2→2→2→3	2.74065	19369
11	2→1→3→1→1→1→1	2.39764	32558
12	3→3→2→2→2→2→2→2→2	2.1528	17468
13	1→3→2→2→3→1	1.53025	31203
14	2→3→3→3→2→2→1→1→1	1.38927	32879
15	3→1→1→1→3	1.20914	20491
16	2→1→1→2→2→1	1.11467	33553
17	3→3→1→3→3→2→1→1→1	0.921699	29130
18	2→2→2→2→2→2→2→2→1	0.879896	18781
19	2→3→2→2→1→1→1→1→1	0.839208	20158
20	1→3→3→3→3→3→1→1	0.659549	20809

TABLE 5.3 – Liste des motifs informatifs sélectionnés (Groenland)

Nous pouvons constater que le support des motifs n'a pas d'impact sur l'ordre de la sélection des motifs, puisqu'un motif est considéré informatif lorsqu'en moyenne ses occurrences apportent un gain d'information important. Nous pouvons également observer que le gain d'information diminue fortement après chaque sélection, notamment lors des premières itérations (cf. Figure 5.2).

Afin d'estimer la complémentarité spatiale entre les motifs sélectionnés, nous proposons la mesure suivante :

$$\theta = \frac{S_\Phi}{\sum_{\beta \in \Phi} support_\beta} \tag{5.19}$$

où S_Φ indique le nombre de séquences couvertes par au moins un motif parmi l'ensemble des motifs sélectionnés Φ. La valeur de θ est comprise dans l'intervalle $[\frac{1}{|\Phi|}, 1]$, elle vaut $\frac{1}{|\Phi|}$ lorsque tous les motifs sélectionnés ont la même couverture, et 1 lorsque les couvertures ne se chevauchent pas. Plus θ est proche de 1, plus les motifs sont complémentaires spatialement et inversement.

Pour ce jeu de données, l'ensemble des 20 motifs les plus informatifs conduit à une valeur de θ à 0.192. Cette mesure, axée sur la couverture spatiale, a l'avantage de pouvoir être également utilisée pour quantifier la complémentarité des motifs sélectionnés par la *swap* randomisation. Pour les mêmes données fournies en entrée, en sélectionnant les 20 motifs avec la méthode de *swap* randomisation (10 motifs à chaque extrémité du classement), nous

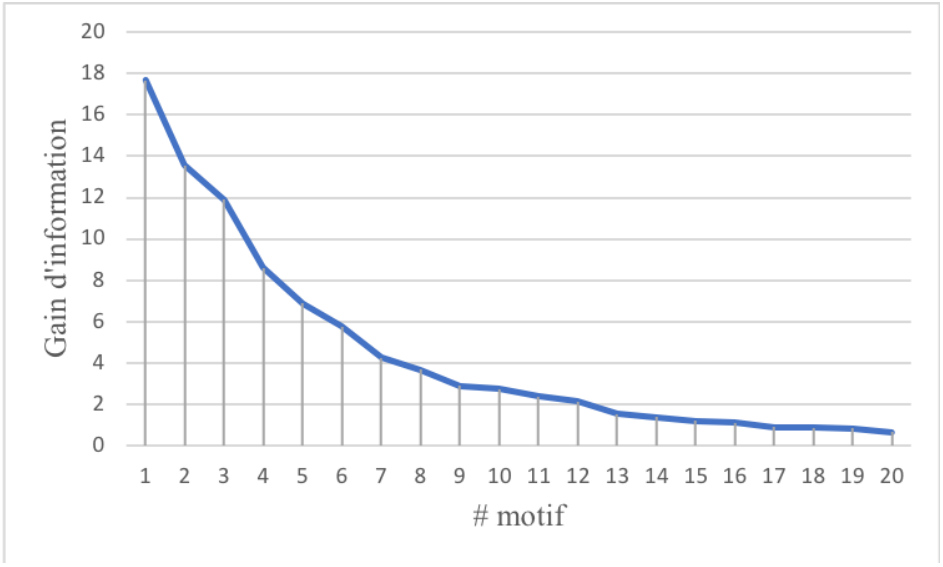


FIGURE 5.2 – Le gain d'information des motifs sélectionnés (Groenland)

obtenons une valeur de θ seulement à 0.1429. Ainsi, au niveau spatial, les motifs sélectionnés par notre proposition sont en moyenne plus complémentaires que ceux sélectionnés par la méthode proposée par Méger *et al.* (2015).

Au niveau temporel, nous ne pouvons pas comparer notre méthode avec la *swap* randomisation. En effet, cette dernière utilise la date de fin au plus tôt de la première occurrence sans qu'aucun type d'occurrence ne soit imposé (minimale ou pas, dates d'occurrence intermédiaire au plus tôt ou pas, etc.). Quant à notre méthode, comme les occurrences de type OMDIT sont utilisées et que pour chaque motif, seule la meilleure occurrence dans chaque séquence qu'il couvre est gardée pour évaluer le gain informationnel, il est possible de définir la complémentarité spatiotemporelle comme suit :

$$\tau = \frac{evt_{\Phi}}{\sum_{\beta \in \Phi} evt_{\beta}} \tag{5.20}$$

où evt_{Φ} indique le nombre d'événements affectés par au moins une contrainte introduite par les occurrences des motifs dans Φ, et evt_{β} correspond au nombre d'événements affectés par des contraintes introduites par les occurrences du motif β. La valeur de τ est comprise dans l'intervalle $[\frac{1}{|\Phi|}, 1]$, à $\frac{1}{|\Phi|}$ lorsque les mêmes événements sont concernés pour tous les motifs dans Φ, et à 1 lorsque ces événements diffèrent complètement d'un motif à un autre. Il est à noter que cette mesure peut également être utilisée pour évaluer la complémentarité de motifs sélectionnés à partir de n'importe quelle BSPP, qu'elle comporte une dimension temporelle ou non.

Pour ce jeu de données, nous avons une valeur de τ à 0.235.

5.5.1.4 Résultats qualitatifs

D'un point de vue qualitatif, les motifs sélectionnés présentent des évolutions de déplacements intéressantes sur différentes régions. Par exemple, les trois premiers motifs sélectionnés sont $3 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1$, $2 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to$

$2 \to 1 \to 1$, et $3 \to 1 \to 3 \to 3 \to 1 \to 3 \to 1 \to 1 \to 1$. Afin de mieux localiser à la fois dans l'espace et dans le temps des occurrences qui ont contribué au gain informationnel de chaque motif séquentiel, nous utilisons une variation des cartes de Localisation Spatio-Temporelle (LST) qui indique la date de fin de la meilleure OMDIT à chaque séquence au lieu de celle de la première occurrence (telles qu'elles sont définies dans la section 4.3). Les cartes LST des trois premiers motifs sélectionnés, présentées dans les figures 5.4, 5.5 et 5.6, montrent une bonne complémentarité spatiale. L'échelle de couleur de ces cartes LST est présentée dans la figure 5.3. Nous retrouvons le motif correspondant à la décélération régionale signalée dans Tedstone *et al.* (2015), $3 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1$, en première position du classement. D'après la figure 5.4, cette décélération a eu lieu au milieu, le long du glacier Nordenskjöld, au nord, au niveau du glacier Alangordliup Sermia, et au sud, dans la région du glacier Polonia. Au niveau temporel, nous pouvons constater que cette évolution de déplacements a duré en général jusqu'à la fin de la période d'étude, entre 2011 et 2013. Cela correspond bien à l'étude de Tedstone *et al.* (2015). Le deuxième motif sélectionné, $2 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1 \to 1$, représente une longue période de déplacements assez stables $2 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2$, suivie d'une décélération $1 \to 1$. Cette évolution est présente en haut des glaciers Sarqardliup Sermia, Alangordliup Sermia, Polonia et aussi au milieu du glacier Nordenskjöld, avec une forte complémentarité spatiale par rapport aux pixels couverts par le premier motif (cf. Figure 5.5). À l'échelle temporelle, globalement, cette évolution de déplacements prend fin plus tôt (entre 2008 et 2010) au niveau du glacier Nordenskjöld que sur les autres glaciers (entre 2010 et 2013). Le troisième motif, $3 \to 1 \to 3 \to 3 \to 1 \to 3 \to 1 \to 1 \to 1$, représente des fluctuations, entre les déplacements forts et les déplacements faibles, que nous pouvons observer sur les bords de l'ensemble des glaciers et aussi sur des régions plus élevées (cf. Figure 5.6). Cette évolution est très peu présente dans les parties centrales des principaux glaciers, montrant ainsi que le déplacement à ces endroits est plus stable. De plus, ce motif, se terminant plus tôt aux bords des glaciers (entre 2008 et 2010) que les autres zones (entre 2011 et 2013), montre également que le déplacement sur les bords est très instable [6].

En considérant l'ensemble de ces trois premiers motifs, nous obtenons des valeurs très élevées pour les mesures de complémentarité spatiale et spatiotemporelle, qui sont respectivement : $\theta = 0.838$ et $\tau = 0.847$. Cette forte complémentarité pour les premiers motifs sélectionnés montre que notre méthode est capable de sélectionner les évolutions de déplacements ayant lieu sur des zones (spatiales et spatiotemporelles) très différentes dans la base de séquences.

Nous retrouvons également le motif contenant une longue période stable à une vitesse élevée suivie d'une décélération soudaine, $3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 1$, présenté dans la section 4.5.1.4 (cf. Figure 5.7). Ce motif souligne principalement les glaciers Polonia, Nordenskjöld et Alangordliup Sermia. Au niveau du glacier Sarqardliup Sermia, sa présence est anecdotique. En effet, le comportement de ce dernier est très différent de celui des autres glaciers. Par exemple, le motif $3 \to 1 \to 1 \to 1 \to 3$, dont la carte LST est présentée dans la figure 5.8, souligne deux courtes périodes de déplacements forts, séparées par une période significative de déplacements faibles en aval du glacier Sarqardliup Sermia. D'après la figure 5.8, cette évolution y prend fin entre 2006 et 2009. En regardant les dates de début des occurrences minimales concernées par cette évolution (cf. Figure 5.9), nous observons que cette évolution débute pendant la période entre 2001 et 2002.

6. Puisqu'il faut aller plus loin dans les séquences dans les parties plus hautes des glaciers pour trouver des occurrences de ces fluctuations.

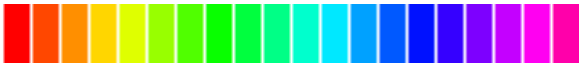


FIGURE 5.3 – Échelle de couleur des cartes LST : depuis 1985 en rouge jusqu'à 2013 en magenta décomposée en 20 périodes

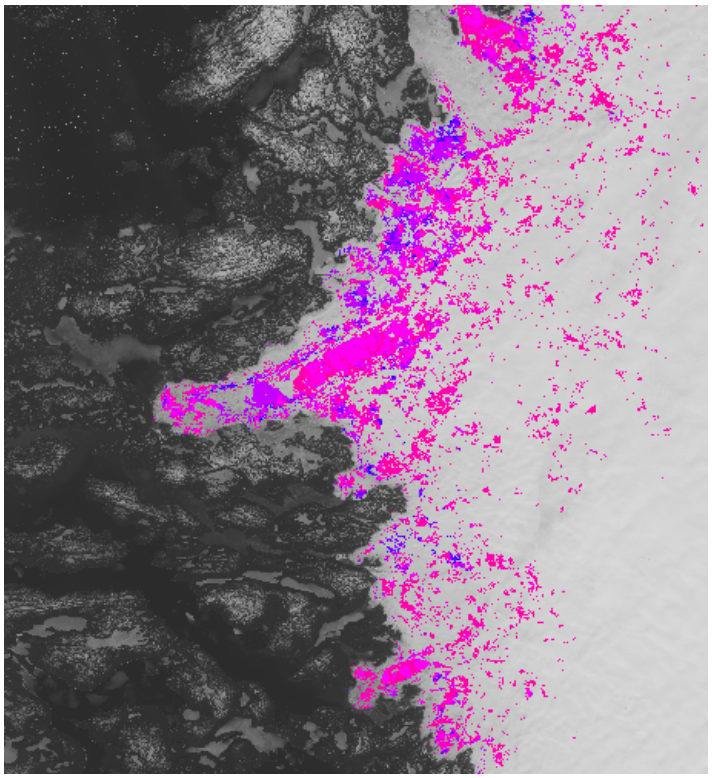


FIGURE 5.4 – Carte LST du motif $3 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1$

5.5.1.5 Comparaison avec une sélection de motifs SFG fiables maximaux

Comme indiqué précédemment, notre méthode peut être appliquée sur n'importe quel ensemble de motifs séquentiels. Dans cette section, nous étudions les résultats obtenus en l'appliquant sur les motifs SFG fiables maximaux, que nous avons extraits avec la méthode présentée dans le chapitre 4, en les comparant avec la sélection sur les motifs SFG maximaux que nous venons de réaliser.

Dans cette configuration, nous retrouvons toujours en première position et en deuxième position les mêmes motifs, $3 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1$ et $2 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1 \to 1$. Au total, il y a 5 motifs en commun entre les deux ensembles. Nous retrouvons également un autre motif représentant des fluctuations de déplacement au niveau des bords des glaciers. Ce motif est $3 \to 3 \to 1 \to 1 \to 3 \to 3 \to 1$, et sa carte LST est présentée dans la figure 5.10. D'ailleurs, au niveau du glacier Sarqardliup Sermia, où nous avons montré des comportements de déplacements très différents par rapport aux autres glaciers, nous avons cette fois-ci le motif $1 \to 1 \to 1 \to 3 \to 1 \to 1$ qui y souligne une longue période de déplacements faibles, interrompue par un déplacement fort. En regardant les Figures 5.11, et 5.12, nous constatons que cette évolution commence au début de la série, pendant la période 1987 et 1991, et qu'elle se termine entre 2004 et 2005. Il y a par conséquent une continuité temporelle entre ce motif et celui révélé avec les motifs

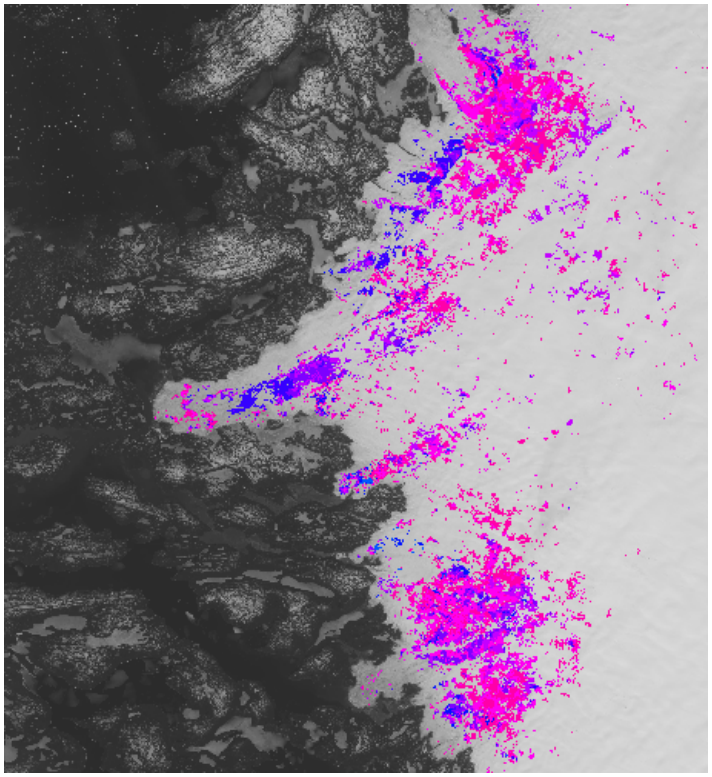


FIGURE 5.5 – Carte LST du motif $2 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1 \to 1$

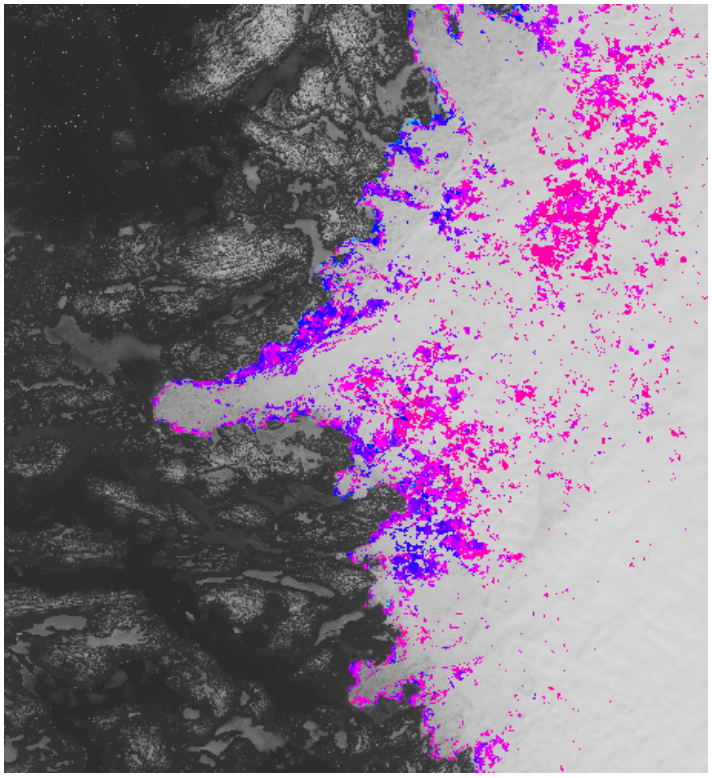


FIGURE 5.6 – Carte LST du motif $3 \to 1 \to 3 \to 3 \to 1 \to 3 \to 1 \to 1 \to 1$

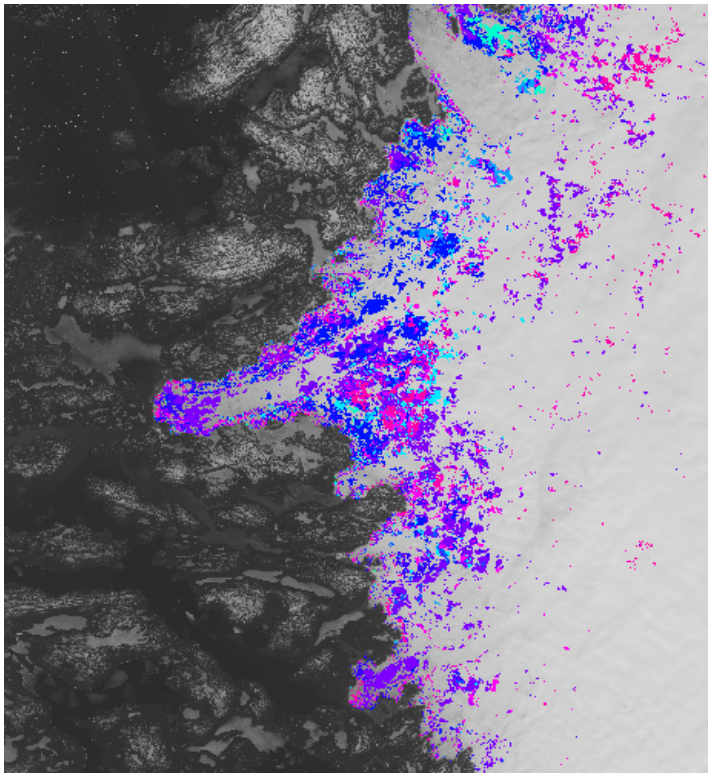


FIGURE 5.7 – Carte LST du motif $3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 2 \rightarrow 1$

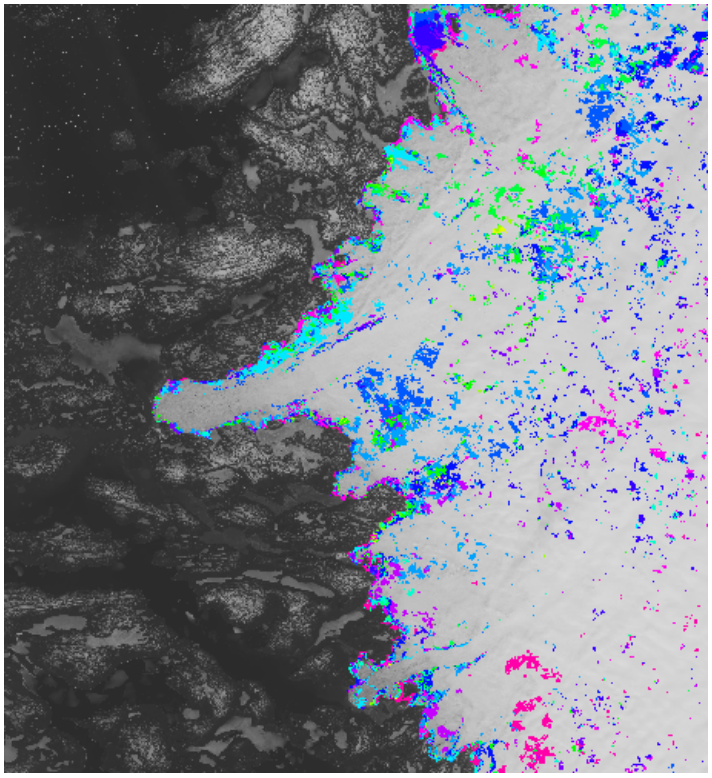


FIGURE 5.8 – Carte LST du motif $3 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 3$

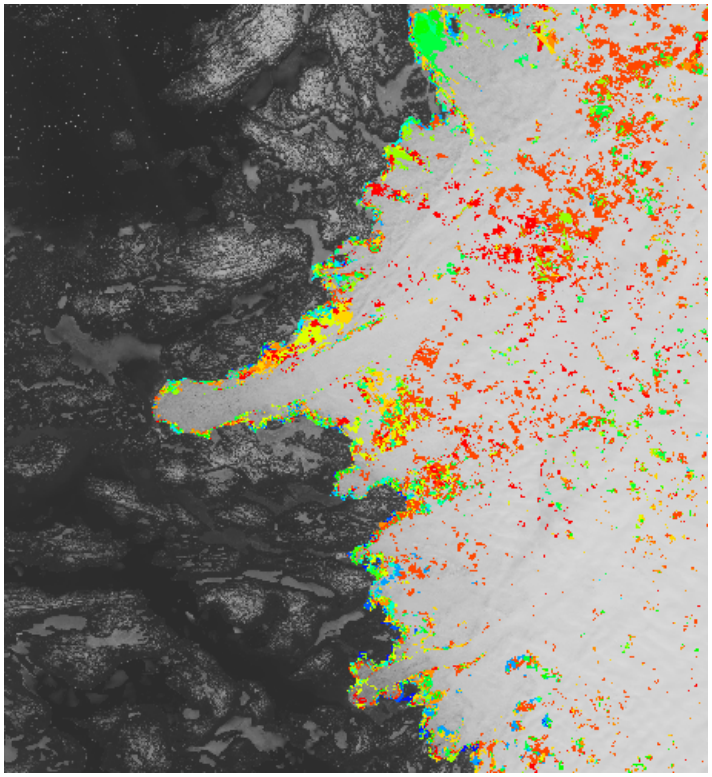


FIGURE 5.9 – Début des occurrences minimales concernées par le motif $3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 1$

normaux, $3 \to 1 \to 1 \to 1 \to 3$, qui, pour rappel, commence entre 2001 et 2002, et se termine entre 2006 et 2009.

La différence qualitative majeure par rapport à la sélection sur les motifs SFG maximaux est l'absence du motif représentant une longue période de déplacements forts suivie d'une décélération soudaine.

Quant aux mesures de complémentarité spatiale et spatiotemporelle, nous avons les valeurs suivantes : $\theta = 0.191$ et $\tau = 0.234$. Ces valeurs sont très proches de celles obtenues sur les motifs SFG maximaux. Cela montre que, pour ces deux types de motifs, notre méthode est capable de sélectionner une liste de motifs disposant d'une complémentarité informationnelle satisfaisante.

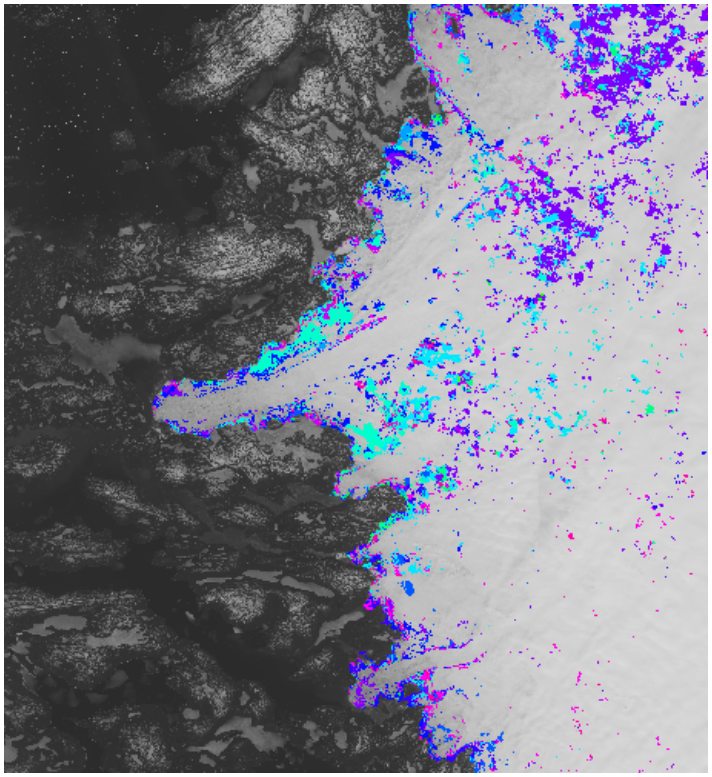


FIGURE 5.10 – Carte LST du motif $3 \rightarrow 3 \rightarrow 1 \rightarrow 1 \rightarrow 3 \rightarrow 3 \rightarrow 1$

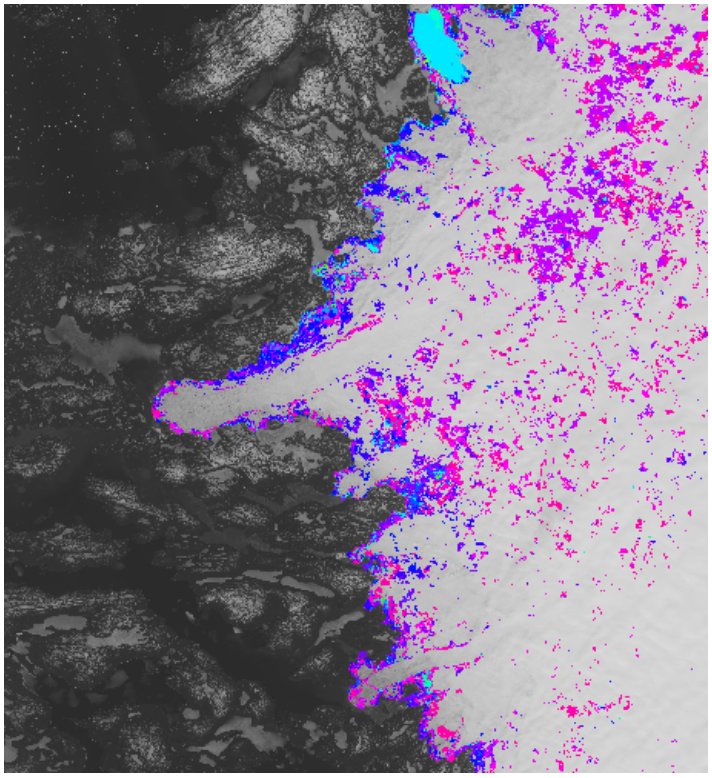


FIGURE 5.11 – Carte LST du motif $1 \rightarrow 1 \rightarrow 1 \rightarrow 3 \rightarrow 1 \rightarrow 1$

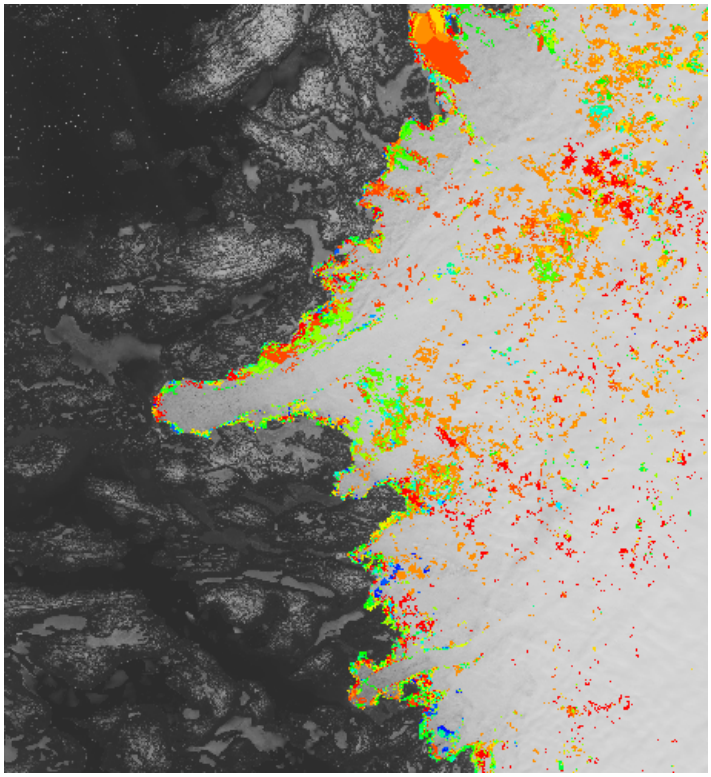


FIGURE 5.12 – Début des occurrences minimales concernées par le motif $1 \to 1 \to 1 \to 3 \to 1 \to 1$

5.5.2 STCD sur le massif du Mont-Blanc provenant de données radar

5.5.2.1 Préparation des données

Pour cette expérience, la BSPP est construite à partir de la STCD utilisée dans la section 4.5.2. Pour rappel, ce jeu de données contient 25 champs de déplacements de taille 3494×3186 pixels, couvrant les glaciers français du massif du Mont-Blanc. Les mesures décrivant les données de déplacement sont les vitesses différentielles médianes, qui représentent, en chaque point et chaque date d'observation, la variation de vitesse par rapport à la vitesse médiane calculée dans la colonne spatiotemporelle centrée sur le point étudié et s'étendant de la première à la dernière date d'observation. L'indice de confiance est, quant à lui, déduit à partir de la stationnarité de la direction de déplacement du glacier. Il est mesuré par le cosinus entre le vecteur de déplacement à une date et une localisation donnée et la direction globale de la vitesse au même point, calculée à l'échelle de la période complète d'observation. Au niveau de la base de séquences, nous avons au total 30062750 données de déplacement, avec une valeur de confiance minimale, maximale et moyenne qui est respectivement de 0.0, 0.999, et 0.487. Comparée à celle du Groenland, cette STCD contient des données qui sont en moyenne beaucoup moins fiables. L'histogramme illustré sur la figure 5.13 montre qu'il y a beaucoup de valeurs de confiance près de 0, et qu'il n'y a plus une dominance de la présence des valeurs près de 1 comme cela a été le cas sur le Groenland. De la même manière que sur le premier jeu de données, nous utilisons ces indices de confiance, notés ρ_{disp}, pour déduire la probabilité des événements probabilistes partiels avec une transformation linaire :

$$\rho = \rho_{disp} \times \frac{2}{3} + \frac{1}{3}$$

Dans cette formule, les mêmes coefficients ont été utilisés puisque nous utilisons toujours 3 symboles pour discrétiser les mesures de déplacement, i.e., $|\mathbb{E}| = 3$.

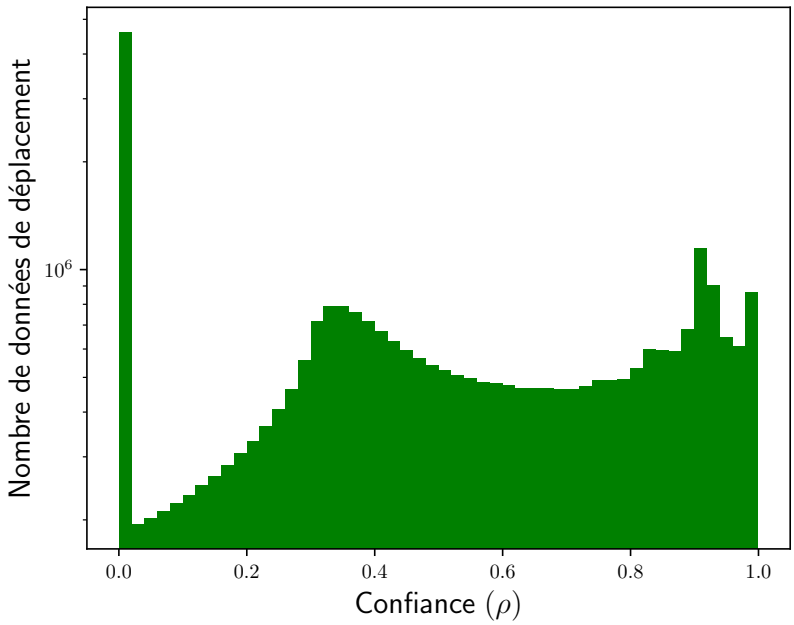


FIGURE 5.13 – Histogramme des indices de confiance (Mont-Blanc)

5.5.2.2 Paramètres utilisés

Comme pour l'expérience sur le Groenland, nous utilisons les motifs SFG maximaux avec les réglages standard de support et de connexité, à partir desquels nous réalisons la sélection des motifs informatifs complémentaires. Comme présenté dans la section 4.5.2.2, le seuil de connexité κ est fixé à 5, et le seuil de support minimal à 4%. Le nombre de motifs à sélectionner sera, quant à lui, fixé à 20.

5.5.2.3 Résultats quantitatifs

Avec les mêmes configurations logicielles et matérielles que l'expérience sur le Groenland, la sélection des 20 motifs les mieux classés a pris environ 25 heures et 45 minutes.

La liste des motifs sélectionnés ainsi que le gain d'information et le support associés sont présentés dans la table 5.4.

De la même manière que lors de l'expérience sur le Groenland, avec l'utilisation du gain d'information moyen le support des motifs n'influence pas directement l'ordre de la sélection. Comme dans l'expérience précédente, le gain d'information diminue fortement au début après chaque sélection (cf. Figure 5.14). Mais comme ce jeu de données est beaucoup moins confiant que celui du Groenland, les gains d'information sont à des niveaux moins élevés.

Quant à la complémentarité spatiale, nous utilisons la mesure θ (cf. Équation 5.19) pour comparer les motifs informatifs sélectionnés et ceux sélectionnés avec l'approche de *swap* randomisation. Elle a une valeur de 0.111 pour les premiers et vaut 0.098 pour les seconds.

#	Motif séquentiel informatif	Gain d'information	Support
1	$2 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1 \to 2 \to 2 \to 2 \to 2$	7.283827	444331
2	$3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1$	4.961673	440564
3	$2 \to 2 \to 2 \to 2 \to 1 \to 1 \to 3 \to 3 \to 3 \to 3$	3.902344	457537
4	$3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 2$	2.826863	594808
5	$3 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1$	2.238475	508275
6	$1 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2$	1.761042	447017
7	$2 \to 3 \to 2 \to 3 \to 2 \to 3 \to 2 \to 3$	1.299521	529660
8	$1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 1 \to 2$	1.025332	629861
9	$3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3$	0.73448	714895
10	$2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1 \to 1$	0.699169	513383
11	$3 \to 3 \to 1 \to 1 \to 1 \to 2 \to 2 \to 2 \to 2 \to 2$	0.438209	483669
12	$1 \to 1 \to 2 \to 1 \to 2 \to 2 \to 3 \to 3$	0.315635	563307
13	$3 \to 3 \to 3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 1$	0.254705	523901
14	$2 \to 1 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2$	0.208004	512526
15	$2 \to 2 \to 3 \to 3 \to 3 \to 1 \to 1 \to 1$	0.176175	581509
16	$1 \to 3 \to 1 \to 3 \to 3 \to 3 \to 3 \to 3$	0.139949	630708
17	$2 \to 2 \to 2 \to 2 \to 2 \to 1 \to 1 \to 1 \to 3 \to 1$	0.111025	473522
18	$1 \to 1 \to 1 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2$	0.08718	500607
19	$3 \to 3 \to 3 \to 3 \to 3 \to 3 \to 1 \to 1$	0.078084	741112
20	$2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 3$	0.065063	498970

TABLE 5.4 – Liste des motifs informatifs sélectionnés (Mont-Blanc)

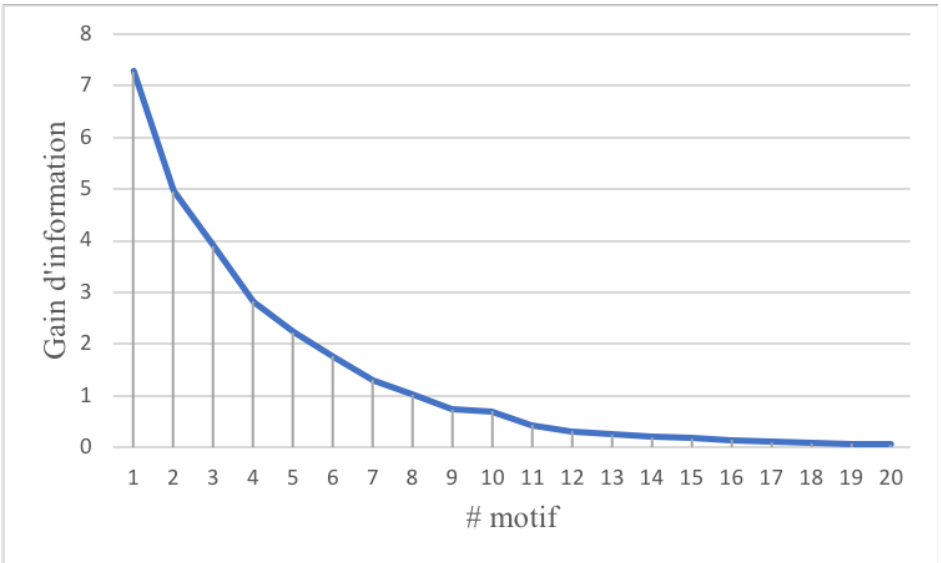


FIGURE 5.14 – Le gain d'information des motifs sélectionnés (Mont-Blanc)

Comme sur la STCD du Groenland, cet écart montre que les motifs informatifs sélectionnés sont en moyenne plus complémentaires spatialement que ceux sélectionnés par la méthode proposée par Méger *et al.* (2015). Pour la mesure de complémentarité spatiotemporelle τ (cf. Équation 5.20), nous obtenons une valeur de 0.153. Ces mesures sont toutes plus petites que celles obtenues sur le Groenland. Cela peut être expliqué par la différence du niveau de confiance sur les données. En effet, les contraintes de type #3, qui affectent les événements

ayant lieu entre les dates d'occurrence des éléments d'une OMDIT, nécessitent des données très confiantes afin de fournir un gain informationnel. La STCD du Mont-Blanc ayant une faible confiance, les contraintes de type #3 fournissent peu d'information sur les événements situés entre les éléments d'une OMDIT. Ceci nécessite donc de sélectionner ensuite d'autres motifs dont les OMDIT vont aussi recouvrir cette zone pour fournir de l'information sur ces événements. Cela entraîne globalement des recouvrements qui vont diminuer la complémentarité.

5.5.2.4 Résultats qualitatifs

Des évolutions de déplacements intéressantes sur différentes régions sont révélées par les motifs sélectionnés. Par exemple, le premier motif sélectionné, $2 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1 \to 2 \to 2 \to 2 \to 2$, souligne deux périodes significatives de déplacements moyens, séparées par une période de déplacements faibles. Au niveau spatial, cette évolution est présente dans la plupart des principaux glaciers, sauf la partie sommitale du glacier d'Argentière et le glacier de Leschaux (cf. Figure 5.16). Le deuxième motif sélectionné, $3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1$, souligne une décélération progressive du déplacement des glaciers. Cette évolution est présente notamment à la fin des glaciers du Taconnaz et des Bossons. On le retrouve également tout au long du glacier du Géant et la Mer de Glace, ainsi que sur la moitié basse du glacier d'Argentière (cf. Figure 5.17). Ensuite, le troisième motif sélectionné est $2 \to 2 \to 2 \to 2 \to 1 \to 1 \to 3 \to 3 \to 3 \to 3$. Il souligne une longue période de déplacements moyens, suivie d'une période courte de déplacements faibles et finalement d'une longue période de déplacements forts (cf. Figure 5.18). Cette évolution est présente notamment au niveau du glacier des Bossons, des séracs du Géant et de la Mer de Glace. Au niveau temporel (échelle de couleur donnée Figure 5.15), elle prend fin plus tôt au niveau de la zone terminale du glacier des Bossons comme aux séracs du Géant (en juin 2011) que sur les autres régions (en août 2011). L'ensemble de ces trois premiers motifs possèdent de bonnes valeurs de complémentarité spatiale et spatiotemporelle, avec $\theta = 0.674$ et $\tau = 0.727$.

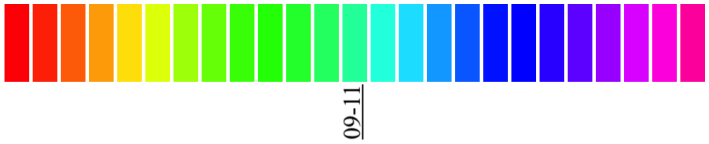


FIGURE 5.15 – Échelle de couleur des cartes LST : de rouge (Mai 2009) à vert (Octobre 2009) et de bleu (Mai 2011) à magenta (Septembre 2011)

5.5.2.5 Comparaison avec une sélection de motifs SFG fiables maximaux

De la même manière que sur la STCD du Groenland, nous effectuons également une sélection des motifs informatifs à partir des motifs SFG fiables maximaux que nous avons extraits dans le chapitre 4.

Les 7 premiers motifs sélectionnés sont identiques à ceux sélectionnés parmi les motifs SFG maximaux. Au total, nous retrouvons 11 motifs en commun dans les deux ensembles. Par ailleurs, nous retrouvons le motif exprimant les déplacements par cycles annuels révélés par Fallourd (2012); Ponton *et al.* (2014). Ce motif est $3 \to 3 \to 2 \to 1 \to 1 \to 1 \to 1 \to 3 \to 2 \to 2 \to 2$, et sa carte LST est présentée dans la figure 5.19. En examinant

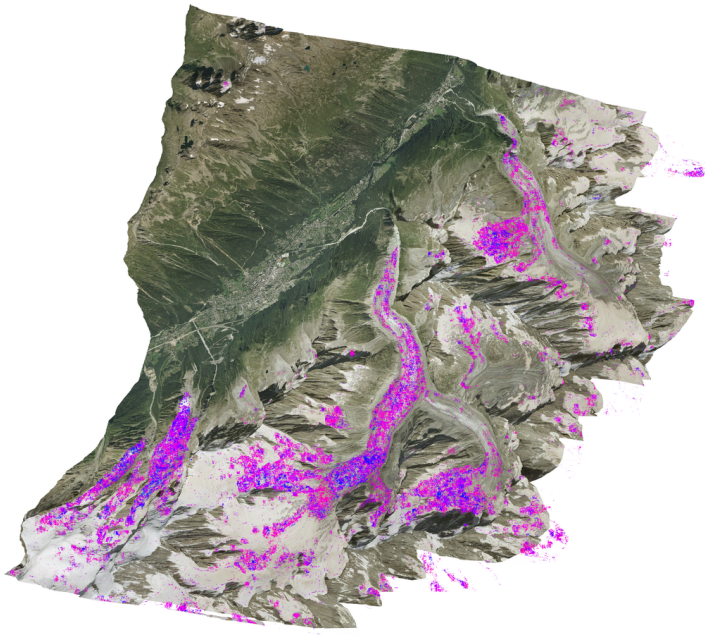


FIGURE 5.16 – Carte LST du motif $2 \to 2 \to 2 \to 1 \to 1 \to 1 \to 1 \to 2 \to 2 \to 2 \to 2$

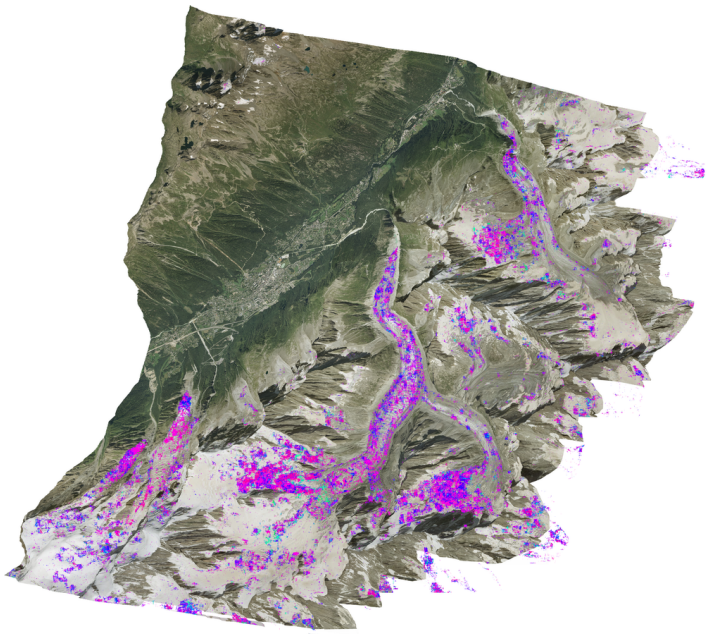


FIGURE 5.17 – Carte LST du motif $3 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 1$

les autres cartes exprimant les dates d'occurrence du premier symbole et celles du huitième symbole (troisième symbole 3), nous constatons que la première période de décélération, i.e., $3 \to 3 \to 2 \to 1 \to 1 \to 1 \to 1$, commence au début de l'été 2009 (cf. Figure 5.20), tandis que la deuxième décélération, $3 \to 2 \to 2 \to 2$, commence également pendant cette période de l'année 2011 (cf. Figure 5.21).

Les motifs sélectionnés ont cette fois-ci des mesures de complémentarité spatiale et spatiotemporelle suivantes : $\theta = 0.115$ et $\tau = 0.157$. Comme sur le jeu de données du Groenland,

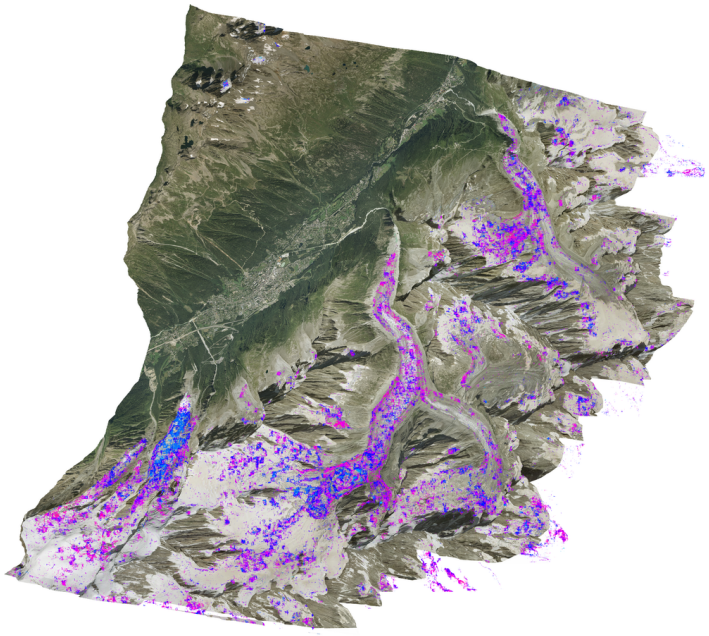


FIGURE 5.18 – Carte LST du motif $2 \rightarrow 2 \rightarrow 2 \rightarrow 2 \rightarrow 1 \rightarrow 1 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$

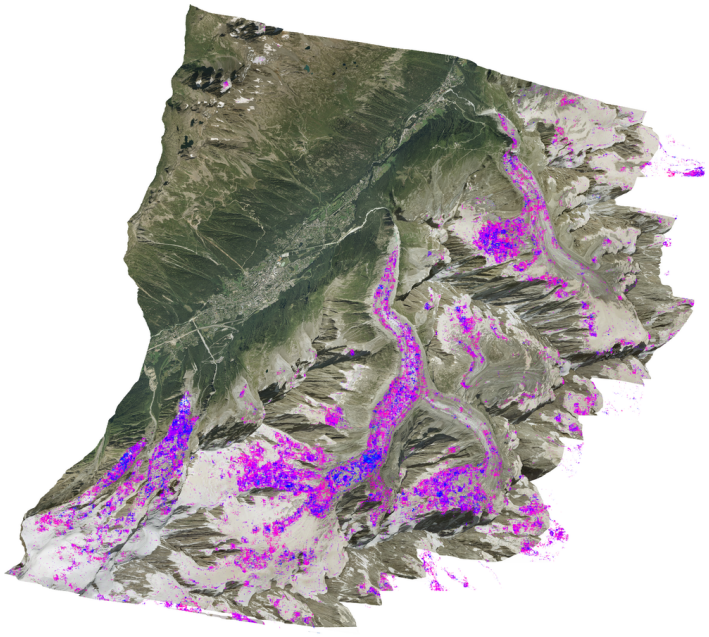


FIGURE 5.19 – Carte LST du motif $3 \rightarrow 3 \rightarrow 2 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 3 \rightarrow 2 \rightarrow 2 \rightarrow 2$

nous obtenons des mesures de complémentarité très proches pour les deux ensembles de motifs (SFG maximaux et SFG fiables maximaux). Cela montre encore une fois que nous arrivons à sélectionner des motifs à partir des ensembles différents avec des niveaux de complémentarité stables.

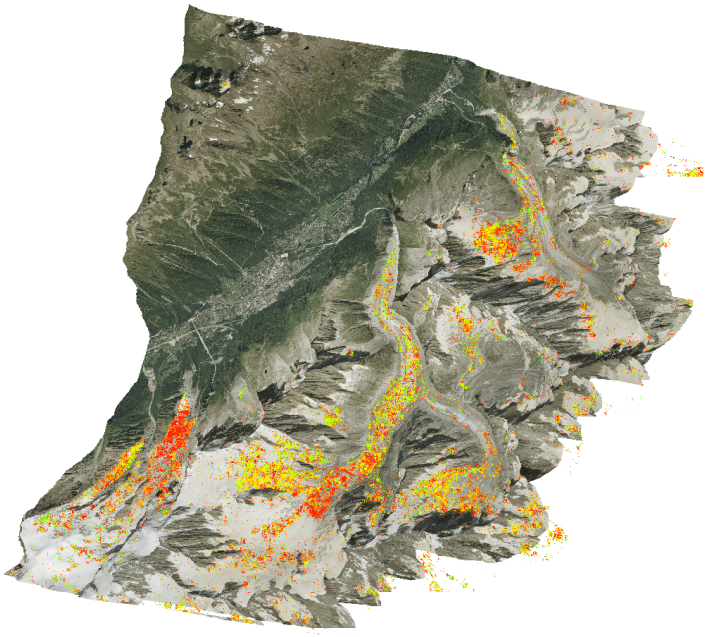


FIGURE 5.20 – Dates d'occurrence du premier symbole du motif
$3 \to 3 \to 2 \to 1 \to 1 \to 1 \to 1 \to 3 \to 2 \to 2 \to 2$

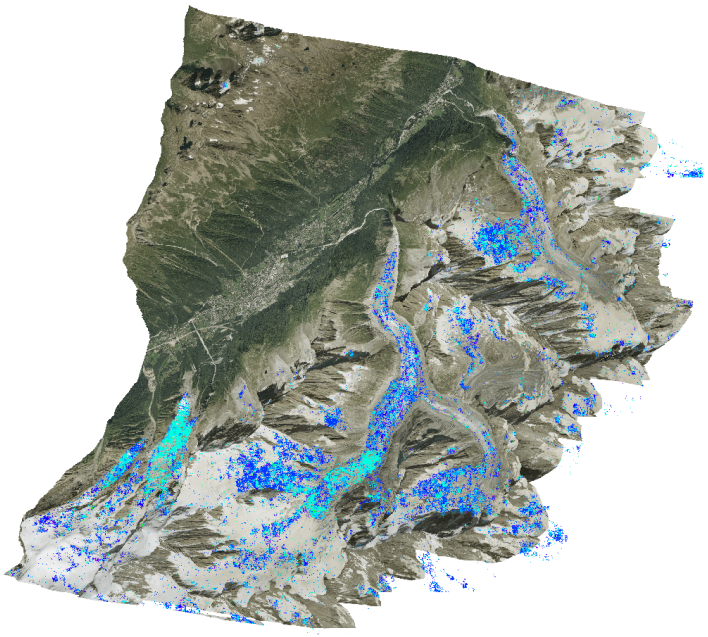


FIGURE 5.21 – Dates d'occurrence du huitième symbole du motif
$3 \to 3 \to 2 \to 1 \to 1 \to 1 \to 1 \to 3 \to 2 \to 2 \to 2$

5.6 Conclusions

Dans ce chapitre, nous avons présenté notre contribution concernant la sélection des motifs basée sur un critère informationnel. Cette tâche s'effectue grâce aux apports suivants :

1 la définition d'une BSPP, dont chacun des événements probabilistes est décrit par une distribution,
2 la définition du gain informationnel apporté par des motifs séquentiels dans une BSPP,
3 l'introduction de contraintes sur les événements probabilistes et obtenues à l'issue de la connaissance des Occurrences Minimales avec Dates Intermédiaires au plus Tôt (OMDIT),
4 la méthode pour combiner les différentes contraintes et pour en déduire le gain informationnel minimal que l'on peut obtenir sur chaque événement,
5 l'heuristique pour déterminer un ensemble de motifs fournissant de façon complémentaire un gain informationnel élevé.

Notre proposition a été appliquée à deux STCD utilisées aussi dans le chapitre précédent : une disposant de bons indices de confiances, couvrant les glaciers du Groenland, et une dont les mesures sont moins fiables, couvrant les glaciers français du massif du Mont-Blanc. Des motifs sélectionnés, à partir des motifs SFG maximaux et des motifs SFG fiables maximaux, offrent des complémentarités spatiales supérieures à la méthode par *swap* randomisation. Avec une localisation précise des occurrences les plus informatives, notre méthode permet ainsi d'analyser les résultats de façon détaillée. Ce niveau de détail, combiné avec la complémentarité entre les motifs, facilite l'interprétation de l'utilisateur. Finalement, en utilisant des probabilités produites à partir des indices de confiances, notre méthode est par nature capable de prendre en compte la confiance des événements constituant les occurrences, lors de la sélection des motifs. Ceci est confirmé expérimentalement par le fort recoupement entre les motifs qui ont été sélectionnés à partir des motifs SFG maximaux, et ceux sélectionnés à partir des motifs SFG fiables maximaux.

Chapitre 6

Conclusions et perspectives

Dans ce mémoire, nous nous sommes intéressés à la découverte de connaissances à partir de Séries Temporelles de Champs de Déplacements (STCD). De telles séries occupent aujourd'hui une place centrale dans l'étude et la surveillance de phénomènes naturels tels que les tremblements de terre, les éruptions volcaniques ou bien encore le déplacement des glaciers. En effet, ces séries sont riches d'informations à la fois spatiales et temporelles et peuvent aujourd'hui être produites régulièrement et à moindre coût grâce à des programmes spatiaux tels que le programme européen Copernicus et ses satellites phares Sentinel.

Afin de favoriser la découverte de connaissances, nous nous sommes orientés vers des méthodes non supervisées. C'est ainsi que nous nous sommes appuyés sur les motifs Séquentiels Fréquents Groupés (SFG) tels que définis par Julea *et al.* (2011). Ces motifs permettent d'extraire des groupes de pixels suffisamment nombreux, groupés spatialement et partageant une même évolution ou sous-évolution. À l'origine définis pour l'analyse de Séries Temporelles d'Images Satellitaires (STIS), ceux-ci ont montré leur potentiel quant à l'analyse de Séries Temporelles de Champs de Déplacements (STCD) grâce aux travaux de Pericault *et al.* (2015). Ils ne sont néanmoins pas capables de prendre en compte les indices de confiance associés aux mesures de déplacement.

Prise en compte des indices de confiance dans l'extraction de motifs

Notre première contribution a donc consisté à ne retenir que les motifs SFG portés par des occurrences s'appuyant sur des données dont la confiance est suffisante en moyenne. Pour chaque occurrence d'un motif, une mesure de fiabilité est établie en sélectionnant la valeur minimale de confiance observée à l'échelle de chacun des événements formant cette occurrence, chaque événement correspondant à la description d'un pixel particulier à une date précise. Cette mesure conservatrice est par la suite utilisée à l'échelle de chaque séquence pour déterminer et retenir, pour chaque motif, l'occurrence ayant la meilleure mesure de fiabilité. Enfin, à l'échelle d'une base de séquences, les mesures obtenues pour chacune des séquences et chaque motif sont moyennées afin d'associer à chacun des motifs une mesure de fiabilité moyenne. Inspirés par les travaux de Pei *et al.* (2007), nous avons proposé un algorithme d'extraction sous contrainte de fiabilité permettant d'extraire les motifs dont la mesure de fiabilité dépasse un seuil fourni par l'utilisateur. Cet algorithme est capable d'identifier, à l'échelle d'une séquence, l'occurrence d'un motif ayant la plus haute fiabilité grâce à un

algorithme de programmation dynamique dont la complexité est linéaire. Par ailleurs, la contrainte de fiabilité est intégrée au sein même de l'algorithme d'extraction grâce à un élagage de l'espace de recherche. Ce dernier s'appuie sur une technique de *push* partiel mise en place grâce à une contrainte exercée sur une borne supérieure de la fiabilité. Cette contrainte impose que cette borne dépasse également le seuil de fiabilité fourni initialement par l'utilisateur. Grâce à son anti-monotonicité, il devient possible d'élaguer des motifs tout en s'assurant de la complétude des résultats retournés. Dans le cas où cette contrainte anti-monotone est validée, une simple vérification est faite afin de s'assurer que la mesure de fiabilité dépasse également le seuil demandé. Il est à noter que la notion de fiabilité développée dans cette contribution peut être utilisée pour qualifier n'importe quel type de motif séquentiel, qu'il soit fréquent ou non, groupé ou non.

La méthode proposée a été implémentée sur la base du prototype *SITS-P2miner* développé par les laboratoires LISTIC et LIRIS pour l'analyse de Séries Temporelles d'Images Satellitaires (STIS). Des expériences sur une STCD de référence produite par Tedstone *et al.* (2015) à partir de données optiques et sur une STCD que nous avons générée à partir de données radar ont permis de montrer que la méthode était d'autant plus efficace que les données étaient peu fiables, ce qui est cohérent avec la stratégie d'élagage proposée. Le gain d'efficacité apporté par cette stratégie d'élagage a permis d'envisager un réglage automatique de l'unique paramètre nécessaire à la prise en compte des indices de confiance, à savoir le seuil minimum de fiabilité moyenne. Par ailleurs, ces expériences ont montré qu'extraire des motifs sous contrainte de fiabilité moyenne permet d'extraire des motifs couvrant en moyenne plus de données que les motifs extraits directement des données après un simple seuillage sur la confiance.

D'un point de vue qualitatif, le ralentissement des glaciers du Groenland rapportés pour quelques transects par Tedstone *et al.* (2015) a été confirmé, spatialisé à l'échelle de toute la zone observée et raffiné dans le temps avec l'identification de périodes distinctes de ralentissement. Par ailleurs, un phénomène d'accélération localisé sur la partie côtière au début et à la fin des années 90 a également été identifié. Concernant la série d'images radar couvrant les glaciers du massif du Mont-Blanc, les résultats obtenus ont permis de confirmer les fluctuations de vitesse rapportées pour les transects étudiés par Fallourd (2012) et Ponton *et al.* (2014). À nouveau, ces comportements ont pu être identifiés à l'échelle de la zone observée tout entière, jusqu'à des altitudes d'environ 3000 m *a.s.l.*, ce qui marque ici une conséquence du réchauffement climatique. En effet, à cette altitude, les glaciers sont considérés comme froid et donc non sujets à de telles fluctuations de vitesse.

Sélection de motifs complémentaires les plus informatifs

Notre deuxième contribution prend également en compte les indices de confiance des STCD. Celle-ci a pour objectif la sélection des motifs SFG (préalablement extraits, fiables ou non) les plus informatifs et les plus complémentaires possible. Cette complémentarité n'était en effet pas vérifiée par la méthode de *swap* randomisation employée par Pericault *et al.* (2015), cette dernière classant chacun des motifs individuellement. En s'inspirant des travaux de Lam *et al.* (2014), nous avons tout d'abord défini le gain informationnel minimum associé à chaque motif. À cette fin, la connaissance initiale de l'utilisateur concernant une STCD est considérée comme se résumant à décrire chaque événement par une variable aléatoire indépendante qui suit une distribution uniforme. Cette hypothèse d'uniformité peut être faite lorsque les valeurs de déplacement sont quantifiées de façon à ce que les fréquences des sym-

boles obtenus soient identiques. La connaissance d'un motif permet, grâce à ses occurrences, de raffiner ces distributions. En effet, pour chaque événement formant une occurrence, l'indice de confiance est interprété comme la probabilité associée au symbole décrivant l'événement. Nous nous plaçons ainsi dans le cas d'une base de séquences probabiliste dont les distributions ne sont que partiellement connues, contrairement aux travaux Muzammal et Raman (2015) où des distributions complètes sont considérées. Par ailleurs, chaque occurrence contient la probabilité minimale observée pour les événements apparaissant durant la période temporelle qu'elle couvre. Cette information est donc également utilisée, de concert avec les informations déduites de l'agencement temporel des occurrences par rapport aux événements de la STCD. Ces dernières permettent d'exclure, pour certains événements, la présence de certains symboles. De façon générale, les informations apportées par les occurrences sont formalisées sous forme de contraintes auxquelles doivent obéir les distributions des variables aléatoires représentant les événements de la STCD. Ces distributions sont estimées de façon à quantifier le gain informationnel minimum, i.e., en minimisant la divergence Kullback-Leibler avec les distributions initialement connues. Sur cette base, une heuristique de type algorithme glouton permet, à chacune de ses itérations, de sélectionner le motif le plus informatif en prenant en compte les informations apportées par les motifs sélectionnés lors des itérations précédentes. De la sorte, les motifs sélectionnés sont complémentaires, et leur nombre, correspondant au nombre d'itérations, est maîtrisé par l'utilisateur final. Il est à noter qu'il s'agit là du seul paramètre de la méthode proposée et que cette contribution peut être utilisée pour sélectionner n'importe quel type de motif séquentiel, qu'il soit fréquent ou non, groupé ou non. De plus, l'algorithme est parallélisable à plusieurs niveaux, ce qui permet d'envisager une mise en production maîtrisée. Des expériences réalisées à l'aide d'un prototype dédié sur la STCD couvrant les glaciers du massif du Mont-Blanc et les glaciers du Groenland ont permis de vérifier que les motifs sélectionnés sont complémentaires, spatialement et temporellement. D'un point de vue qualitatif, ils mettent en avant les phénomènes les plus importants, phénomènes qui avaient été sélectionnés manuellement par les experts à partir des classements établis par *swap* randomisation après extraction des motifs séquentiels fréquents groupés et fiables.

Perspectives

En ce qui concerne les perspectives ouvertes par ces travaux, la première concerne les indices de confiance en eux-mêmes. En effet, qu'ils soient issus du calcul même des champs de déplacements ou d'une qualification *a posteriori* des propriétés spatiotemporelles des champs produits, ceux-ci ne représentent qu'une des étapes permettant d'obtenir les symboles de déplacement qu'ils qualifient *in fine*. Plus précisément, les déplacements sont calculés à partir de données par nature sujettes à des incertitudes systémiques liées aux instruments d'observation et aux chaînes de traitement, et à des incertitudes stochastiques liées aux bruits et aux conditions d'observation. De plus, les mesures de déplacement obtenues à partir de ces données incertaines sont quantifiées pour obtenir les symboles de déplacement. Il conviendrait donc d'intégrer aux calculs des indices de confiance actuellement effectués un terme permettant d'exprimer jusqu'à quel point il est possible d'avoir confiance dans la donnée d'origine et un terme permettant de pénaliser les situations où la valeur de déplacement mesurée est proche des seuils de quantification utilisés, en utilisant par exemple des fonctions d'appartenance (Zadeh, 1965). Les nouveaux indices de confiance pourraient alors être directement utilisés par l'algorithme d'extraction proposé. Il est à noter que celui-ci impose une contrainte stricte sur le seuil de fiabilité alors que cette mesure est basée sur des indices de confiance dont on sait qu'ils n'expriment qu'une approximation imparfaite des incertitudes liées aux données et

aux calculs. Une solution serait alors par exemple de s'inspirer des travaux de Ugarte *et al.* (2015) afin de relaxer cette contrainte de fiabilité à l'aide de techniques de programmation par contraintes.

Une deuxième perspective a trait à l'adaptation de nos propositions en vue d'extraire des *itemsets* tels que définis par Agrawal *et al.* (1993). En effet, pour ces derniers, la définition d'une mesure de fiabilité pourrait être directement calquée sur la mesure de fiabilité proposée dans ce mémoire. L'algorithme d'extraction correspondant pourrait également s'appuyer sur la stratégie d'élagage que nous avons proposée. Concernant la sélection de motifs séquentiels sur critère informationnel, celle-ci pourrait également être directement adaptée aux *itemsets*, avec néanmoins, la perte des informations liées à l'agencement temporel des occurrences dans les données, un *itemset* étant simplement présent ou non dans une transaction.

Une troisième perspective se dessine quant à la sélection de motifs sur critère informationnel. Il serait en effet intéressant d'intégrer l'utilisateur au sein de l'algorithme glouton de façon à orienter la découverte de connaissance en fonction de son intérêt et de la tâche associée. Plus précisément, à chaque itération, le classement des motifs en fonction de leur gain informationnel pourrait être présenté à l'utilisateur, chaque motif étant accompagné de sa carte de localisation spatiotemporelle. L'utilisateur pourrait alors choisir le motif lui paraissant le plus intéressant et indiquer quels sont les motifs qu'il "aime" ou "n'aime pas". Sur cette base, un modèle de préférence pourrait être construit, affiné et utilisé à chaque itération de façon à pondérer les classements construits sur critère informationnel. La construction de ce modèle de préférence reste un problème ouvert et important comme expliqué par Van Leeuwen (2014). De nombreux travaux recensés dans Crémilleux *et al.* (2016) peuvent servir de support à une première réflexion. Par exemple, il serait possible de procéder comme dans Bhuiyan *et al.* (2012) en calculant un score pour chaque symbole qui soit fonction du nombre de motifs "aimés" par l'utilisateur. Les scores obtenus pour chaque symbole pourraient alors être utilisés pour pondérer le gain informationnel de chaque motif. Des stratégies plus fines et adaptées aux données séquentielles sont cependant à envisager afin de se concentrer plutôt sur la notion d'évolution symbolique que sur les symboles en eux-mêmes. Cette interaction avec l'utilisateur final est d'autant plus envisageable que la visualisation des données et des motifs peut se faire sur la base de cartes spatiotemporelles et que l'algorithme de sélection est linéaire et parallélisable à plusieurs niveaux. Si toutefois, des problèmes de performance venaient à se présenter, des approches de type échantillonnage pourraient être employées comme dans Toivonen (1996).

Une quatrième perspective concerne l'extraction directe de motifs informatifs et complémentaires. Pour ce faire, à l'image des travaux visant à extraire les motifs qui compressent le jeu de données tels que ceux de Tatti et Vreeken (2012) ou de Lam *et al.* (2014), des algorithmes de type glouton pourraient être envisagés. L'idée générale serait de générer des motifs par ajouts successifs de types d'événement, ces ajouts étant guidés par maximisation d'une fonction objectif de type gain d'information, et ce jusqu'à obtenir un nombre de motifs fixé par l'utilisateur. Une telle approche permettrait ainsi de supprimer le coût de l'extraction des motifs tout en étant capable de fournir des motifs informatifs et complémentaires. Par ailleurs, dans la lignée des travaux de Bascol *et al.* (2016), il serait intéressant de voir dans quelle mesure il serait possible d'extraire des motifs à partir d'un *autoencoder* tout en prenant en compte les indices de confiance. En effet, les motifs fournis par un *autoencoder* permettent de couvrir au mieux les données, à l'image des motifs sélectionnés sur critère de compression comme dans Lam *et al.* (2014) ou sur critère informationnel comme proposé dans ce mémoire.

Table des figures

Liste des tableaux

Acronymes

ACP Analyse en Composantes Principales 11, 61

ASAR *Advanced Synthetic Aperture Radar* 12

BSPP Base de Séquences Probabiliste Partielle 86–88, 92, 105, 106, 108, 115, 122, 151

DInSAR Interférométrie Différentielle SAR (en anglais *Differential Interferometry SAR*) 17, 19, 20, 22, 23

DP *Data Point cover* 80

ECD Extraction de Connaissances à partir des Données 3

ESA Agence spatiale européenne (en anglais *European Space Agency*) 2

FreeSpan *Frequent pattern-projected Sequential pattern mining* 38

GPS *Global Positioning System* 2, 15, 19

GRD *Ground Range Detected* 12, 13

GSP *Generalized Sequential Pattern* 38

KL Kullback-Leibler 96, 98–101

LEO Orbite Terrestre Basse (en anglais *Low Earth Orbit*) 12

LOS ligne de visée (en anglais *Line Of Sight*) 12, 48, 71

LST Localisation Spatio-Temporelle 49, 51, 64, 66, 67, 75, 78, 82, 109, 110, 118, 147, 148

MDL Longueur de Description Minimale (en anglais *Minimal Description Length*) 42, 45, 84

MDP *Mean Data Point cover* 80

MDS vitesse médiane différentielle (en anglais *Median Differential Speed*) 62, 63, 72, 73, 105

MERIS *Medium Resolution Imaging Spectrometer* 19

MNT Modèle Numérique de Terrain 15, 18

MODIS *Moderate Resolution Imaging Spectroradiometer* 19

NASA Administration nationale de l'aéronautique et de l'espace des États-Unis (en anglais *National Aeronautics and Space Administration*) 2

NDVI indice de végétation par différence normalisée (en anglais *Normalized Difference Vegetation Index*) 21

NMI Information Mutuelle Normalisée (en anglais *Normalized Mutual Information*) 147, 149

OLI *Operational Land Imager* 2

OMDIT Occurrence Minimale avec dates Intermédiaires au plus Tôt 92–95, 101–103, 106, 108, 109, 118, 151

PrefixSpan *Prefix-projected Sequential pattern mining* 38, 40

PS *Persistent Scatterers* 20

PSP *Prefix tree for Sequential Pattern* 38

SAR Radar à Synthèse d'Ouverture (en anglais *Synthetic Aperture Radar*) viii, 2, 7, 10, 12, 13, 15–17, 19–24, 28, 48, 61, 71, 82, 83

SBAS *Small BASeline* 20, 23

SFG Séquentiel Fréquent Groupé v, viii, 5, 6, 47–52, 54, 56, 58, 60, 62–64, 66, 68, 70, 72, 74–76, 78–83, 88, 89, 106, 110, 113, 116, 118, 120, 122–124, 128, 147, 149

SLC *Single Look Complex* 12, 13, 23

SNR Rapport Signal sur Bruit (en anglais *Signal to Noise Ratio*) 23, 24, 61

SOT Satellite d'Observation de la Terre 2, 10

SPADE *Sequential PAttern Discovery using Equivalence classes* 38, 40

SPAM *Sequential PAttern Mining* 38

SPIRIT *Sequential Pattern mIning with Regular expressIon consTraints* 39, 40

STCD Série Temporelle de Champs de Déplacements v, vii, viii, 2, 3, 5–10, 12, 14, 16, 18, 20, 22, 24–29, 33, 34, 48–51, 61, 63, 64, 71–74, 80, 82–88, 105, 115, 117, 118, 122–125, 147, 148

STIS Série Temporelle d'Images Satellitaires v, vii, 2–6, 9, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 33, 124

VVC cohérence des vecteurs de déplacements (en anglais *Velocity Vector Coherence*) 61, 105

Bibliographie

Abiteboul, S., Kanellakis, P., et Grahne, G. (1987). On the Representation and Querying of Sets of Possible Worlds. In *Proceedings of the 1987 ACM SIGMOD International Conference on Management of Data*, SIGMOD '87, (pp. 34–48). New York, NY, USA : ACM.

Aggarwal, C. C. (2009). *Managing and Mining Uncertain Data*. Springer Publishing Company, Incorporated.

Agrawal, R., Imieliński, T., et Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*(2), 207–216.

Agrawal, R., et Srikant, R. (1995). Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, (pp. 3–14). IEEE.

Akbari, V., Doulgeris, A. P., et Eltoft, T. (2014). Monitoring Glacier Changes Using Multitemporal Multipolarization SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(6), 3729–3741.

Albert-Lorincz, H., et Boulicaut, J.-F. (2003). Mining frequent sequential patterns under regular expressions : a highly adaptative strategy for pushing constraints. *Proceedings SIAM DM 2003*, (pp. 316–320).

Altena, B., Scambos, T., Fahnestock, M., et Kääb, A. (2018). Extracting recent short-term glacier velocity evolution over Southern Alaska from a large collection of Landsat data. *The Cryosphere Discussions*, (pp. 1–27).

Aminikhanghahi, S., et Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, *51*(2), 339–367.

Andrews, L. C., Catania, G. A., Hoffman, M. J., Gulley, J. D., Lüthi, M. P., Ryser, C., Hawley, R. L., et Neumann, T. A. (2014). Direct observations of evolving subglacial drainage beneath the Greenland Ice Sheet. *Nature*, *514*(7520), 80–83.

Ayres, J., Flannick, J., Gehrke, J., et Yiu, T. (2002). Sequential pattern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 429–435). ACM.

Ball, J. E., Anderson, D. T., et Chan, C. S. (2018). Special Section Guest Editorial : Feature and Deep Learning in Remote Sensing Applications. *Journal of Applied Remote Sensing*, *11*(4), 042601.

Bascol, K., Emonet, R., Fromont, E., et Odobez, J.-M. (2016). Unsupervised Interpretable Pattern Discovery in Time Series Using Autoencoders. In A. Robles-Kelly, M. Loog, B. Biggio, F. Escolano, et R. Wilson (Eds.) *Structural, Syntactic, and Statistical Pattern Recognition*, (pp. 427–438). Cham : Springer International Publishing.

Bhuiyan, M., Mukhopadhyay, S., et Hasan, M. A. (2012). Interactive Pattern Mining on Hidden Data : A Sampling-based Solution. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, (pp. 95–104). New York, NY, USA : ACM.

Bioucas-Dias, J. M., et Valadão, G. (2005). Phase Unwrapping via Graph Cuts. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. S. Marques, N. Pérez de la Blanca, et P. Pina (Eds.) *Pattern Recognition and Image Analysis*, vol. 3522, (pp. 360–367). Berlin, Heidelberg : Springer Berlin Heidelberg.

Bonchi, F., et Lucchese, C. (2005). Pushing tougher constraints in frequent pattern mining. In *Advances in Knowledge Discovery and Data Mining*, (pp. 114–124). Springer.

Bonchi, F., van Leeuwen, M., et Ukkonen, A. (2011). Characterizing uncertain data using compression. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, (pp. 534–545). SIAM.

Burdick, D., Calimlim, M., Flannick, J., Gehrke, J., et Yiu, T. (2005). MAFIA : a maximal frequent itemset algorithm. *IEEE Transactions on Knowledge and Data Engineering*, *17*(11), 1490–1504.

Calders, T., Garboni, C., et Goethals, B. (2010). Efficient pattern mining of uncertain data with sampling. *Advances in Knowledge Discovery and Data Mining*, (pp. 480–487).

Casu, F., Manconi, A., Pepe, A., et Lanari, R. (2011). Deformation Time-Series Generation in Areas Characterized by Large Displacement Dynamics : The SAR Amplitude Pixel-Offset SBAS Technique. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(7), 2752–2763.

Chadwick Jr, W. W., Geist, D. J., Jónsson, S., Poland, M., Johnson, D. J., et Meertens, C. M. (2006). A volcano bursting at the seams : Inflation, faulting, and eruption at Sierra Negra volcano, Galapagos. *Geology*, *34*(12), 1025–1028.

Chao, C. F., Chen, K. S., et Lee, J. S. (2013). Refined Filtering of Interferometric Phase From InSAR Data. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(12), 5315–5323.

Chen, C. W., et Zebker, H. A. (2000). Network approaches to two-dimensional phase unwrapping : intractability and two new algorithms. *JOSA A*, *17*(3), 401–414.

Chen, C. W., et Zebker, H. A. (2002). Phase unwrapping for large SAR interferograms : Statistical segmentation and generalized network models. *IEEE Transactions on Geoscience and Remote Sensing*, (pp. 1709–1719).

Chen, E., Cao, H., Li, Q., et Qian, T. (2008). Efficient strategies for tough aggregate constraint-based sequential pattern mining. *Information Sciences*, *178*(6), 1498–1518.

Christy, S. (1998). Localisation et modélisation tridimensionnelles par approximations successives du modèle perspectif de caméra. (p. 169).

Cobb, G. W., et Chen, Y.-P. (2003). An Application of Markov Chain Monte Carlo to Community Ecology. *The American Mathematical Monthly*, *110*(4), pp. 265–288.

Cormode, G., Li, F., et Yi, K. (2009). Semantics of Ranking Queries for Probabilistic Data and Expected Ranks. In *2009 IEEE 25th International Conference on Data Engineering*, (pp. 305–316).

Cover, T. M., et Thomas, J. A. (2006). *Elements of information theory*. Hoboken, N.J : Wiley-Interscience, 2nd ed.

Crémilleux, B., Plantevit, M., et Soulet, A. (2016). Preference-based pattern mining. Tutorial presented at European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2016.

Dehecq, A., Gourmelen, N., et Trouvé, E. (2015). Deriving large-scale glacier velocities from a complete satellite archive : Application to the Pamir–Karakoram–Himalaya. *Remote Sensing of Environment*, *162*, 55–66.

Derauw, D. (1999). DInSAR and Coherence Tracking Applied to Glaciology : The Example of Shirase Glacier. *99*, 8.

Doin, M.-P., Lasserre, C., Peltzer, G., Cavalié, O., et Doubre, C. (2009). Corrections of stratified tropospheric delays in SAR interferometry : Validation with global atmospheric models. *Journal of Applied Geophysics*, *69*(1), 35–50.

Eppler, J., et Rabus, B. (2012). Monitoring Urban Infrastructure With An Adaptive Multi-looking INSAR Technique. vol. 697, (p. 68).

Fallourd, R. (2012). *Suivi des glaciers alpins par combinaison d'informations hétérogènes : images SAR Haute Résolution et mesures terrain*. Ph.D. thesis, Université de Grenoble.

Fallourd, R., Harant, O., Trouvé, E., Nicolas, J. M., Gay, M., Walpersdorf, A., Mugnier, J. L., Serafini, J., Rosu, D., Bombrun, L., Vasile, G., Cotte, N., Vernier, F., Tupin, F., Moreau, L., et Bolon, P. (2011). Monitoring Temperate Glacier Displacement by Multi-Temporal TerraSAR-X Images and Continuous GPS Measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *4*(2), 372–386.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., et others (1996). Knowledge Discovery and Data Mining : Towards a Unifying Framework. In *Discovery and Data Mining*, vol. 96, (pp. 82–88).

Fernando, B., Fromont, E., et Tuytelaars, T. (2014). Mining Mid-level Features for Image Classification. *International Journal of Computer Vision*, *108*(3), 186–203.

Ferraioli, G., Shabou, A., Tupin, F., et Pascazio, V. (2009). Multichannel Phase Unwrapping With Graph Cuts. *IEEE Geoscience and Remote Sensing Letters*, *6*(3), 562–566.

Ferretti, A., Fumagalli, A., Novali, F., Prati, C., Rocca, F., et Rucci, A. (2011). A New Algorithm for Processing Interferometric Data-Stacks : SqueeSAR. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(9), 3460–3470.

Forkuor, G., Conrad, C., Thiel, M., Ullmann, T., et Zoungrana, E. (2014). Integration of Optical and Synthetic Aperture Radar Imagery for Improving Crop Mapping in Northwestern Benin, West Africa. *Remote Sensing*, *6*(7), 6472–6499.

Foster, J., Brooks, B., Cherubini, T., Shacat, C., Businger, S., et Werner, C. L. (2006). Mitigating atmospheric noise for InSAR using a high resolution weather model. *Geophysical Research Letters*, *33*(16).

Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., et Thomas, R. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition*, *1*(1), 54–77.

García-Hernández, R. A., Martínez-Trinidad, J. F., et Carrasco-Ochoa, J. A. (2006). A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. In *Computational Linguistics and Intelligent Text Processing*, (pp. 514–523). Springer, Berlin, Heidelberg.

Garofalakis, M. N., Rastogi, R., et Shim, K. (1999). SPIRIT : Sequential pattern mining with regular expression constraints. In *VLDB*, vol. 99, (pp. 7–10).

Gatelli, F., Guamieri, A. M., Parizzi, F., Pasquali, P., Prati, C., et Rocca, F. (1994). The wavenumber shift in SAR interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, *32*(4), 855–865.

Ge, J., Xia, Y., et Wang, J. (2015). Towards Efficient Sequential Pattern Mining in Temporal Uncertain Databases. In T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, D. Cheung, et H. Motoda (Eds.) *Advances in Knowledge Discovery and Data Mining*, vol. 9078, (pp. 268–279). Cham : Springer International Publishing.

Ge, J., Xia, Y., Wang, J., Nadungodage, C. H., et Prabhakar, S. (2017). Sequential pattern mining in databases with temporal uncertainty. *Knowledge and Information Systems*, *51*(3), 821–850.

Geerts, F., Goethals, B., et Mielikäinen, T. (2004). Tiling Databases. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, E. Suzuki, et S. Arikawa (Eds.) *Discovery Science*, vol. 3245, (pp. 278–289). Berlin, Heidelberg : Springer Berlin Heidelberg.

Ghiglia, D. C., et Romero, L. A. (1994). Robust two-dimensional weighted and unweighted phase unwrapping that uses fast transforms and iterative methods. *Journal of the Optical Society of America A*, *11*(1), 107.

Giles, A. B., Massom, R. A., et Warner, R. C. (2009). A method for sub-pixel scale feature-tracking using Radarsat images applied to the Mertz Glacier Tongue, East Antarctica. *Remote Sensing of Environment*, *113*(8), 1691–1699.

Gionis, A., Mannila, H., Mielikäinen, T., et Tsaparas, P. (2007). Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(3), 14.

Goldstein, R. M., et Werner, C. L. (1998). Radar interferogram filtering for geophysical applications. *Geophysical Research Letters*, *25*(21), 4035–4038.

Good, P. (2000). *Permutation tests : a practical guide to resampling methods for testing hypotheses*. Springer series in statistics. Springer.

Gouda, K., et Zaki, M. J. (2001). Efficiently mining maximal frequent itemsets. In *Proceedings 2001 IEEE International Conference on Data Mining*, (pp. 163–170).

Goyal, V., Sureka, A., et Patel, D. (2008). Efficient Skyline Itemsets Mining. (pp. 119–124). ACM Press.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press. Google-Books-ID : mbU6T7oUrBgC.

Gueguen, L., et Datcu, M. (2007). Image Time-Series Data Mining Based on the Information-Bottleneck Principle. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(4), 827–838.

Guns, T., Nijssen, S., et Raedt, L. D. (2013). k-Pattern Set Mining under Constraints. *IEEE Transactions on Knowledge & Data Engineering*, *25*(2), 402–418.

Guyet, T., et Nicolas, H. (2016). Long term analysis of time series of satellite images. *Pattern Recognition Letters*, *70*, 17 – 23.

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., et Hsu, M.-C. (2000). FreeSpan : Frequent Pattern-projected Sequential Pattern Mining. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, (pp. 355–359). New York, NY, USA : ACM.

Hanssen, R. F. (2001). *Radar Interferometry*, vol. 2 of *Remote Sensing and Digital Image Processing*. Dordrecht : Springer Netherlands.

Harant, O., Bombrun, L., Vasile, G., Ferro-Famil, L., et Gay, M. (2011). Displacement Estimation by Maximum-Likelihood Texture Tracking. *IEEE Journal of Selected Topics in Signal Processing*, *5*(3), 398–407.

Heas, P., et Datcu, M. (2005). Modeling trajectory of dynamic clusters in image time-series for spatio-temporal reasoning. *IEEE Transactions on Geoscience and Remote Sensing*, *43*(7), 1635–1647.

Heid, T., et Kääb, A. (2012). Evaluation of existing image matching methods for deriving glacier surface displacements globally from optical satellite imagery. *Remote Sensing of Environment*, *118*, 339–355.

Hooper, A. (2008). A multi-temporal InSAR method incorporating both persistent scatterer and small baseline approaches. *Geophysical Research Letters*, *35*(16).

Hooper, A., Bekaert, D., Spaans, K., et Arıkan, M. (2012). Recent advances in SAR interferometry time series analysis for measuring crustal deformation. *Tectonophysics*, *514-517*, 1–13.

Hooper, A., Segall, P., et Zebker, H. (2007). Persistent scatterer interferometric synthetic aperture radar for crustal deformation analysis, with application to Volcán Alcedo, Galápagos. *Journal of Geophysical Research : Solid Earth*, *112*(B7).

Hooper, A., et Zebker, H. A. (2007). Phase unwrapping in three dimensions with application to InSAR time series. *JOSA A*, *24*(9), 2737–2747.

Hooper Andrew, Zebker Howard, Segall Paul, et Kampes Bert (2004). A new method for measuring deformation on volcanoes and other natural terrains using InSAR persistent scatterers. *Geophysical Research Letters*, *31*(23).

Hooshsadat, M., Bayat, S., Naeimi, P., Mirian, M. S., et ZaïAne, O. R. (2012). Uapriori : an algorithm for finding sequential patterns in probabilistic data. In *Uncertainty Modeling in Knowledge Engineering and Decision Making*, vol. 7, (pp. 907–912). World Scientific.

Huss, M., Bauder, A., Werder, M., Funk, M., et Hock, R. (2007). Glacier-dammed lake outburst events of Gornersee, Switzerland. *Journal of Glaciology*, *53*(181), 189–200.

Ibrahim, A., Sastry, S., et Sastry, P. S. (2016). Discovering compressing serial episodes from event sequences. *Knowledge and Information Systems*, *47*(2), 405–432.

Ienco, D., Gaetano, R., Dupaquier, C., et Maurel, P. (2017). Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, *14*(10), 1685–1689.

Jolivet, R., Agram, P. S., Lin, N. Y., Simons, M., Doin, M.-P., Peltzer, G., et Li, Z. (2014). Improving InSAR geodesy using Global Atmospheric Models. *Journal of Geophysical Research : Solid Earth*, *119*(3), 2324–2341.

Jolivet, R., Grandin, R., Lasserre, C., Doin, M.-P., et Peltzer, G. (2011). Systematic InSAR tropospheric phase delay corrections from global meteorological reanalysis data. *Geophysical Research Letters*, *38*(17).

Julea, A., Méger, N., Bolon, P., Rigotti, C., Doin, M.-P., Lasserre, C., Trouvé, E., et Lazarescu, V. N. (2011). Unsupervised Spatiotemporal Mining of Satellite Image Time Series Using Grouped Frequent Sequential Patterns. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(4), 1417–1430.

Julea, A., Méger, N., Rigotti, C., Trouvé, E., Jolivet, R., Bolon, P., et others (2012). Efficient Spatio-temporal Mining of Satellite Image Time Series for Agricultural Monitoring. *Trans. MLDM*, *5*(1), 23–44.

Kazarinoff, N. D. (1961). *Analytic inequalities*. Courier Corporation.

Ketterlin, A., et Gançarski, P. (2007). Sequence similarity and multi-date image segmentation. In *Analysis of Multi-temporal Remote Sensing Images, 2007. MultiTemp 2007. International Workshop on the*, (pp. 1–4). IEEE.

Khiali, L., Ienco, D., et Teisseire, M. (2018). Object-oriented satellite image time series analysis using a graph-based representation. *Ecological Informatics*, *43*, 52 – 64.

Kullback, S., et Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lam, H. T., Mörchen, F., Fradkin, D., et Calders, T. (2012). Mining compressing sequential patterns. In *SDM*, (pp. 319–330).

Lam, H. T., Mörchen, F., Fradkin, D., et Calders, T. (2014). Mining Compressing Sequential Patterns. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, *7*(1), 34–52.

Lanari, R., Mora, O., Manunta, M., Mallorqui, J. J., Berardino, P., et Sansosti, E. (2004). A small-baseline approach for investigating deformations on full-resolution differential SAR interferograms. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(7), 1377–1386.

Leung, C. K.-S. (2011). Mining uncertain data. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, *1*(4), 316–329.

Leung, C. K. S., et Hao, B. (2009). Mining of Frequent Itemsets from Streams of Uncertain Data. In *2009 IEEE 25th International Conference on Data Engineering*, (pp. 1663–1670).

Leung, C. K.-S., et Jiang, F. (2011). Frequent itemset mining of uncertain data streams using the damped window model. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, (pp. 950–955). ACM.

Leung, C. K.-S., Mateo, M. A. F., et Brajczuk, D. A. (2008). A tree-based approach for frequent pattern mining from uncertain data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, (pp. 653–661). Springer.

Leung, C.-S., MacKinnon, R., et Jiang, F. (2014). Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data. In *2014 IEEE International Congress on Big Data (BigData Congress)*, (pp. 315–322).

Li, Z., Fielding, E. J., Cross, P., et Preusker, R. (2009). Advanced InSAR atmospheric correction : MERIS/MODIS combination and stacked water vapour models. *International Journal of Remote Sensing*, *30*(13), 3343–3363.

Lofgren, J. S., Bjorndahl, F., Moore, A. W., Webb, F. H., Fielding, E. J., et Fishbein, E. F. (2010). Tropospheric correction for InSAR using interpolated ECMWF data and GPS Zenith Total Delay from the Southern California Integrated GPS Network. (pp. 4503–4506). IEEE.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, (pp. 1150–1157 vol.2).

Luo, C., et Chung, S. M. (2004). A scalable algorithm for mining maximal frequent sequences using sampling. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, (pp. 156–165). IEEE.

Mannila, H., Toivonen, H., et Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. *Data mining and knowledge discovery*, *1*(3), 259–289.

Masseglia, F., Cathala, F., et Poncelet, P. (1998). The PSP approach for mining sequential patterns. In *Principles of Data Mining and Knowledge Discovery*, (pp. 176–184). Springer.

Massonnet, D., et Feigl, K. L. (1998). Radar interferometry and its application to changes in the Earth's surface. *Reviews of Geophysics*, *36*(4), 441–500.

Massonnet, D., et Souyris, J.-C. (2008). *Imaging with Synthetic Aperture Radar*. CRC Press.

Mauro, N. D., Vergari, A., Basile, T. M. A., Ventola, F. G., et Esposito, F. (2017). End-to-end Learning of Deep Spatio-temporal Representations for Satellite Image Time Series Classification. In *DC@PKDD/ECML*.

Méger, N., Jolivet, R., Lasserre, C., Trouvé, E., Rigotti, C., Lodge, F., Doin, M., Guillaso, S., Julea, A., et Bolon, P. (2011). Spatio-Temporal Mining of ENVISAT SAR Interferogram Time Series over the Haiyuan Fault in China. In *2011 6th International Workshop on the Analysis of Multi-temporal Remote Sensing Images (Multi-Temp)*, (pp. 10.1109/Multi–Temp.2011.6005067). Trento, Italy.

Méger, N., Rigotti, C., et Pothier, C. (2015). Swap Randomization of Bases of Sequences for Mining Satellite Image Times Series. In *Machine Learning and Knowledge Discovery in Databases*, (pp. 190–205). Springer.

Muzammal, M., et Raman, R. (2010). On Probabilistic Models for Uncertain Sequential Pattern Mining. In L. Cao, Y. Feng, et J. Zhong (Eds.) *Advanced Data Mining and Applications*, vol. 6440, (pp. 60–72). Berlin, Heidelberg : Springer Berlin Heidelberg.

Muzammal, M., et Raman, R. (2011). Mining sequential patterns from probabilistic databases. *Advances in Knowledge Discovery and Data Mining*, (pp. 210–221).

Muzammal, M., et Raman, R. (2015). Mining sequential patterns from probabilistic databases. *Knowledge and Information Systems*, *44*(2), 325–358.

Nakamura, K., Doi, K., et Shibuya, K. (2007). Estimation of seasonal changes in the flow of Shirase Glacier using JERS-1/SAR image correlation. *Polar Science*, *1*(2-4), 73–83.

Ng, R. T., Lakshmanan, L. V. S., Han, J., et Pang, A. (1998). Exploratory Mining and Pruning Optimizations of Constrained Associations Rules. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, (pp. 13–24). New York, NY, USA : ACM.

Nguyen, T., Méger, N., Rigotti, C., Pothier, C., Trouvé, E., et Gourmelen, N. (2017). Handling coherence measures of displacement field time series : Application to Greenland ice sheet glaciers. In *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, (pp. 1–4).

Nicolas, J.-M., Trouvé, E., Fallourd, R., Vernier, F., Tupin, F., Harant, O., Gay, M., et Moreau, L. (2012). A first comparison of Cosmo-Skymed and TerraSAR-X data over Chamonix Mont-Blanc test-site. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, (pp. 5586–5589). IEEE.

NunaGIS (2018). Asiaq's maps : Topographic map 1 :500000 and 1 :250000. `http://nunagis.gl/en/`. Consulté le : 01/07/2018.

Pasquier, N., Bastide, Y., Taouil, R., et Lakhal, L. (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems*, *24*(1), 25–46.

Pathier, E., Fielding, E. J., Wright, T. J., Walker, R., Parsons, B. E., et Hensley, S. (2006). Displacement field and slip distribution of the 2005 Kashmir earthquake from SAR imagery. *Geophysical Research Letters*, *33*(20).

Pei, J., Han, J., et Lakshmanan, L. V. (2004a). Pushing convertible constraints in frequent itemset mining. *Data Mining and Knowledge Discovery*, *8*(3), 227–252.

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., et Hsu, M.-C. (2004b). Mining sequential patterns by pattern-growth : The prefixspan approach. *Knowledge and Data Engineering, IEEE Transactions on*, *16*(11), 1424–1440.

Pei, J., Han, J., et Wang, W. (2007). Constraint-based sequential pattern mining : the pattern-growth methods. *Journal of Intelligent Information Systems*, *28*(2), 133–160.

Pericault, Y., Pothier, C., Méger, N., Rigotti, C., Vernier, F., Pham, H. T., et Trouvé, E. (2015). A swap randomization approach for mining motion field time series over the Argentiere glacier. In *2015 8th International Workshop on the Analysis of Multitemporal Remote Sensing Images (Multi-Temp)*, (pp. 1–4).

Petitjean, F., Inglada, J., et Gancarski, P. (2012). Satellite Image Time Series Analysis Under Time Warping. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(8), 3081–3095.

Ponton, F., Trouvé, E., Gay, M., Walpersdorf, A., Fallourd, R., Nicolas, J.-M., Vernier, F., et Mugnier, J.-L. (2014). Observation of the Argentière Glacier Flow Variability from 2009 to 2011 by TerraSAR-X and GPS Displacement Measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(8), 3274–3284.

Quincey, D. J., Glasser, N. F., Cook, S. J., et Luckman, A. (2015). Heterogeneity in Karakoram glacier surges. *Journal of Geophysical Research : Earth Surface*, *120*(7), 1288–1300.

Raissi, C., Poncelet, P., et Teisseire, M. (2006). Speed : Mining maximal sequential patterns over data streams. In *International Conference on Intelligent Systems-ICIS*, (pp. 1–8). IEEE.

Raucoules, D., de Michele, M., Malet, J. P., et Ulrich, P. (2013). Time-variable 3D ground displacements from high-resolution synthetic aperture radar (SAR). application to La Valette landslide (South French Alps). *Remote Sensing of Environment*, *139*, 198–204.

Reiche, J., Verbesselt, J., Hoekman, D., et Herold, M. (2015). Fusing Landsat and SAR time series to detect deforestation in the tropics. *Remote Sensing of Environment*, *156*, 276–293.

Rényi, A., et al. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Contributions to the Theory of Statistics*. The Regents of the University of California.

Rigotti, C., Lodge, F., Méger, N., Pothier, C., Jolivet, R., et Lasserre, C. (2014). Monitoring of Tectonic Deformation by Mining Satellite Image Time Series. In *Reconnaissance de Formes et Intelligence Artificielle (RFIA) 2014*, (p. 6). Rouen, France.

Romani, L. A. S., de Avila, A. M. H., Chino, D. Y. T., Zullo, J., Chbeir, R., Traina, C., et Traina, A. J. M. (2013). A New Time Series Mining Approach Applied to Multitemporal Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(1), 140–150.

Rosenau, R., Scheinert, M., et Dietrich, R. (2015). A processing system to monitor Greenland outlet glacier velocity variations at decadal and seasonal time scales utilizing the Landsat imagery. *Remote Sensing of Environment*, *169*, 1–19.

Scambos, T. A., Dutkiewicz, M. J., Wilson, J. C., et Bindschadler, R. A. (1992). Application of image cross-correlation to the measurement of glacier velocity using satellite image data. *Remote Sensing of Environment*, *42*(3), 177–186.

Scherler, D., Leprince, S., et Strecker, M. (2008). Glacier-surface velocities in alpine terrain from optical satellite imagery—Accuracy improvement and quality assessment. *Remote Sensing of Environment*, *112*(10), 3806–3819.

Singleton, A., Li, Z., Hoey, T., et Muller, J.-P. (2014). Evaluating sub-pixel offset techniques as an alternative to D-InSAR for monitoring episodic landslide movements in vegetated terrain. *Remote Sensing of Environment*, *147*, 133–144.

Soulet, A., Raïssi, C., Plantevit, M., et Cremilleux, B. (2011). Mining Dominant Patterns in the Sky. (pp. 655–664). IEEE.

Srikant, R., et Agrawal, R. (1996). *Mining sequential patterns : Generalizations and performance improvements*. Springer.

Strozzi, T., Kouraev, A., Wiesmann, A., Wegmüller, U., Sharov, A., et Werner, C. (2008). Estimation of Arctic glacier motion with satellite L-band SAR data. *Remote Sensing of Environment*, *112*(3), 636–645.

Strozzi, T., Luckman, A., Murray, T., Wegmuller, U., et Werner, C. L. (2002). Glacier motion estimation using SAR offset-tracking procedures. *IEEE Transactions on Geoscience and Remote Sensing*, *40*(11), 2384–2391.

Suciu, D., et Dalvi, N. (2005). Foundations of probabilistic answers to queries. (p. 963). ACM Press.

Sun, L., Cheng, R., Cheung, D. W., et Cheng, J. (2010). Mining uncertain data with probabilistic guarantees. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 273–282). ACM.

Suo, Z., Li, Z., et Bao, Z. (2010). A new strategy to estimate local fringe frequencies for InSAR phase noise reduction. *IEEE Geoscience and Remote Sensing Letters*, *7*(4), 771–775.

Tatti, N., et Vreeken, J. (2012). The long and the short of it : summarising event sequences with serial episodes. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 462–470). ACM.

Tedstone, A. J., Nienow, P. W., Gourmelen, N., Dehecq, A., Goldberg, D., et Hanna, E. (2015). Decadal slowdown of a land-terminating sector of the Greenland Ice Sheet despite warming. *Nature*, *526*(7575), 692–695.

Tobita, M., Murakami, M., Nakagawa, H., Yarai, H., Fujiwara, S., et Rosen, P. A. (2001). 3-D surface deformation of the 2000 Usu Eruption measured by matching of SAR images. *Geophysical Research Letters*, *28*(22), 4291–4294.

Toivonen, H. (1996). Sampling Large Databases for Association Rules. In *Proceedings of the 22th International Conference on Very Large Data Bases*, VLDB '96, (pp. 134–145). San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Trouvé, E., Nicolas, J. M., et Maitre, H. (1998). Improving phase unwrapping techniques by the use of local frequency estimates. *IEEE Transactions on Geoscience and Remote Sensing*, *36*(6), 1963–1972.

Trouvé, E., Vasile, G., Gay, M., Bombrun, L., Grussenmeyer, P., Landes, T., Nicolas, J.-M., Bolon, P., Petillot, I., Julea, A., Valet, L., Chanussot, J., et Koehl, M. (2007). Combining Airborne Photographs and Spaceborne SAR Data to Monitor Temperate Glaciers : Potentials and Limits. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(4), 905–924.

Ugarte, W., Boizumault, P., Loudni, S., Crémilleux, B., et Lepailleur, A. (2015). Soft constraints for pattern mining. *Journal of Intelligent Information Systems*, *44*(2), 193–221.

Vaglio Laurin, G., Liesenberg, V., Chen, Q., Guerriero, L., Del Frate, F., Bartolini, A., Coomes, D., Wilebore, B., Lindsell, J., et Valentini, R. (2013). Optical and SAR sensor synergies for forest and land cover mapping in a tropical site in West Africa. *International Journal of Applied Earth Observation and Geoinformation*, *21*, 7–16.

Van Leeuwen, M. (2014). Interactive Data Exploration Using Pattern Mining. In A. Holzinger, et I. Jurisica (Eds.) *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics : State-of-the-Art and Future Challenges*, (pp. 169–182). Berlin, Heidelberg : Springer Berlin Heidelberg.

Vernier, F., Fallourd, R., Friedt, J. M., Yan, Y., Trouvé, E., Nicolas, J.-M., et Moreau, L. (2011). Fast correlation technique for glacier flow monitoring by digital camera and spaceborne SAR images. *EURASIP Journal on Image and Video Processing*, *2011*(1), 1–15.

Vincent, C., et Moreau, L. (2016). Sliding velocity fluctuations and subglacial hydrology over the last two decades on Argentière glacier, Mont Blanc area. *Journal of Glaciology*, *62*(235), 805–815.

Vreeken, J., van Leeuwen, M., et Siebes, A. (2011). Krimp : mining itemsets that compress. *Data Mining and Knowledge Discovery*, *23*(1), 169–214.

Wan, L., Chen, L., et Zhang, C. (2013). Mining frequent serial episodes over uncertain sequence data. (p. 215). ACM Press.

Wang, J., Han, J., et Li, C. (2007). Frequent Closed Sequence Mining without Candidate Maintenance. *IEEE Transactions on Knowledge and Data Engineering*, *19*(8), 1042–1056.

Werner, C., Wegmüller, U., Strozzi, T., et Wiesmann, A. (2005). Precision estimation of local offsets between pairs of SAR SLCs and detected SAR images. In *Geoscience and Remote Sensing Symposium, 2005. IGARSS'05. Proceedings. 2005 IEEE International*, vol. 7, (pp. 4803–4805). IEEE.

Yan, X., Han, J., et Afshar, R. (2003). CloSpan : Mining : Closed Sequential Patterns in Large Datasets. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, Proceedings, (pp. 166–177). Society for Industrial and Applied Mathematics.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353.

Zaki, M. J. (2000). Sequence mining in categorical domains : incorporating constraints. In *Proceedings of the ninth international conference on Information and knowledge management*, (pp. 422–429). ACM.

Zaki, M. J. (2001). SPADE : An efficient algorithm for mining frequent sequences. *Machine learning*, *42*(1-2), 31–60.

Zebker, H. A., et Villasenor, J. (1992). Decorrelation in interferometric radar echoes. *IEEE Transactions on Geoscience and Remote Sensing*, *30*(5), 950–959.

Zhang, Q., Li, F., et Yi, K. (2008). Finding frequent items in probabilistic data. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, (pp. 819–832). ACM.

Zhao, Z., Yan, D., et Ng, W. (2012). Mining probabilistically frequent sequential patterns in uncertain databases. In *Proceedings of the 15th international conference on extending database technology*, (pp. 74–85). ACM.

Zhenhong Li, Fielding, E., et Cross, P. (2009). Integration of InSAR Time-Series Analysis and Water-Vapor Correction for Mapping Postseismic Motion After the 2003 Bam (Iran) Earthquake. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(9), 3220–3230.

Zhou Zhao, Da Yan, et Ng, W. (2014). Mining Probabilistically Frequent Sequential Patterns in Large Uncertain Databases. *IEEE Transactions on Knowledge and Data Engineering*, *26*(5), 1171–1184.

Annexe A

Cette annexe présente une méthode de classement des motifs Séquentiels Fréquents Groupés (SFG) (cf. Chapitre 4). Ce classement, proposé par Méger *et al.* (2015), utilise les cartes de Localisation Spatio-Temporelle (LST). Une carte LST permet de visualiser l'emplacement des occurrences d'un motif SFG dans l'espace et dans le temps. Chacun des pixels affecté par le motif contient en effet la date de fin au plus tôt de la première occurrence du motif observée pour la zone géographique correspondante. Une valeur spéciale telle que 0 est attribuée aux pixels qui ne sont pas affectés par le motif. Les valeurs portées par les pixels sont généralement associées à une échelle de couleur afin que les cartes puissent être interprétées par les experts. Les motifs SFG les plus représentatifs sont sélectionnés en utilisant une technique de *swap* randomisation et un classement basé sur la mesure d'Information Mutuelle Normalisée (en anglais *Normalized Mutual Information*) (NMI).

A.1 *Swap* randomisation et classement des motifs SFG

Selon la Série Temporelle de Champs de Déplacements (STCD) considérée, il est possible d'obtenir de nombreux motifs SFG pour des valeurs standard de support minimal et de connexité minimale. La localisation spatiotemporelle de chaque motif extrait est, pour rappel, donnée par sa carte LST. Afin de mettre en avant les motifs les plus prometteurs, Méger *et al.* (2015) ont proposé de sélectionner un ensemble de cartes LST représentatives. À cette fin, deux types de cartes LST sont considérés :

— Les cartes dont l'information diffère de ce qui pourrait être observé en considérant une STCD randomisée ayant la même structure spatio-temporelle en termes de la fréquence des symboles.
— Les cartes dont l'information est similaire à ce qui pourrait être observé en considérant la même STCD randomisée.

La STCD randomisée est obtenue par une technique de *swap* randomisation, présentée dans la section A.1.1. Les cartes LST sont ensuite comparées avec celles obtenues en utilisant la série randomisée. Cette comparaison est basée sur la mesure NMI. Elle permet de produire un classement à partir duquel il est possible de sélectionner les deux types souhaités de cartes. La construction de ce classement est présenté en section A.1.1.

A.1.1 *Swap* randomisation de la série

De nombreux travaux tels que ceux de Good (2000) ont montré l'intérêt de la randomisation pour les tests d'hypothèse. En ce qui concerne la *swap* randomisation, elle est principalement appliquée aux matrices booléennes (e.g., Cobb et Chen (2003); Gionis *et al.* (2007)). Le principe repose sur une comparaison entre les résultats obtenus sur un jeu de données et ceux obtenus sur un autre jeu de données randomisé ayant la même structure en termes de marges en ligne et en colonne. Autrement dit, nous ne nous intéressons pas à l'information apportée par ces marges, mais la disposition du jeu de données lui-même. Cette disposition peut être exprimée par des motifs tels que des rectangles de symboles 1. Méger *et al.* (2015) ont adapté cette approche au cadre des matrices symboliques et ont montré que les jeux de données randomisés sont toujours équiprobables. Soit S la STCD symbolique contenant pour chaque pixel et chaque date un symbole décrivant le déplacement observé pour une zone géographique. Soit S' la série randomisée. Elle est générée à partir de S en appliquant une série d'échanges de symboles, chaque échange étant appliqué à la matrice obtenue par les échanges précédents. Considérons la STCD comme une matrice symbolique de taille $m \times n$ dont chaque ligne correspond à la position d'un pixel et chaque colonne indique une date. Soit D une telle matrice symbolique. Soit u et v les positions de deux pixels choisis au hasard. Soit i et j deux dates également choisies au hasard. Si $D_{u,i} = D_{v,j} = \alpha$ et $D_{u,j} = D_{v,i} = \beta$, avec α et β deux symboles distincts, alors ceux-ci sont modifiés de sorte que $D_{u,i} = D_{v,j} = \beta$ and $D_{u,j} = D_{v,i} = \alpha$ (i.e., les symboles α et β sont échangés). Dans le cadre des STCD, ces échanges peuvent être interprétés comme des échanges spatiotemporels. Par construction, comme pour les matrices booléennes, la fréquence des symboles est maintenue à la fois temporellement (en colonne) et spatialement (en ligne). Comme proposé dans Cobb et Chen (2003); Gionis *et al.* (2007), la procédure de *swap* randomisation a les caractéristiques suivantes :

— Toutes les positions des pixels et toutes les dates ont la même probabilité d'être choisies, et peuvent être choisies plusieurs fois.
— Les échanges peuvent être annulés par d'autres échanges.

En pratique, le nombre minimal d'échanges à appliquer doit être de l'ordre de 10 fois le nombre de données contenues dans la série.

A.1.2 Classement des cartes LST par mesure NMI

Après la *swap* randomisation, pour chacun des motifs extrait, sa carte LST, notée C, construite dans les données d'origine est comparée avec celle construite dans les données randomisées, notée C'. Cette comparaison permettrait de révéler si :

— C et C' partagent peu d'informations, i.e., le contenu informationnel de C est singulier car il ne peut pas être obtenu pour un jeu de données randomisé de même structure en termes de fréquences des symboles,
— ou C et C' partagent beaucoup d'informations, i.e., la *swap* randomisation n'arrive pas à détruire les occurrences du motif extrait, ce qui signifie qu'il exprime des phénomènes importants.

Soit Ω l'univers contenant toutes les dates de fin des occurrences. Considérons chaque date de fin x d'une occurrence présente dans la carte C comme la réalisation d'une variable

aléatoire discrète X et chaque date de fin y d'une occurrence présente dans la carte C' comme la réalisation d'une variable aléatoire discrète Y. Afin d'évaluer le contenu d'information partagé par X et Y, avec des valeurs dans l'intervalle $[0, 1]$, Méger *et al.* (2015) ont proposé d'utiliser une version normalisée de la mesure d'information mutuelle (Cover et Thomas, 2006), appelée Information Mutuelle Normalisée (en anglais *Normalized Mutual Information*) (NMI) :

$$NMI(X,Y) = \frac{\sum_{x,y\in\Omega^2} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}}{min(H(X), H(Y))} \tag{A.1}$$

où $P(x, y)$ représente la probabilité de cooccurrence des deux dates de fin x et y à la même position, dans C et C'.

Cette mesure quantifie le contenu d'information partagé par les deux variables aléatoires. En d'autres termes, connaissant la réalisation de la variable aléatoire X, la NMI indique dans quelle mesure la réalisation de la variable Y peut être déduite de celle de X, et vice versa. Par conséquent, elle peut être considérée comme une mesure de la dépendance mutuelle entre X et Y. De plus, en s'appuyant sur la NMI, aucune hypothèse sur la relation entre les dates de fin n'est faite. Au lieu d'utiliser, par exemple, un coefficient de corrélation de Pearson pour vérifier l'existence d'une relation linéaire, les dates de fin sont simplement évaluées en considérant leurs co-occurrences. Pour chacun des motifs extraits, cette mesure NMI permet de classer les motifs (et les cartes associées). Si l'on est intéressé par des motifs SFG montrant des phénomènes qui ne peuvent pas être obtenus sur une série randomisée, il faut alors regarder ceux avec des mesures NMI faibles. Inversement, si l'on est intéressé par des motifs montrant des phénomènes importants qui sont toujours présents dans la série randomisée, alors il faut considérer les motifs avec des scores NMI élevés. Afin de sélectionner un ensemble contenant les motifs les plus représentatifs, Méger *et al.* (2015) considèrent les deux extrémités du classement, i.e., les motifs avec les scores NMI les plus bas et les plus élevés.

Annexe B

Nous allons dans cette annexe examiner la complexité de l'algorithme 2. Pour chaque Occurrence Minimale avec dates Intermédiaires au plus Tôt (OMDIT), un parcours sur tous les événements situés dans sa fenêtre temporelle est effectué pour le calcul du gain informationnel. Pour chaque événement, ce gain est mesuré en combinant la nouvelle contrainte apportée par la connaissance de l'occurrence concernée avec la contrainte actuelle sur cet événement. Ce calcul étant considéré comme une seule opération, il y a par conséquent autant d'opérations que le nombre d'événements parcourus. Soit N le nombre d'événements de chaque séquence [1]. Pour chaque séquence couverte par un motif β, une borne supérieure du nombre d'Occurrences Minimales avec Dates Intermédiaires au plus Tôt (OMDIT) consiste à considérer qu'il y a autant d'occurrences que d'événements, que ces occurrences s'étalent jusqu'à la fin de la séquences, et que le début des occurrence se décale d'un événement l'un après l'autre. Il y aura donc au maximum N occurrences contenant $N, N-1, N-2, \ldots, 1$ événements. Dans ce cas, le nombre d'événements à parcourir est :

$$\begin{aligned} X &= N + N - 1 + N - 2 + \cdots + 1 \\ &= \frac{N \times (N+1)}{2} \end{aligned} \tag{B.1}$$

Pour chaque motif β, le support maximal est le nombre de séquences de la Base de Séquences Probabiliste Partielle (BSPP) $\mathcal{B}$, noté $|\mathcal{B}|$. Soit $|P|$ le nombre de motifs dans l'ensemble P des motifs à sélectionner. Pour sélectionner le meilleur motif à la première itération, nous avons à effectuer, dans le pire des cas le nombre d'opérations suivant :

$$|P| \times |\mathcal{B}| \times \frac{N \times (N+1)}{2} \tag{B.2}$$

Pour la deuxième itération, comme le motif le plus informatif est exclu de l'ensemble P, le nombre maximal d'opérations sera :

$$(|P| - 1) \times |\mathcal{B}| \times \frac{N \times (N+1)}{2} \tag{B.3}$$

Pour la $i^{ième}$ itération :

$$(|P| - i + 1) \times |\mathcal{B}| \times \frac{N \times (N+1)}{2} \tag{B.4}$$

1. Nous considérons que toutes les séquences ont la même taille.

Au total, pour sélectionner K motifs, nous avons dans le pire des cas le nombre d'opérations suivant :

$$\begin{aligned} &\sum_{i=1...K} (|P| - i + 1) \times |\mathcal{B}| \times \frac{N \times (N+1)}{2} \\ &= \frac{(2|P| - K + 1) \times K}{2} \times |\mathcal{B}| \times \frac{N \times (N+1)}{2} \end{aligned} \tag{B.5}$$

Supposons que $K \ll P$, alors $2|P| - K \approx 2|P|$ et une borne supérieure de la complexité dans le pire des cas de l'algorithme est :

$$\mathcal{O}(|P| \times K \times |\mathcal{B}| \times N^2) \tag{B.6}$$

Notre algorithme peut donc être très gourmand en temps de calcul, surtout pour de grands jeux de données, contenant de longues séquences, et avec beaucoup de motifs extraits. Cependant, comme à chaque itération l'information obtenue par les occurrences de chaque motif est indépendante de celle des autres motifs, ce calcul peut être effectué de façon parallèle sur différents cœurs disponibles. De la même manière, il est également possible d'utiliser les *frameworks* comme *MapReduce* sur des *clusters* de calcul pour accélérer le processus.