Résumés linguistiques de données numériques : interprétabilité et périodicité de séries

Gilles Moyse

Résumés linguistiques de données numériques : interprétabilité et périodicité de séries

Thèse de doctorat de l'université de Paris 6

présentée pour obtenir le grade de

Docteur de l'Université Paris 6

(spécialité informatique)

par

Gilles Moyse

Soutenance prévue le 19 juillet 2016

devant le jury composé de

Janusz Kacprzyk	Rapporteur
Trevor Martin	Rapporteur
Bernadette Bouchon-Meunier	Examinatrice
Jean-Gabriel Ganascia	Examinateur
Anne Laurent	Examinatrice
Adrien Revault d'Allonnes	Examinateur
Marie-Jeanne Lesot	Directrice de thèse

Résumé

Nos travaux s'inscrivent dans le domaine des résumés linguistiques flous (RLF) qui permettent la génération de phrases en langage naturel, descriptives de données numériques, et offrent ainsi une vision synthétique et compréhensible de grandes masses d'information.

Nous nous intéressons d'abord à l'interprétabilité des RLF, capitale pour fournir une vision simplement appréhendable de l'information à un utilisateur humain et complexe du fait de sa formulation linguistique. En plus des travaux existant à ce sujet sur les composants élémentaires des RLF, nous proposons une approche globale de l'interprétabilité des résumés vus comme un ensemble de phrases et nous intéressons plus spécifiquement à la question de leur cohérence. Afin de la garantir dans le cadre de la logique floue standard, nous introduisons une formalisation originale de l'opposition entre phrases de complexité croissante. Ce formalisme nous permet de démontrer que les propriétés de cohérence sont vérifiables par le choix d'un modèle de négation spécifique. D'autre part, nous proposons sur cette base un cube en 4 dimensions mettant en relation toutes les oppositions possibles entre les phrases d'un RLF et montrons que ce cube généralise plusieurs structures d'opposition logiques existantes.

Nous considérons ensuite le cas de données sous forme de séries numériques et nous intéressons à des résumés linguistiques portant sur leur périodicité : les phrases que nous proposons indiquent à quel point une série est périodique et proposent une formulation linguistique appropriée de sa période. La méthode d'extraction proposée, nommée DPE pour Detection of Periodic Events, permet de segmenter les données de manière adaptative et sans paramètre utilisateur, en utilisant des outils issus de la morphologie mathématique. Ces segments sont ensuite utilisés pour calculer la période de la série temporelle ainsi que sa périodicité, calculée comme un degré de qualité sur le résultat renvoyé mesurant à quel point la série est périodique. Enfin, DPE génère des phrases comme « Environ toutes les 2 heures, l'afflux de client est important ». Des expériences sur des données artificielles et réelles confirment la pertinence de l'approche.

D'un point de vue algorithmique, nous proposons une implémentation incrémentale et efficace de DPE, basée sur l'établissement de formules permettant le calcul de mises à jour des variables. Cette implémentation permet le passage à l'échelle de la méthode ainsi que l'analyse en temps réel de flux de données.

Nous proposons également une extension de DPE basée sur le concept de périodicité locale permettant d'identifier les sous-séquences périodiques d'une série temporelle par

l'utilisation d'un test statistique original. La méthode, validée sur des données artificielles et réelles, génère des phrases en langage naturel permettant d'extraire des informations du type « Toutes les deux semaines sur le premier semestre de l'année, les ventes sont élevées ».

Mots-clés Résumés, Génération automatique de texte, Logique floue, Big data, Séries numériques, Périodicité, Périodicité locale

Abstract

Our research is in the field of fuzzy linguistic summaries (FLS) that allow to generate natural language sentences to describe very large amounts of numerical data, providing concise and intelligible views of these data.

We first focus on the interpretability of FLS, crucial to provide end-users with an easily understandable text, but hard to achieve due to its linguistic form. Beyond existing works on that topic, based on the basic components of FLS, we propose a general approach for the interpretability of summaries, considering them globally as groups of sentences. We focus more specifically on their consistency. In order to guarantee it in the framework of standard fuzzy logic, we introduce a new model of oppositions between increasingly complex sentences. The model allows us to show that these consistency properties can be satisfied by selecting a specific negation approach. Moreover, based on this model, we design a 4-dimensional cube displaying all the possible oppositions between sentences in a FLS and show that it generalises several existing logical opposition structures.

We then consider the case of data in the form of numerical series and focus on linguistic summaries about their periodicity: the sentences we propose indicate the extent to which the series are periodic and offer an appropriate linguistic expression of their periods. The proposed extraction method, called DPE, standing for Detection of Periodic Events, splits the data in an adaptive manner and without any prior information, using tools from mathematical morphology. The segments are then exploited to compute the period and the periodicity, measuring the quality of the estimation and the extent to which the series is periodic. Lastly, DPE returns descriptive sentences of the form "Approximately every 2 hours, the customer arrival is important". Experiments with artificial and real data show the relevance of the proposed DPE method.

From an algorithmic point of view, we propose an incremental and efficient implementation of DPE, based on established update formulas. This implementation makes DPE scalable and allows it to process real-time streams of data.

We also present an extension of DPE based on the local periodicity concept, allowing the identification of local periodic subsequences in a numerical series, using an original statistical test. The method validated on artificial and real data returns natural language sentences that extract information of the form "Every two weeks during the first semester of the year, sales are high".

Keywords Summaries, Natural Language Generation, Fuzzy logic, Big data, Time series, Periodicity, Local periodicity

Table des matières

Introduction générale

> L'esprit de l'homme a trois clefs qui ouvrent tout : le chiffre, la lettre, la note. Savoir, penser, rêver. Tout est là.
>
> —Victor Hugo, *Les Rayons et les Ombres*

Contexte

Le résumé linguistique de données propose une représentation *textuelle* et *synthétique* de l'information, permettant une interprétation simple en un temps raisonnable pour un utilisateur. Ce type de représentation est d'autant plus utile aujourd'hui que la quantité de données créées explose et que leur analyse exhaustive par des humains uniquement n'est plus envisageable.

Dans sa forme la plus générale, le résumé de données est une représentation condensée d'un type de données vers un autre. Les résumés sont caractérisés par la forme qu'ils prennent et les données qu'ils traitent.

Certains résumés sont de même type que les données qu'ils résument : résumés de texte (Amini et al., 2005), de vidéo (Detyniecki & Marsala, 2007) ou de musique (Peeters et al., 2002), qui renvoient des versions raccourcies des données notamment par sélection des parties identifiées comme importantes. Par extension, une moyenne peut aussi être vue comme le résumé numérique d'une série de données numériques, au même titre que certaines des techniques de réduction de la dimensionalité présentées dans la thèse.

Lorsque le format du résumé est différent de celui des données qu'il traite, il est alors le plus souvent textuel ou linguistique. Différents types de données peuvent être résumés de la sorte : une image au travers de la génération d'une légende descriptive (Vinyals et al., 2015), une pièce musicale par la création d'annotations qui en décrivent les mouvements (Ros et al., 2016), un graphe via des phrases en détaillant certaines relations (Castelltort & Laurent, 2015), des données numériques enfin (Yager, 1982; Kacprzyk & Zadrozny, 2013a,b). Les travaux présentés dans cette thèse concernent deux types de données, décrites sous la forme attribut-valeur et sous la forme de séries.

Il est parfois objecté qu'un résultat graphique est plus adapté : ne dit-on pas qu'une image vaut mieux qu'un long discours ? Cette assertion ne se vérifie pas néanmoins dans

un certain nombre de cas. Ainsi, lorsque la quantité d'information est importante, hétérogène, sur un grand nombre de dimensions ou plus généralement difficile à représenter graphiquement, une phrase peut être plus adaptée pour la synthétiser (Yu et al., 2003). D'autre part, l'interprétation qui est faite de la langue est souvent plus intuitive que celle d'un graphique et permet de prendre des décisions plus rapidement (Law et al., 2005). A l'inverse, une image comme une carte météo nécessite une expertise pour son analyse tandis que le bulletin généré est directement appréhendable (Sripada et al., 2003). Par ailleurs, la phrase peut être lue, par un humain ou par synthèse vocale, la rendant particulièrement adaptée aux situations où l'attention visuelle ne doit pas être dérangée (Arguelles & Triviño, 2013) ou lorsqu'elle est endommagée ou manquante (Thomas et al., 2012). Enfin, cette représentation est préférable dans les cas où l'information ne peut être affichée de manière graphique (e.g. SMS) ou lorsque la bande-passante de la connexion utilisée pour transférer les données est faible (Reiter & Dale, 1997).

L'intérêt de la représentation textuelle d'informations est d'ailleurs reconnue tant par les acteurs académiques, comme le montre le nombre de publications, de sessions spéciales dans différentes conférences et de numéros spéciaux de journaux sur le domaine, que par les acteurs privés qui développent également la recherche en ce sens et la mettent en œuvre dans leur activité. De nombreuses entreprises proposent des services de ce type, la plupart d'entre elles étant issues de la recherche académique : Narrative Science [1], Automated Insights [2], Yseop [3] Arria NLG [4] ou Phedes Lab [5] par exemple.

Différentes disciplines sont également liées aux résumés linguistiques. Le *data storytelling* désigne de manière générale les approches dédiées à la présentation textuelle de données, automatique ou non (Nussbaumer Knaflic, 2015). Le *data to text* est plus spécifiquement lié aux méthodes algorithmiques de conversion automatique de données en texte (Ramos-Soto et al., 2016). Le robot journalisme, porté à la connaissance du grand public à l'occasion notamment de la publication par le site d'informations `lemonde.fr` d'articles générés automatiquement décrivant les résultats des élections départementales de 2015, est un exemple d'application des méthodes de *data to text*.

Si l'intérêt de la production automatique de texte ne fait pas de doute, les modalités de sa mise en œuvre soulèvent un ensemble de questions. Nous nous intéressons dans cette thèse à celles de leur interprétabilité et de leur utilisation pour des données temporelles.

Interprétabilité

L'interprétabilité d'un texte évalue son caractère compréhensible, intelligible et la capacité qu'il a de transmettre une ou plusieurs informations. Ce sujet est central dans la communauté de l'intelligence artificielle et celle du flou en particulier (Gacto et al., 2011;

1. `https://www.narrativescience.com`
2. `https://automatedinsights.com`
3. `http://yseop.com`
4. `http://www.arria.com`
5. `http://phedes.com`

Alonso et al., 2009; Hüllermeier, 2015) qui s'attache notamment à modéliser l'interprétabilité par les objets qu'elle manipule. Le sujet couvre un nombre importants de domaines comme les sciences cognitives, la stylistique ou la linguistique, situés très en dehors du cadre de cette thèse. L'interprétabilité telle que nous l'entendons dans ce travail est celle d'un type particulier de résumés linguistiques, les *résumés linguistiques flous* ou RLF.

Proposés par Yager (1982) puis développés par Kacprzyk et al. (2000), Delgado et al. (2014) ou Novák (2015), les résumés linguistiques flous sont basés sur les principes de la logique floue et permettent par exemple la génération de phrases comme « La plupart des jeunes sont grands » ou « Peu de températures en hiver sont élevées ». Ces phrases sont basés sur des modèles appelés *protoformes* et utilisent des *modalités* de *variables linguistiques* comme *Jeune* ou *Grand* et des *quantificateurs* comme *Peu* ou *La plupart*.

Les variables linguistiques (Zadeh, 1975) et les quantificateurs (Zadeh, 1983) sont des concepts très utilisés de la logique floue qui permettent l'expression d'éléments numériques sous forme de mots selon des définitions intuitives données par l'utilisateur. Ils constituent le vocabulaire de base des RLF.

L'étude de l'interprétabilité de ces résumés passe par celle du vocabulaire sur lequel ils sont construits, avec notamment la question de sa couverture, de sa taille ou du sens qu'il induit pour l'utilisateur. Comment définir la variable linguistique *Taille* par exemple, composée des modalités *Petit*, *Moyen* et *Grand* ? Une personne mesurant moins d'un mètre peut être qualifiée de petite, mais qu'en est-il d'un chat ? De la même manière, dans le cadre de l'étude plus spécifique de la cohérence que nous menons dans cette thèse, peut-on qualifier la taille d'une personne de *Petite* et *Moyenne* ? *Petite* et *Grande* ?

Des questions du même ordre se posent concernant les quantificateurs utilisés, qui servent à donner un décompte des objets vérifiant la propriété étudiée dans la phrase. *La plupart* par exemple, peut-il désigner tous les éléments étudiés ou bien faut-il utiliser *Tous* en ce cas précis ? *Beaucoup* fait probablement référence à plus d'éléments que *La moitié*, mais combien ?

En plus de l'interprétabilité du vocabulaire, nous proposons de manière nouvelle dans cette thèse d'étudier l'interprétabilité des groupes de phrases générées, plus spécifiquement dans le cadre de leur cohérence. Comment interpréter un résumé contenant les deux phrases « Peu de jeunes sont grands » et « Peu de jeunes sont non grands » ? Toutes les données sont-elles couvertes ? Manque-t-il une modalité pour l'évaluation de la taille ? Quid alors de « Peu de jeunes sont grands » et « Peu de jeunes sont petits » ? Le résumé semble plus correct s'il existe une modalité supplémentaire entre *Petit* et *Grand*.

La souplesse offerte par la logique floue dans la définition du vocabulaire, des protoformes et donc des phrases qu'elle permet, est également la source d'un ensemble de questions concernant leur interprétabilité, dont l'étude est au centre de la première partie de cette thèse.

Données temporelles

Dans la seconde partie, nous nous intéressons plus spécifiquement au cas des résumés linguistiques de *données temporelles*, i.e. composées de valeurs numériques associées à des dates. L'étude de ces séries en particulier est motivée par leur nombre en croissance exponentielle notamment liée au développement de l'internet des objets, à la présence toujours plus importante de capteurs dans les objets manufacturés ou encore de l'enregistrement régulier de phénomènes économiques.

De nombreuses propositions ont été réalisées pour traiter ces séries dans le cadre des résumés linguistiques flous, par exemple pour l'analyse de leurs tendances (Kacprzyk et al., 2008) ou celle de leurs propriétés sur des intervalles temporels hiérarchiques (Castillo-Ortega et al., 2011b).

Dans cette thèse, nous nous intéressons plus spécifiquement à leur *périodicité*, c'est-à-dire à leur caractère répétitif, majoritairement ignoré des travaux liés aux résumés linguistiques flous, à l'exception de l'approche de Novák et al. (2008). Une série dont la périodicité est élevée est composée de motifs identiques répétés à intervalles réguliers, dont la longueur mesure la *période*. Un résumé linguistique de périodicité peut être « Tous les jours, les valeurs de la série sont élevées ».

Les problématiques que nous avons identifiées pour la génération de ces phrases se sont moins portées sur leurs aspects linguistiques, plus simples dans ce cas et déjà explorés dans la première partie, que sur le calcul de la période et de la périodicité à proprement parler. Bien que ces questions soient l'objet de nombreux travaux, ces derniers ne tiennent compte la plupart du temps que de la période sans la périodicité et ne proposent pas de méthode de rendu linguistique. Ils sont de plus majoritairement basés sur des modèles a priori ou sur des hypothèses fortes concernant les données.

L'approche linguistique que nous proposons ici permet la génération de phrases décrivant localement ou globalement la période et la périodicité de séries temporelles. Elle se base sur une méthode originale sans paramètres utilisant des outils issus de la morphologique mathématique.

Structure de la thèse

La thèse est structurée en deux parties, la première dédiée à l'interprétabilité des résumés linguistiques flous et la seconde à l'étude de la périodicité dans les séries temporelles. Ces deux parties sont précédées d'une présentation générale sur les résumés linguistiques flous, leurs principes et leurs composantes, détaillés au chapitre 1.

Dans la première partie, le chapitre 2 s'attache à définir la notion de qualité d'un résumé linguistique flou, et plus spécifiquement son interprétabilité. Cette dernière est étudiée à différents niveaux : vocabulaire, phrases, résumé et système de génération.

Le chapitre 3 décrit nos propositions concernant la cohérence des résumés, nécessaire à leur interprétabilité et basée sur les différents types d'opposition pouvant exister entre phrases. Nous analysons en détail ces oppositions et en déduisons un cube en 4 dimensions

les résumant toutes ainsi qu'un grand nombre de structures logiques d'opposition préexistantes. Nous utilisons également cette formalisation pour étendre la validité de certaines propriétés de cohérence à des protoformes étendus.

La seconde partie est dédiée à l'étude de la périodicité dans les séries temporelles. Le chapitre 4 présente un état de l'art des approches existantes de calcul de la période dans les séries temporelles, passant en revue les méthodes temporelles, fréquentielles, temporo-fréquentielles et symboliques.

A la suite de cette étude, nous proposons dans le chapitre 5 la nouvelle méthode DPE destinée à l'évaluation de la période et de la périodicité des séries temporelles, ainsi qu'à la génération d'une phrase la décrivant. Cette méthode est basée sur des approches de morphologie mathématique ainsi que sur une nouvelle transformation, le score d'érosion.

Ce dernier nécessitant un calcul complexe, nous prouvons dans le chapitre 6 qu'il peut être considérablement simplifié par l'usage de structures intermédiaires autorisant son calcul par niveaux, ainsi que par l'accès aux donnés de manière incrémentale. Une méthode tirant parti de ces deux approches, incrémentale et par niveaux, est proposée. D'autre part, une étude plus poussée du traitement de la série, par lot, par flux incrémental et par flux fenêtré, est réalisée dans le chapitre.

Le chapitre 7 détaille les nombreuses expériences réalisées afin de valider la méthode, du point de vue de sa pertinence puis de sa performance. Dans les deux cas, les résultats sont concluants : la pertinence de la série est démontrée sur des jeux de données artificielles très nombreux ainsi que sur un jeu de données réelles et sa performance est illustrée via le calcul du score d'érosion de séries de taille croissante jusqu'à un million de points, traités en 1,5 seconde.

Le chapitre 8 clôt la seconde partie par la présentation d'une méthode *contextuelle* d'analyse de la période et de la périodicité qui permet la génération de phrases comme « Les deux premiers trimestres de l'année, la série a des valeurs hautes toutes les semaines ». Cette méthode est une extension de DPE qui l'applique localement à différentes sous-parties de la série, déterminées automatiquement via l'utilisation d'un test d'hypothèse original.

Le chapitre 9 présente les conclusions et les perspectives ouvertes par cette thèse.

Chapitre 1

Résumés linguistiques flous : composantes et principes

> Presque tout ce qui caractérise l'humanité se résume par le mot culture.
>
> —François Jacob, *Le Jeu des possibles*

Ce chapitre présente les résumés linguistiques flous (RLF) en décrivant leurs différentes composantes et leur principe. De par leur caractère linguistique, ils s'inscrivent dans le cadre plus large du Traitement Automatique des Langues (TAL ou *NLP* pour *Natural Language Processing*), qui se décompose en Compréhension Automatique des Textes (CAT ou *NLU* pour *Natural Language Understanding*) et Génération Automatique de Texte (GAT ou *NLG* pour *Natural Language Generation*) (Dale et al., 1998). Les RLF sont donc liés aux méthodes de GAT, mais développés dans deux communautés différentes selon des problématiques et des approches différentes.

Après avoir donné un aperçu de leurs principes respectifs dans la section 1.1, la suite du chapitre est dédiée aux RLF, sur lesquels portent les travaux décrits dans cette thèse. La section 1.2 p. 10 détaille les différentes composantes d'un RLF ainsi que leur fonctionnement, décrivant en particulier leur évaluation par le calcul de leur degré de vérité dans le cadre classique, séminal, des données attributs / valeurs. La section 1.3 p. 15 présente ensuite le cas des données décrites sous la forme de séries temporelles et les expressions linguistiques de connaissances qui peuvent en être extraites.

Plusieurs parties de ce chapitre ont été publiées dans les articles (Bouchon-Meunier & Moyse, 2012; Almeida et al., 2013; Moyse et al., 2015; Lesot et al., 2016).

1.1 Approches pour les résumés linguistiques

La production de résumés linguistiques est abordée aujourd'hui avec les RLF d'un côté et les méthodes de GAT de l'autre. Bien que les RLF soient formellement des méthodes de génération automatique de texte, nous les distinguons des méthodes de GAT dans la suite

du document car ces dernières, issues d'autres communautés de recherche, procèdent de manière très différente des RLF. Précisément, les RLF proposent des méthodes évoluées d'analyse des données basées sur des représentations linguistiques sommaires tandis qu'à l'inverse les méthodes de GAT sont peu développées sur l'extraction de données mais plus riches sur la génération linguistique.

Les deux approches, détaillées dans les deux sous-sections suivantes, ont évolué de manière indépendante jusqu'en 2010 où certains articles dans la communauté floue ont souligné leur complémentarité et l'intérêt d'ajouter aux capacités linguistiques des méthodes de GAT celles d'extraction de données des RLF (Kacprzyk & Zadrozny, 2010; Bouchon-Meunier & Moyse, 2012; Ramos-Soto et al., 2016).

1.1.1 Résumés linguistiques flous

Les résumés linguistiques flous (RLF) ont été proposés par Yager (1982) et Zadeh (1983) puis notamment développés par Bosc et al. (1999); Kacprzyk & Zadrozny (2002); Kacprzyk et al. (2008); Castillo-Ortega et al. (2009); Ramos-Soto et al. (2016). Un RLF est un ensemble de phrases décrivant chacune un aspect particulier des données étudiées. Chaque phrase est une instance d'un schéma générique appelé « protoforme » pour lequel un degré de vérité est calculé, indiquant dans quelle mesure la phrase est en adéquation avec les données étudiées.

Les deux protoformes de base des RLF proposés par Yager (1982) sont « Qx sont P » et « QRx sont P », où Q, R et P sont des sous-ensemble flous (sef) appelés respectivement quantificateur (*quantifier*), qualifieur (*qualifier*) et résumeur (*summariser*) et les x sont les données prises dans un un ensemble $X = \{x_1, ..., x_n\}$ à résumer. Ces RLF sont dits « standards » car ils sont les plus couramment utilisés dans le domaine du résumé linguistique flou.

Le protoforme « Qx sont P » peut par exemple être instancié en « La plupart des individus sont grands » et « QRx sont P » en « La plupart des jeunes sont grands », avec $Q = La\ plupart$, $P = grand$ et $R = jeune$ dans le second cas.

Une extension directe de ces protoformes standards est celle proposée par Liétard (2008) avec « $Q(C_1$ et C_2 et ... et $C_3)x$ sont P », par exemple « La plupart des individus sont grands et jeunes et amateurs de sport ». D'autres extensions incluent l'extraction d'information sur des dépendances floues, pour des résumés du type « La plupart des R ont des P similaires » (Bosc et al., 1998; Cubero et al., 1999), des dépendances graduelles (Rasmussen & Yager, 1999; Bosc et al., 1999), des règles graduelles floues avec des résumés de la forme « plus R est élevé, plus P est élevé » (Di-Jorio et al., 2009), les règles graduelles enrichies comme « plus R est élevé, plus P est élevé, particulièrement si S » (Oudni et al., 2013) et les règles graduelles par rapport à la moyenne « plus R est proche de la normale, plus P est élevé » (Hüllermeier, 2002). De plus, des protoformes du type « Qx sont P, et si possible R » peuvent également être évalués avec les résumés linguistiques bipolaires (Dubois & Prade, 2002; Dziedzic et al., 2013).

1.1.2 Approches GAT

L'autre famille d'approches destinées à la création de résumés linguistiques est basée sur les méthodes de GAT. Ces dernières s'inscrivent dans le cadre défini par Reiter & Dale (1997, 2000) qui préconisent les six étapes décrites ci-dessous pour la production automatique de texte.

1. L'extraction des données (*content determination*) permet d'identifier les données utiles au résumé. Elle est souvent réalisée par un moyen assez simple, e.g. une requête en base de données. Comme mentionné précédemment, les approches de GAT se concentrent plus spécifiquement sur la génération de phrases et de textes, et moins sur l'analyse des données en elles-mêmes au contraire des méthodes floues.

2. L'organisation du discours (*discourse planning*) a pour objet l'ordonnancement et la structuration des phrases. Cette étape vise par exemple à exprimer que la moyenne de tel attribut doit apparaître dans un premier paragraphe, puis que les informations des autres attributs doivent apparaître dans le second (Reiter, 1996).

3. L'agrégation des phrases (*sentence aggregation*) assure la représentation sous une forme condensée de phrases partageant certains critères. Par exemple, les phrases « Le prochain train part à 10h » et « Le prochain train est un train Corail » peuvent être représentées par « Le prochain train, qui part à 10h, est un train Corail ».

4. La lexicalisation (*lexicalisation*) correspond au choix des mots pour exprimer les concepts identifiés. Par exemple, « Le prochain train, qui est un un train Corail, part à 10h » peut également être rendu par « Le prochain train, un Corail, quitte la gare à 10h ».

5. La lexicalisation des entités (*referring expression generation*) reprend le principe de lexicalisation pour des entités nommées, identifiées par des constantes dans le programme. Par exemple TRAIN_CORAIL est lexicalisé en « train corail » en français mais peut avoir d'autres représentations dans d'autres langues ou contextes.

6. La réalisation linguistique (*linguistic realisation*) transforme en phrases l'ensemble des concepts obtenus à l'issue des étapes précédentes.

Différentes approches peuvent être utilisées pour chacune de ces étapes. Danlos & El Ghali (2002) par exemple utilisent un système de logique descriptive pour la phase d'extraction des données, une extension d'une théorie de la représentation du discours (SRDT) pour l'étape d'organisation du discours et la grammaire G-TAG pour les étapes de réalisation linguistique.

Cette décomposition en six étapes a été utilisée dans de nombreuses applications, commerciales ou non, comme EasyText qui produit des analyses de sondages (Danlos et al., 2011) ou SumTime Mousam qui crée des bulletins à partir de prévisions météorologiques (Sripada et al., 2003). Une liste plus complète est donnée en annexe A p. 209 (voir aussi Ramos-Soto et al. (2016)).

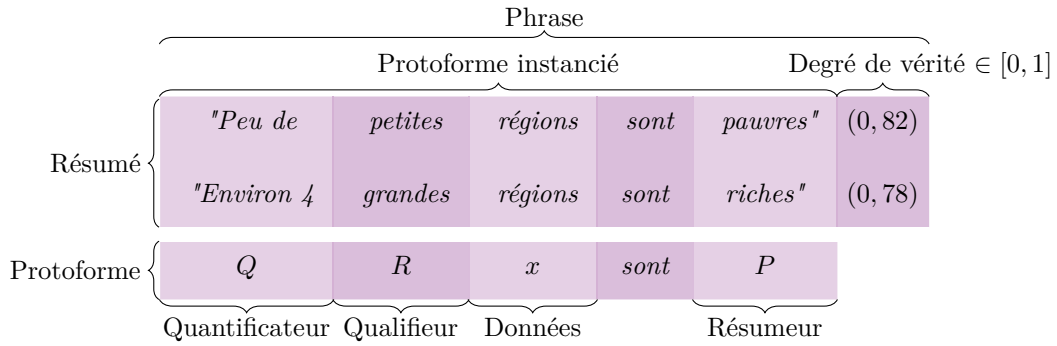


FIGURE 1.1 – Résumé linguistique flou et ses différentes composantes

1.2 Composantes des résumés linguistiques flous

Les éléments constitutifs des RLF brièvement introduits dans la section 1.1 p. 7 sont ici présentés en détail.

Un résumé linguistique flou est un ensemble de phrases instanciées à partir de protoformes et de variables linguistiques associées à des valeurs d'appartenance qui mesurent leur adéquation aux données en entrée.

Nous présentons dans cette section les RLF « standards », i.e. ceux décrits par Yager (1982) pour le traitement de données numériques ou catégorielles sous la forme attributs / valeurs. Les éléments d'un RLF standard sont présentés en détail dans les sous-sections suivantes et illustrés sur la figure 1.1 avec des données fictives décrivant les régions d'un pays.

1.2.1 Données

Les données considérées dans les RLF standards sont des vecteurs x_i de dimension m (attributs) pour $i = 1...n$. Elles peuvent être représentées par un tableau de n lignes et m colonnes.

Chaque attribut est défini sur un ensemble de valeurs également appelé univers, qui peut être continu, discret, fini ou infini.

1.2.2 Variable linguistique

Pour définir le vocabulaire à l'aide duquel les résumés sont formulés, les variables linguistiques (VL) proposées par Zadeh (1975) sont utilisées. Plus précisément, une variable linguistique notée V regroupe un ensemble de m modalités $v_1, ..., v_m$ représentées par les sous-ensembles flous (sef) $\mu_1, ..., \mu_m$ définis sur l'univers de V. Les sef $\mu_1, ..., \mu_m$ peuvent également être notés avec le nom de la modalité qu'ils décrivent (cf. exemple plus bas). Pour rappel, un sef A est défini par une fonction d'appartenance μ_A qui associe à chaque élément de l'univers un degré d'appartenance dans [0,1] (Zadeh, 1965).

Ainsi, une VL est définie pour chaque attribut des données et ses modalités permettent de segmenter l'univers sur lequel l'attribut est défini.

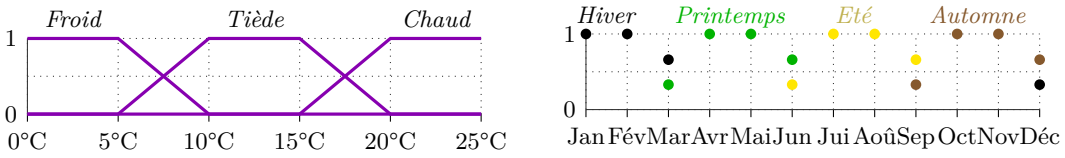


FIGURE 1.2 – La VL *Température* (à gauche) et la VL *Saison* (à droite)

Les variables Q, R et P des protoformes « Qx sont P » et « QRx sont P » sont des modalités de variables linguistiques.

Exemples Un attribut *Température* à valeurs dans $\mathbb{R}^+$ peut par exemple être représenté par la VL illustrée à gauche sur la figure 1.2. En ce cas, $m = 3$ et les modalités $v_1 =$ *Froid*, $v_2 =$ *Tiède* et $v_3 =$ *Chaud* ont pour fonctions d'appartenance :

$$\mu_{Froid}(x) = \begin{cases} 1 & \text{si } x \leq 5 \\ (10 - x)/5 & \text{si } 5 < x < 10 \\ 0 & \text{sinon} \end{cases}$$

$$\mu_{Tiède}(x) = \begin{cases} (x - 5)/5 & \text{si } 5 < x < 10 \\ 1 & \text{si } 10 < x < 15 \\ (20 - x)/5 & \text{si } 15 < x < 20 \\ 0 & \text{sinon} \end{cases}$$

$$\mu_{Chaud}(x) = \begin{cases} 1 & \text{si } x \geq 20 \\ (x - 15)/5 & \text{si } 15 < x < 20 \\ 0 & \text{sinon} \end{cases}$$

Ainsi, une température de 9°C appartient avec un degré de 0,2 à la modalité *Froid,* un degré de 0,8 à la modalité *Tiède* et un degré nul à la modalité *Chaud.*

Un attribut *Saison* à valeurs discrètes dans {*Jan*, *Fév*, *Mar*, *Avr*, *Mai*, *Jun*, *Jui*, *Aoû*, *Sep*, *Oct*, *Nov*, *Déc*} peut être représenté par la VL à droite sur la figure 1.2. Ici, le mois de janvier a un degré d'appartenance de 1 à *Hiver* et 0 aux autres saisons tandis que le mois de mars appartient à un degré $2/3$ à *Hiver*, un degré $1/3$ à *Printemps* et un degré nul aux autres saisons.

Les deux VL *Température* et *Saison* définissent des partitions floues de l'univers de discours. Elles ont de plus la particularité d'être « de Ruspini » car la somme des fonctions d'appartenance de chaque point de l'univers à l'ensemble des modalités est toujours égale à 1, i.e. $\forall x \in X$, $\sum_{j=1}^{m} \mu_j(x) = 1$ (Ruspini, 1969). Elles sont par ailleurs normales car $\forall j = 1...m$, $\exists x \in X$ tq $\mu_j(x) = 1$. Nous verrons au chapitre 2 que ces partitions jouent un rôle particulier dans l'étude de l'interprétabilité des RLF.

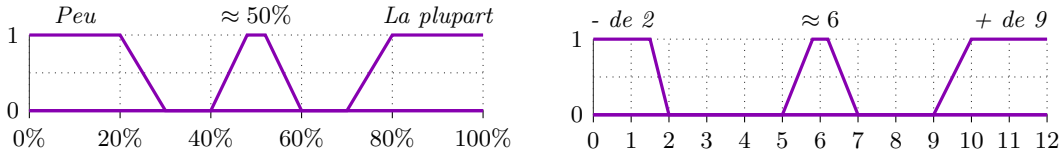


FIGURE 1.3 – Quantificateurs relatifs (à gauche) et absolus (à droite)

1.2.3 Quantificateur flou

Un quantificateur flou est un sef correspondant à une expression linguistique décrivant un décompte d'éléments (Zadeh, 1983).

Un quantificateur *relatif* rapporte le décompte au nombre total d'éléments considérés, comme par exemple *La plupart* ou *Peu*. Un quantificateur *absolu* ne tient pas compte du nombre total d'éléments, comme *Environ 10* ou *Plusieurs*.

Dans le cadre des RLF, les quantificateurs relatifs sont des sef définis sur l'univers [0,1] tandis que les quantificateurs absolus sont définis sur $\mathbb{R}^+$. La figure 1.3 en donne des exemples et la section 2.1.2 p. 28 en discute les propriétés.

1.2.4 Protoforme

De manière générale, un protoforme est un modèle de phrase composé de parties constantes en langage naturel et de parties variables pouvant faire référence à des expressions linguistiques, des expressions numériques ou des sef. L'instanciation d'un protoforme en phrase revient selon le type de partie variable à calculer sa valeur et/ou à évaluer une mesure de qualité pour la phrase instanciée étant donné ses parties variables.

Dans le cadre des RLF standards, les protoformes « Qx sont P » et « QRx sont P » contiennent respectivement deux et trois parties variables sous forme de sef : un quantificateur Q, une modalité P et une modalité R dans le second cas. L'instanciation d'un protoforme est associée à une mesure de qualité liée à son adéquation aux données appelée degré de vérité comme détaillé au paragraphe suivant.

Avec l'exemple de données décrivant des individus, les protoformes « Qx sont P » et « QRx sont P » peuvent être instanciés par exemple en « La plupart des individus sont grands » et « La plupart des jeunes individus sont grands » respectivement, avec $Q =$ *La plupart*, $R =$ *jeunes* et $P =$ *grands*. Dans le premier cas, tous les individus présents dans les données sont pris en compte, tandis que dans le second seuls ceux qui sont jeunes le sont. Ainsi, l'utilisation de la modalité R permet d'évaluer la pertinence de P sur un univers de discours réduit. En ce sens, le protoforme « QRx sont P » peut être vu comme une généralisation de « Qx sont P », ce dernier étant limité au cas $R = X$. Pour cette raison et par souci de clarté du discours, seul le protoforme « QRx sont P » est considéré par la suite.

1.2.5 Valeur de vérité

La mesure du degré d'adéquation des donnés permet le calcul du degré pour le protoforme « QRx sont P » a été proposé par Yager (1982). Il est réalisé en trois étapes décrites ci-après.

1. Appartenance des données aux modalités Les degrés d'appartenance $R(x)$ et $P(x)$ de chacun des x aux modalités R et P respectivement sont calculés dans un premier temps. Si $R = X$ alors $\forall x$, $R(x) = 1$ puisque les x sont les éléments de X.

2. Fonction de comptage La fonction de comptage $\nu(R, P)$ permet de déterminer le nombre d'éléments dans l'univers de discours R qui vérifient P, ou de manière équivalente le nombre d'éléments dans l'univers de discours P qui vérifient R, donc plus généralement les éléments de X qui sont R et P. Cette fonction est donc symétrique par symétrie de la conjonction logique. Elle utilise deux concepts clés de la logique floue, la cardinalité $|\cdot|$ et la t-norme $\top$ d'un sef, et se définit par :

$$\nu(R, P) = |\top (R, P)| \tag{1.1}$$

T-norme La t-norme représente une conjonction en logique floue. C'est une fonction de $[0, 1]^2$ dans $[0, 1]$ qui renvoie le sef intersection de deux sef. Plusieurs t-normes ont été proposées : $\top_Z(x, y) = \min(x, y)$ pour celle de Zadeh, $\top_P(x, y) = xy$ pour la t-norme probabiliste ou $\top_L(x, y) = \max(x + y - 1, 0)$ pour celle de Łukasiewicz (Bouchon-Meunier, 2007). Les propriétés des t-normes sont étudiées exhaustivement dans (Klement et al., 2000). Enfin, celle de Zadeh est utilisée dans le cadre des RLF standards.

Enfin, une t-conorme $\perp$ est définie comme l'opération duale de la t-norme. La t-conorme de Zadeh est définie par $\perp_Z(x, y) = \max(x, y)$ et la t-conorme probabiliste par $\perp_P(x, y) = x + y - xy$.

Cardinalité La cardinalité désigne le nombre d'éléments d'un ensemble. Dans un ensemble classique ou crisp, un élément appartient ou non à l'ensemble. Sa cardinalité est donc à valeurs dans $\mathbb{N}$. Dans un sef au contraire, les éléments appartiennent à un certain degré à l'ensemble : sa cardinalité est donc à valeurs dans $\mathbb{R}$.

La cardinalité floue la plus classique, qui est également celle utilisée dans les RLF standards, est appelée σ-count et définie par Deluca & Termini (1972) pour un sef A comme $\sigma\text{-count}\,(A) = \sum_x \mu_A(x)$. D'autres cardinalités floues sont présentées dans l'annexe C p. 215.

3. Calcul de la valeur de vérité La valeur de vérité du protoforme est calculée comme l'adéquation au quantificateur choisi du compte calculé à l'étape précédente, i.e. la valeur d'appartenance de $\nu(R, P)$ à Q. Elle dépend du type de quantificateur utilisé, les quantificateurs absolus étant définis sur $\mathbb{R}^+$ et les quantificateurs relatifs sur [0,1]. Dans ce dernier

cas, la valeur de $\nu(R, P)$ est divisée par la taille de l'univers de discours $|R|$. La valeur de vérité t d'un protoforme est donc calculée comme :

$$t(QRx \text{ sont } P) = Q\left(\rho\left(R, Q\right) \cdot \nu\left(R, P\right)\right) \tag{1.2}$$

où ρ est le facteur permettant d'adapter le calcul au quantificateur utilisé. Il est défini comme :

$$\rho\left(R, Q\right) = \begin{cases} 1 & \text{si } Q \text{ est absolu} \\ 1/\left|R\right| & \text{sinon} \end{cases} \tag{1.3}$$

1.2.6 Exemple

À titre d'illustration, les deux variables linguistiques *Température* et *Saison* illustrées sur la figure 1.2 p. 11, le protoforme « QRx sont P » et les données de températures mensuelles moyennes données dans le tableau 1.1. Le calcul de la valeur de vérité pour les modalités $P = \mathit{Hiver}$ et $R = \mathit{Froid}$ commence par l'évaluation de l'appartenance de chacune des données aux deux modalités dont les résultats sont donnés dans le tableau, i.e. $\sigma\text{-count}(P) = 3,33$ et $\sigma\text{-count}(R) = 3,46$.

La fonction de comptage donne alors $\nu(\mathit{Froid}, \mathit{Hiver}) = \sum_x \top_Z(\mathit{Froid}(x), \mathit{Hiver}(x)) = 2,83$ ce qui peut s'interpréter par « 2,83 mois froids sont en hiver ».

Avec le quantificateur relatif $Q = \mathit{La\ plupart}$ représenté sur la figure 1.3 p. 12, la valeur de vérité du résumé est :

$$\begin{aligned} t(QRx \text{ sont } P) &= \mathit{LaPlupart}(\nu(\mathit{Froid}, \mathit{Hiver})/|\mathit{Froid}|) \\ &= \mathit{LaPlupart}(2,83/3,46) \\ &= \mathit{LaPlupart}(0,82) = 1 \end{aligned}$$

avec $\rho(\mathit{Froid}, \mathit{LaPlupart}) = 1/|\mathit{Froid}| = 1/3,46$ car *LaPlupart* est un quantificateur relatif.

Même en remplaçant x par « températures », la transformation du protoforme en phrase n'est pas réalisable directement car l'utilisation directe du nom des modalités dans le protoforme donne « La plupart Froid températures sont Hiver ». D'une manière générale, la transformation linguistique du protoforme en phrase est peu détaillée dans les articles sur les RLF. Ici, la phrase adéquate serait « La plupart des températures froides sont en hiver (1,00) ».

Afin de générer les phrases simplement, la solution la plus simple consiste à utiliser un protoforme construit de sorte à pouvoir être directement transformé, ici, ce pourrait être « Q des températures R sont mesurées en P ». Dans le cas cependant où $P = \mathit{Printemps}$ ou $Q = \mathit{Peu}$, la formulation doit également être adaptée. Une réponse à la problématique de conversion des protoformes en phrase pourrait être trouvée dans l'utilisation des techniques de GAT.

Tableau 1.1 – Données d'exemple pour le calcul de la valeur de vérité d'un protoforme

X	Temp. (°C)	Mois	*Froid*(x)	*Hiver*(x)	$\top_Z$(*Froid*(x), *Hiver*(x))
x_1	6,1	Mar	0,78	0,33	0,33
x_2	3,2	Jan	1,00	1,00	1,00
x_3	1,8	Fév	1,00	1,00	1,00
x_4	7,4	Avr	0,18	0,00	0,00
x_5	4,0	Fév	0,50	1,00	0,50
			$\sum = 3,46$	$\sum = 3,33$	$\sum = 2,83$

1.3 RLF de séries temporelles

La section précédente a considéré le cas général des données sous forme attributs / valeurs, celle-ci décrit les RLF utilisés pour l'analyse de séries temporelles. Ces données, étudiées dans la partie 2 de cette thèse, sont un ensemble ordonné de valeurs x_i associées à des étiquettes temporelles t_i. Cette définition est formalisée dans la section 4.1 p. 68.

Les différentes approches présentées dans cette section sont réparties en fonction du type des séries temporelles étudiées, univariées ou multivariées, selon qu'une ou plusieurs séries sont analysées simultanément.

Marín & Sánchez (2016) proposent un état de l'art complet et récent structuré, contrairement à celui-ci, autour des composantes fonctionnelles du système de résumé.

1.3.1 Séries univariées

Différents exemples de génération de résumés linguistiques à partir d'une seule série temporelle sont présentés ici.

Association d'une variable linguistique à l'échelle temporelle Dans le cadre du traitement de données médicales, décrivant la fréquentation d'un hôpital au cours du temps en fonction d'informations météorlogiques, Castillo-Ortega et al. (2009) proposent d'utiliser une VL *Mois* ainsi qu'une VL décrivant la température habituellement constatée aux dates de la série temporelle. Ils génèrent ainsi des RLF comme « Au moins 70% des jours où le temps est doux, le nombre de patients est élevé » ou « La plupart des jours de mai, le nombre de patients est moyen ».

Calcul d'un attribut temporel Kacprzyk et al. (2008) proposent de calculer un attribut *Tendance* égal à la pente de la plus grande droite de régression de points consécutifs et d'erreur inférieure à un seuil. Cette tendance est par la suite associée à une VL composée des modalités *Très croissante*, *Croissante*, *Plate*, *Décroissante* et *Très décroissante*. L'intervalle de temps et le variabilité des tendances sont ensuite associés à des VL. Des RLF quantifiant le nombre et la durée des tendances sont ensuite générés, comme par exemple « Les tendances qui durent le plus longtemps sont les plus variables » ou « Les tendances décroissant lentement qui ont la plus grande durée ont une variabilité importante ».

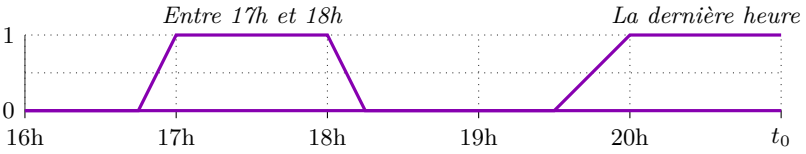


FIGURE 1.4 – Exemples d'intervalles temporels, t_0 représente l'heure actuelle

Fuzzy Temporal Propositions Cariñena et al. (1999, 2000) proposent des protoformes permettant l'étude de l'occurrence ou de la durée d'un phénomène dans une série temporelle, permettant la génération de phrases comme « À un moment durant les 30 dernières minutes, la température était élevée » ou « Durant les 30 dernières minutes, la température était élevée ».

Pour ce faire, en plus de la définition habituelle d'une VL pour représenter les valeurs de la série, les intervalles temporels d'intérêt sont également représentés par des sef définis par l'utilisateur comme ceux illustrés sur la figure 1.4.

En notant μ_A la fonction d'appartenance des valeurs de la série temporelle et μ_T celle de l'intervalle temporel considéré, le degré de vérité du protoforme « A un moment durant T, A » est calculé par :

$$\max_{i=1,\dots,n} \left(\min\left(\mu_A(x_i), \mu_T(t_i)\right)\right)$$

et celui du protoforme « Durant T, A » par :

$$\max_{i=1,\dots,n} \left(\min\left(\mu_A(x_i), 1 - \mu_T(t_i)\right)\right)$$

Cette seconde expression renvoie la plus grande valeur de l'implication $T \Rightarrow A$.

Périodicité Nous définissons au chapitre 5 le protoforme « M toutes les p unités, les valeurs sont élevées » dédié à la périodicité des données. Un exemple de phrase issue de ce protoforme est « Environ toutes les heures, les valeurs sont élevées », associé à un degré de périodicité.

Règles floues Štěpnička et al. (2010, 2011) proposent de décomposer la série temporelle en un ensemble de composantes par une transformée floue (F-Transform) puis d'apprendre des règles sur ces dernières. Ces règles ont une forme particulière du fait de l'utilisation d'expressions linguistiques construites avec un modificateur comme *très*, *précisément*, *approximativement* ou *environ* suivies d'une modalité de variable linguistique.

Les règles d'inférence sont utilisées pour prédire les valeurs suivantes de la série temporelle et portent une connaissance exprimable sous forme linguistique, comme par exemple « si la valeur du mois précédent est moyenne et celle du mois actuel environ grande alors la valeur du mois à venir est à peu près environ grande ». L'interprétabilité des modificateurs dans ce contexte est discutée au section 2.4.1 p. 38.

GLMP La description linguistique granulaire des perceptions (*GLMP* pour *Granular Linguistic Modular Perception*), proposée par Eciolaza et al. (2011) et étudiée de manière théorique par Triviño & Sugeno (2013) , repose sur la théorie computationnelle des perceptions (*CTP* pour *Computational Theory of Perceptions*) proposée par Zadeh (2002).

La GLMP fonctionne de manière hiérarchique en convertissant dans un premier temps les valeurs numériques d'une série temporelle en VL comme dans les méthodes standards basées sur les RLF et en exploitant de plus ces VL pour créer de nouvelles VL, appelées perceptions de second ordre (*2-CP* pour *Second order Computational Perceptions*). La génération des 2-CP à partir des VL de premier niveau est réalisée à l'aide de règles d'inférence données par un utilisateur expert.

De nombreuses applications sont proposées sur ce modèle. Dans l'article introduisant les GLMP, Eciolaza et al. (2011) décrivent le comportement d'un véhicule à partir de données issues d'un simulateur de conduite. Par la suite, Triviño & Sanchez-Valdes (2015) génèrent des recommandations personnalisées visant à réduire la consommation énergétique des foyers étudiés. Sanchez-Valdes et al. (2016) décrivent un système permettant de générer des rapports d'activité physique d'un utilisateur en fonction des données d'accélération enregistrées par son téléphone portable. A partir d'une VL décrivant les valeurs d'accélération, la 2-CP « état de l'activité physique » est calculée à partir d'une machine à états finis, où le passage d'un état à l'autre est donnée par une règle, e.g. « Si l'état actuel est *arrêté* et que l'accélération est *faible* et que la durée d'inaction est *suffisante* alors l'état suivant est *arrêté* ». De la même manière, d'autres CP de niveaux supérieurs sont construites, notamment par agrégation des états successifs au cours de la semaine, permettant finalement des résumés comme « Cette semaine l'utilisateur a eu une activité vigoureuse, car lundi il/elle a eu une activité extrême. Cependant, durant le week-end, l'activité a décru à un niveau modéré *»*.

1.3.2 Séries multivariées

Les séries multivariées sont vues ici comme un ensemble de séries temporelles décrites sur le même intervalle de temps dont les similarités et les différences sont étudiées. Wilbik (2010, chap. 5) décrit un ensemble d'approches utilisées dans ce cadre.

Échelles hiérarchiques temporelles Différentes approches prennent en compte plusieurs séries temporelles pour en réaliser la comparaison. Par exemple, Castillo-Ortega et al. (2011a) proposent d'étudier deux séries temporelles par la création d'une troisième qui compare la synchronie de leurs évolutions, permettant des RLF du type « La plupart des jours de 2001, les deux séries évoluent localement dans la même direction ».

Les auteurs proposent également trois VL hiérarchisées pour décrire l'axe temporel, l'une annuelle avec les modalités *2001*, *2002*..., la seconde utilisant des périodes de 5 ans et la troisième des décennies. Les résumés extraits sont ceux qui couvrent l'intervalle le plus large, puisque si « QRx sont P » est *vrai* sur 5 ans il l'est également sur chacune des 5 années. L'intérêt de cette approche en termes de brièveté du résumé est discuté dans

la section 2.4.1 p. 37.

Mesure d'exceptionnalité Van der Heide & Triviño (2009) proposent de calculer pour chaque intervalle de temps considéré la moyenne et l'écart-type de la consommation électrique d'un ensemble de foyers. Un attribut est ensuite calculé comme l'écart à la moyenne de la consommation de chaque foyer, permettant ainsi de ne renvoyer que les RLF concernant des foyers ayant une consommation particulière.

Analyse d'une base de données médicale Dans le cadre du séjour de deux mois de Rui Jorge Almeida au LIP6, nous avons collaboré pour étudier la base Medan (Paetz et al., 2003) en proposant des protoformes originaux. Ce travail a donné lieu à la publication (Almeida et al., 2013).

La base Medan contient des données médicales de patients admis dans les services de réanimation de 70 hôpitaux allemands entre 1998 et 2002, pour des chocs septiques abdominaux. Les données étudiées portent sur les dernières 24 heures du patient en réanimation, dont il peut sortir mort ou guéri.

Les quantificateurs *Très peu*, *Peu*, *La moitié*, *La plupart* et *Presque tous* sont définis ainsi que la VL *Rythme cardiaque* avec les modalités *Très bas*, *Bas*, *Moyen*, *Élevé* et *Très élevé* et la VL non floue *État* égale à *Mort* ou *Vivant*.

Nous avons introduit deux types de protoformes : le premier, temporel, s'écrit « Qx sont P, Q_T fois », où Q est un quantificateur classique et Q_T est un quantificateur temporel comme *La plupart du temps* ou *Rarement*. Le second, dit « différentiel », s'écrit « QR_1x sont P contrairement aux R_2x ». Ces protoformes permettent l'instanciation de phrases comme « Peu d'individus ont une tension faible la plupart du temps » pour le premier ou « Peu d'observations réalisées sur des hommes montrent un rythme cardiaque faible contrairement à celles réalisées sur des femmes » pour le second.

Le degré de vérité du premier type est calculé comme deux RLF imbriqués, i.e. le degré de vérité par série est calculé suivi par celui sur l'ensemble des séries. Le degré d'adéquation du second type de protoforme est évalué comme la différence entre les degrés de vérité de ses deux parties, i.e. « QR_1x sont P » et « QR_2x sont P ».

L'ensemble des résumés générés permet notamment de déterminer que très peu de patients sortis vivants de l'unité ont eu des rythmes cardiaques élevés, contrairement à ceux qui y sont décédés.

1.4 Bilan

Ce chapitre a présenté brièvement les différentes méthodes de génération de texte en langue naturelle sur des bases de données numériques et/ou temporelles. Il a détaillé plus particulièrement les approches à bases de résumés linguistiques flous et donné un certain nombre de celles permettant l'analyse de séries temporelles.

Cet état de l'art permet de mettre en lumière les problématiques qui sous-tendent

les deux parties de cette thèse. La première est liée à l'interprétabilité des RLF. En effet, il apparaît à l'issue de ce chapitre que les protoformes peuvent traiter d'un nombre potentiellement illimité de sujets par l'usage d'attributs auxiliaires calculés, de règles d'inférences données par des utilisateurs et de manière plus générale par toutes les techniques de traitement de données pouvant être converties en phrases. L'énumération de tous les protoformes possibles étant irréalisable, l'étude de leurs propriétés générales et de leur qualité paraît alors plus pertinente. Est-il possible d'identifier des règles générales permettant de garantir que les protoformes sont transformés en phrases intelligibles, interprétables ? Peut-on s'assurer de la cohérence des différentes phrases d'un résumé ainsi généré ? Une étude de ces questions et des approches pour y répondre sont au centre de la première partie de cette thèse.

D'autre part, les différentes méthodes d'analyse des séries temporelles présentées dans la section 1.3 p. 15 de ce chapitre n'abordent pas la question pourtant cruciale de leur périodicité, i.e. de la répétition de motifs en leur sein. De plus, si différentes approches classiques comme la transformée de Fourier permettent dans un certain nombre de cas d'y apporter une réponse, leur rendu linguistique reste peu documenté. C'est donc de la problématique de la périodicité des séries temporelles et de leur caractérisation en langue naturelle qu'il est question dans la seconde partie du document.

Première partie

Cohérence des résumés linguistiques flous

Introduction

La question de la qualité des RLF apparaît centrale : les protoformes peuvent être créés indéfiniment et leur qualité doit être évaluée afin d'écarter ceux qui ne vérifient pas certains critères minimaux d'acceptabilité.

Différentes mesures de qualité pour les RLF sont présentées dans le premier chapitre de cette partie, organisées selon les différents parties auxquelles elles sont attachées.

Les premières concernent le vocabulaire, qui rassemble les variables linguistiques et les quantificateurs, et qui constitue les mots élémentaires à l'aide desquels les phrases sont générées. Y sont attachées les mesures de couverture, de nombre ou de spécificité par exemple. Les propriétés des phrases sont ensuite étudiées, avec des mesures comme l'imprécision, la pertinence ou la longueur. Le degré de vérité, mesure centrale de la qualité d'une phrase, fait l'objet d'une section spécifique dans ce premier chapitre. Enfin, les propriétés des résumés dans leur ensemble et des systèmes en permettant la génération sont présentées.

Les propriétés de cohérence occupent une place particulière parmi les différentes mesures de qualité présentées car elles s'intéressent à la question de l'interprétabilité du résumé dans son ensemble. En effet, un résumé qui contiendrait des phrases comme « La plupart des jeunes sont grands » et « Peu de jeunes sont grands » ou encore affichant des valeurs de vérité différentes pour les phrases « La plupart des jeunes ne sont pas grands » et « Peu de jeunes sont grands » serait peu compréhensible car ne respectant pas les règles élémentaires de la logique.

De manière plus précise, nous montrons dans le chapitre 3 que la question de la cohérence entre phrases est débattue de longue date et nous proposons de présenter l'ensemble des propositions faites à ce sujet en organisant les différents types d'opposition possibles entre phrases selon une échelle de complexité croissante. A l'aide de cette mise en perspective, nous présentons une nouvelle structure recensant l'ensemble des oppositions possibles entre protoformes dans le cadre des résumés linguistiques flous. D'autre part, nous montrons que certaines propriétés de cohérence peuvent être vérifiées pour tous les protoformes classiques.

Chapitre 2

Qualité des RLF

> Ce qui me tue, dans l'écriture, c'est qu'elle est trop courte. Quand la phrase s'achève, que de choses sont restées au-dehors !
>
> —Jean-Marie Gustave Le Clézio, *Le Livre des fuites*

Un système de RLF prend en entrée des données ainsi qu'un vocabulaire composé de quantificateurs et de variables linguistiques et renvoie un résumé contenant plusieurs phrases construites sur des protoformes.

L'annexe B p. 211 illustre un tel système appliqué à des données réelles pour établir un résumé linguistique des corrélations entre le nombre de pages des livres vendus sur Amazon et leur position dans les classements de meilleures ventes.

Ce chapitre propose un état de l'art des mesures qui ont été proposées pour évaluer la qualité d'un résumé, tâche complexe qui fait intervenir de nombreuses composantes. Nous proposons de structurer ces mesures selon trois niveaux, selon qu'elles s'appliquent au vocabulaire, aux phrases, ou au résumé globalement. Il convient de noter que ces mesures, au-delà du sens commun, exploitent également des travaux issus du domaine de la philosophie comme ceux de Grice (Grice, 1970) sur les maximes conversationnelles ou de la logique avec ceux du groupe Gamut (1991)

La section 2.1 est consacrée aux mesures appliquées au vocabulaire, qui évaluent sa pertinence en fonction de ses propriétés intrinsèques, comme le nombre de modalités d'une variable linguistique ou la forme des sef utilisés par exemple. Ces mesures sont indépendantes des données et peuvent être calculées avant la génération du résumé.

Les mesures associées aux phrases, détaillées dans la section 2.2 p. 29, sont calculées a posteriori et dépendent donc des données. Parmi celles-ci, le degré de vérité, qui évalue l'adéquation aux données comme détaillé dans la section 1.2.5 p. 13, joue un rôle central. Les propriétés souhaitées pour cette mesure sont discutées dans la section 2.3 p. 32, ainsi que certaines extensions du cadre classique développées pour les vérifier.

Les mesures liées au résumé dans son ensemble, présentées dans la section 2.4 p. 37, sont la plupart du temps des agrégations des mesures de niveaux vocabulaire et phrase.

La dernière section de ce chapitre se concentre sur les méthodes de génération du

système de RLF permettant d'optimiser certaines de ces mesures.

2.1 Vocabulaire

Cette section décrit différentes mesures qui ont été proposées pour mesurer la qualité du vocabulaire, i.e. des VL et des quantificateurs. Certaines d'entre elles sont calculées a priori, indépendamment des données, et d'autres a posteriori, avec les données.

La plupart de ces mesures provient du domaine de la modélisation floue dans lequel les règles d'un système flou sont apprises à partir de données étiquetées fournies en entrée. L'interprétabilité des sef flous construits dans ce cadre est un sujet largement étudié dont la complexité provient du nécessaire compromis entre leur interprétabilité et leur précision (Casillas et al., 2003).

Ces mesures sont également applicables aux résumés linguistiques flous (Kacprzyk & Zadrozny, 2013a; Lesot et al., 2016) et présentées ci-dessous.

2.1.1 Modalités

Différentes propriétés et mesures de qualité pour les sef sont proposées en modélisation floue et ici transposées au cadre des RLF pour les modalités des VL.

Propriétés

Sef On note V une VL définie sur l'univers X et V_j, $j = 1...m$, les m fonctions d'appartenance de ses modalités.

La propriété de *normalité* est vérifiée si $\exists x \in X$ tq $V_j(x) = 1$. Cette contrainte garantit qu'il existe au moins un point de l'univers qui appartient complètement au sef. Elle est souvent avancée comme un critère nécessaire d'interprétabilité dans le domaine des systèmes de règles floues (Casillas et al., 2003; Mencar & Fanelli, 2008; Zhou & Gan, 2008; Gacto et al., 2011).

Mencar & Fanelli (2008) proposent d'autres contraintes sur la forme du sef comme sa concavité, de laquelle découle directement son unimodalité pour un sef normal, ainsi que la continuité de sa fonction d'appartenance.

Partitions Les propriétés concernant les partitions d'une VL sont présentées ci-après.

Distingabilité La distingabilité (Zhou & Gan, 2008; Alonso et al., 2009; Gacto et al., 2011) aussi appelée non redondance (Casillas et al., 2003) vise à assurer que les sef de la partition sont suffisamment séparés et clairement différenciables. Différentes mesures de similarité peuvent être utilisées pour identifier les sef non distingables car trop similaires. Pulkkinen & Koivisto (2010) proposent par exemple de fixer des distance minimales entre les centroïdes de deux sef successifs et des maxima sur l'appartenance simultanée d'un point à plusieurs sef.

Nombre La plupart des travaux suggèrent le *magic number* de Miller 7 ± 2 pour le nombre idéal de sef de la partition (Zhou & Gan, 2008; Alonso et al., 2009; Gacto et al., 2011).

Couverture La propriété de couverture est vérifiée si l'ensemble des sef de la partition couvre l'ensemble de l'univers de discours. L'approche la plus simple pour la satisfaire est de s'assurer que chaque élément de l'univers appartient à un degré non nul (Zhou & Gan, 2008) ou, dans une variante plus exigeante, supérieur à α (Mencar & Fanelli, 2008; Gacto et al., 2011), à au moins un sef de la partition.

Pour Dubois & Prade (1985b), la propriété de couverture est vérifiée si ses modalités vérifient $|V_j| + |\overline{V_j}| = |X|$.

D'autres contraintes plus spécifiques ont également été proposées. Par exemple, la propriété de granulation uniforme (Mencar & Fanelli, 2008) impose que les sef de la partition soient de cardinalités à peu près égales. Celle sur la première et la dernière modalité de la partition requiert que la plus petite et la plus grande valeurs de l'univers appartiennent respectivement à la première et à la dernière modalités de la partition.

La couverture du vocabulaire ne doit pas être confondue avec celle des phrases, calculée avec les données (cf. section 2.2.2 p. 30).

Complémentarité La complémentarité impose que la somme des appartenances d'un point à l'ensemble des modalités de la partition égale 1 (Mencar & Fanelli, 2008; Zhou & Gan, 2008; Alonso et al., 2009; Gacto et al., 2011). Les partitions vérifiant cette propriété sont donc des partitions de Ruspini (cf. section 1.2.2 p. 10). Ces dernières sont très couramment utilisées dans le cadre des RLF car vérifiant de nombreuses propriétés présentées dans ce chapitre (Bodenhofer & Bauer, 2005; Alonso et al., 2009; Díaz-Hermida & Bugarín, 2010).

Ces partitions ont néanmoins certains inconvénients. Mencar & Fanelli (2008) par exemple rappellent que les degrés d'appartenance à différents sef ne sont pas nécessairement additifs par nature et que la propriété de complémentarité n'est pertinente qu'avec la cardinalité σ-count (voir la section 2.3.2 p. 34 et l'annexe C p. 215 pour une discussion plus complète sur les cardinalité floues). Sanchez-Valdes & Triviño (2013) suggèrent de s'affranchir de la complémentarité afin de permettre une mesure de *non interprétabilité*, définie comme 1 moins la somme des appartenances d'un point au sef de la partition.

Nous pensons néanmoins sur la base de nos expériences sur les systèmes de RLF (cf. annexe B p. 211) et de nos attentes concernant leur interprétabilité que les partitions de Ruspini sont adéquates dans ce cadre.

Labels Les noms associés aux modalités doivent être cohérents avec la compréhension que l'on en a intuitivement. Par exemple, les 3 sefs *Petit*, *Moyen* et *Grand* d'une VL *Taille* doivent être construits de sorte à respecter l'ordre habituellement induit par ces trois adjectifs, i.e. $Petit \prec Moyen \prec Grand$.

Díaz-Hermida & Bugarín (2010) suggèrent également de systématiquement définir l'antonyme associé à un terme, par exemple *Petit* si *Grand* est défini et *Très petit* si *Très grand*

l'est également. Une discussion complète des différents types d'oppositions, dont l'antonyme, est proposée au chapitre 3.

Mesures de spécificité, d'imprécision et de flou

Les mesures pour les modalités de variables linguistiques évaluent leur faculté à référencer précisément un élément de l'univers. Yager (1982) définit le degré de spécificité comme la somme de l'inverse du nombre d'éléments des α-coupes du sef. Le degré d'imprécision détaillé par Kacprzyk & Zadrozny (2005b) est construit comme la taille du support du sef rapporté à celle de l'univers sur lequel il est défini. Wilbik (2010, p. 78) considère le degré de flou (*fuzziness*), calculé comme la distance entre le sef et l'ensemble crisp le plus proche, égal à 0 lorsque le sef est inférieur à 0,5 et 1 sinon.

Il est intéressant de noter que ces mesures, issues des approches RLF, sont étroitement liées à la propriété de distingabilité présentée plus haut. D'une manière générale l'objectif est d'optimiser le compromis couverture / spécificité (ou distingabilité) / nombre de modalités (Alonso et al., 2009; Gacto et al., 2011). En effet, plus la couverture d'un sef est importante, moins il est spécifique et donc moins il est distingable des autres modalités. A l'inverse, l'utilisation de sef plus spécifiques entraîne une diminution de leur couverture et donc la nécessité d'en introduire de nouveaux.

2.1.2 Quantificateurs

Le concept d'interprétabilité des quantificateurs fait appel à des mesures spécifiques : certaines, comme la couverture, étendent des notions présentées ci-dessus pour les sef en général, et d'autres s'appliquent spécifiquement aux quantificateurs, indépendants ou définis dans des familles.

Couverture des quantificateurs Díaz-Hermida & Bugarín (2010) proposent de mesurer la couverture d'un quantificateur relatif de manière différente de la mesure des sef d'une partition floue présentée dans la section 2.1.1, i.e. sans se baser sur la taille du support mais sur le nombre d'individus pris en compte par le quantificateur.

Ainsi, un quantificateur comme *La Plupart*, dont le support est défini par exemple sur [0,7 ;1] a une couverture plus importante que *Peu* dont le support est défini sur [0 ;0,3] et qui ou, bien que les deux supports aient la même taille. Cette définition de la couverture pour un quantificateur permet donc de favoriser les phrases décrivant une quantité importante de données.

Quantificateurs indépendants Un quantificateur indépendant est défini de manière autonome, contrairement aux familles de quantificateurs présentées ci-dessous. Un exemple de quantificateurs indépendants est donné dans la section 1.2.3 p. 12 et illustré sur la figure 1.3 p. 12.. Certaines études dans le domaine de la cognition apportent un éclairage utilisateur intéressant pour leur définition. Laurent et al. (2004) montrent par exemple que les quantificateurs *Presque tous*, *La plupart*, *Peu*, *Environ la moitié / un quart / un tiers*,

sont spontanément utilisés par des utilisateurs à qui il est demandé de décrire des tableaux de données. On peut donc penser que leur utilisation est pertinente avec des RLF.

Newstead et al. (1987) ont étudié l'impact de la taille du jeu de données sur l'interprétation des quantificateurs. Si *Tous*, *La plupart*, *Beaucoup*, *La moitié* et *Aucun* sont interprétés de manière constante et ne dépendent pas de la taille des données, *Plusieurs*, *Quelques* et *Peu* sont analysés comme représentant une proportion d'autant plus petite que le jeu de données est grand. *Peu* par exemple représente 26% pour un ensemble de 12 éléments mais seulement 9% pour un autre de 10 000.

Famille de quantificateurs Les quantificateurs peuvent également être définis comme des instanciations de familles paramétriques (Castillo-Ortega et al., 2011a; Díaz-Hermida & Bugarín, 2010). Dans le premier article par exemple, les auteurs proposent d'utiliser un ensemble ordonné de q quantificateurs Q_i non décroissants tels que $Q_j \preceq Q_k \leftrightarrow \mu_{Q_j} \leq \mu_{Q_k}$: le plus grand d'entre eux est Q_1 et représente $\exists$ et les suivants correspondent à *Au moins 10%*, *Au moins 20%*, etc. jusqu'au dernier, Q_q, représentant $\forall$. La connaissance induite par cet ordre est mise à profit pour n'extraire que les résumés associés au quantificateur le plus précis.

2.1.3 Adéquation du vocabulaire

Si les méthodes de modélisation floue permettent la définition du vocabulaire à partir de données étiquetées en optimisant les différentes propriétés décrites dans la section 2.1.1 p. 26 (Mencar & Fanelli, 2008; Gacto et al., 2011), la définition automatique du vocabulaire dans le cadre des RLF ne peut se baser sur ces approches car les données en ce cas ne sont pas étiquetées. De plus, les méthodes à base de règles construisent le vocabulaire à l'aide des données tandis que les RLF utilisent un vocabulaire prédéfini par un expert.

Afin de faciliter ce travail de définition, Lesot et al. (2013) proposent d'utiliser une approche de clustering pour adapter le vocabulaire utilisateur afin d'en améliorer les caractéristiques en termes de spécificité et de distingabilité à l'aide d'indices de qualité sur les clusters obtenus.

2.2 Phrases et protoformes

Les mesures présentées dans cette section sont directement conçues pour les RLF. Celles concernant les protoformes sont calculées a priori et présentées dans le premier paragraphe. Celles concernant les phrases, plus nombreuses et calculées a posteriori, sont décrites dans le second. Le degré de vérité, mesure essentielle pour une phrase, est présenté dans la section 2.3 p. 32 qui lui est dédiée.

2.2.1 Protoforme

Imprécision

La seule mesure proposée pour le protoforme est celle d'imprécision, basée sur le même principe que celles utilisées sur les sef, décrite dans la section 2.1.1 p. 28. Pour le protoforme « Qx sont P », Castillo-Ortega et al. (2012) définissent l'imprécision comme la moyenne des aires sous la courbe des fonctions d'appartenance de Q et P. Kacprzyk & Zadrozny (2005b); Wilbik (2010) proposent quant à eux de le calculer pour le protoforme « QRx sont P » comme la moyenne pondérée des degrés d'imprécision de ses composantes Q, R et P.

2.2.2 Phrase

Focus, couverture

Le degré de focus s'applique aux phrases générées à partir du protoforme « QRx sont P » et donne la représentativité de R dans la base, calculée comme $|R|/n$ (Kacprzyk & Zadrozny, 2005b; Wilbik, 2010, p. 82). Supposons par exemple que le degré de vérité de la phrase « Tous jeunes sont grands » soit élevé mais que son degré de focus soit faible car la base ne contient qu'un seul élément *jeune* : cette phrase est trompeuse car elle semble faire état d'une règle générale dans la base alors qu'elle ne décrit qu'un cas. Le degré de focus permet de l'écarter, au même titre que le support utilisé dans les règles d'association.

Il est à noter que ce degré ne peut être utilisé qu'avec un quantificateur relatif puisque le quantificateur absolu porte le nombre considéré dans son expression. Ainsi, sur une base de 1000 individus, la phrase « Environ 3 jeunes sont grands » n'est pas trompeuse puisque le nombre de 3, bien que faible, est annoncé. Ces phrases sont en revanche pauvres en termes d'information, comme remarqué dans l'analyse de l'expérience FFS détaillée dans l'annexe B p. 211.

Le degré de couverture est similaire au degré de focus à ceci près qu'il est calculé sur la conjonction R et P au lieu de R uniquement (Kacprzyk & Zadrozny, 2005b; Wilbik, 2010, p.81). Son interprétation et son usage sont les mêmes, i.e. il permet d'ignorer les phrases portant sur un nombre trop faible de données, jugé non significatif.

Pertinence, exceptionnalité

Le degré de pertinence (*appropriateness*) est élevé si deux attributs sont dépendants et faible sinon (Kacprzyk & Yager, 2001; Kacprzyk & Zadrozny, 2005b). L'hypothèse est faite qu'une phrase dont les deux attributs R et P sont dépendants est plus intéressante qu'une autre dont les attributs ne le sont pas. Supposons par exemple que 50% des individus soient *jeunes* et que 50% soit *très qualifiés*. Si 25% des *jeunes* sont *très qualifiés*, alors la phrase « Environ un quart des jeunes sont très qualifiés » est peu pertinente car l'âge et la qualification sont indépendants. A l'inverse, si « Environ 80% des jeunes sont très qualifiés », alors la phrase est pertinente car les attributs sont corrélés.

Il convient toutefois de noter que le degré de pertinence n'est justifié que lorsqu'en effet deux attributs sont a priori considérés comme indépendants. Avec par exemple les deux attributs *Niveau d'éducation* et *Salaire* pour lesquels il est raisonnable d'attendre une corrélation, le degré de pertinence est contre-intuitif puisqu'il renvoie un score d'autant plus faible que les attributs sont indépendants alors qu'en ce cas justement cette propriété serait surprenante et mériterait d'être signalée à l'utilisateur.

Le degré de pertinence peut être rapproché de la mesure d'exceptionnalité proposée par Van der Heide & Triviño (2009) et présentée dans la section 1.3.2 p. 18. Dans le cadre de la comparaison de séries temporelles, cette dernière permet de conserver des phrases relatives à des valeurs très différentes de la moyenne constatée sur les autres séries.

De la même manière, le score de différenciation que nous présentons au section 1.3.2 p. 18 et dans (Almeida et al., 2013) permet de mettre en avant un groupe de données associé à une phrase ayant des propriétés sensiblement différentes de celles d'autres groupes associés à d'autres phrases, entraînant par exemple la génération de la phrase « Peu d'observations réalisées sur des hommes ont une valeur faible de rythme cardiaque contrairement à celles observées sur des femmes ».

Degré d'informativité

Le degré d'informativité proposé par Yager (1982) permet d'exploiter l'information associée aux phrases ayant une faible valeur de vérité, habituellement supprimées des résumés. L'auteur souligne toutefois que de telles phrases peuvent être informatives. Il propose donc de combiner le degrés de vérité de la phrase avec la spécificité de Q et P d'une part et d'autre part de combiner 1 moins la valeur de vérité de la phrase avec la spécificité des négations de Q et P et de retenir la phrase ayant la plus grand score des deux.

Longueur d'une phrase

Kacprzyk & Yager (2001) proposent d'évaluer la taille d'une phrase issue d'un protoforme « Qx sont P » comme $2 \times 0,5^{|P|}$ et Wilbik (2010, p.87) celle d'une phrase issue de « QRx sont P » comme $|R| + |P|$. Une phrase plus courte est valorisée par rapport à une autre plus longue.

Mesures agrégées

Kacprzyk & Zadrozny (2005b) introduisent un degré de vérité total d'un phrase calculé comme la somme pondérée de sa longueur et des degrés de pertinence (cf. section 2.2.2), d'imprécision (cf. section 2.2.1) et de couverture (cf. section 2.2.2). La détermination des poids reste à la charge de l'utilisateur. Ils proposent également de définir la meilleure phrase comme celle dont le degré de vérité total est maximal parmi l'ensemble des phrases possibles.

Díaz-Hermida & Bugarín (2010) proposent un index combinant le degré de vérité de la phrase (cf. section 2.3), la spécificité (cf. section 2.1.1 p. 28) et la couverture du quantificateur (cf. section 2.1.2 p. 28), la phrase la plus spécifique avec la plus grande couverture et le plus grand degré de vérité étant retenue.

2.3 Degré de vérité

Le degré de vérité capture l'adéquation aux données d'une phrase. Elle est sa mesure de qualité la plus courante et la plus étudiée, et ainsi détaillée dans cette section spécifiquement.

Un certain nombre de ses propriétés, présentées dans un premier temps, ont été définies pour en cadrer la mise en œuvre. Il en ressort notamment que les systèmes de RLF standards ne permettent pas d'en vérifier certaines, notamment liées à la cohérence du résumé. Le second paragraphe présente donc un ensemble de méthodes alternatives permettant de les satisfaire dans le cadre de la logique floue habituelle. Enfin, le troisième paragraphe introduit des méthodes basées sur des extensions ou des réinterprétations du paradigme flou, dans le but également de vérifier ces propriétés.

2.3.1 Propriétés du calcul du degré de vérité

Un nombre important d'articles décrivent les propriétés attendues pour le degré de vérité calculé des phrases d'un RLF (Glöckner, 1997; Delgado et al., 2000; Blanco et al., 2002; Barro et al., 2003; Delgado et al., 2014). Ces propriétés sont résumées ci-dessous.

Propriétés générales Tout d'abord, les résumés doivent être insensibles à l'ordre des données, c'est-à-dire que leurs permutations doivent produire les mêmes résumés avec les mêmes évaluations. Cette propriété peut aussi être vérifiée pour les séries temporelles en ajoutant un attribut contenant la date à laquelle sont associées la ou les valeurs.

De plus, l'évaluation du degré de vérité doit être « continue » dans le sens où une légère variation de la définition du vocabulaire et/ou des données doit n'entraîner qu'une légère variation du degré de vérité.

Elle doit également ne pas être trop stricte, i.e. pour un ensemble de quantificateurs il doit exister au moins un résumeur permettant d'obtenir un degré différent de 0 ou 1.

D'un point de vue algorithmique enfin, les méthodes proposées doivent être efficaces, i.e. de complexité comprise entre $O(n)$ et $O(n \log n)$.

Propriétés du vocabulaire Une propriété requise concernant le qualifieur R est qu'il doit agir comme une restriction de l'univers de discours X et donc que si $R = X$, la valeur de vérité de « QRx sont P » doit être égale à celle de « Qx sont P ». Cette propriété est vérifiée pour les RLF standards (cf. section 1.2.4 p. 12).

Par ailleurs, le calcul du degré de vérité doit être monotone par rapport au quantificateur et au résumeur, i.e. $Q \subseteq Q' \Rightarrow t(QRx \text{ sont } P) \leq t(Q'Rx \text{ sont } P))$ et $P \subseteq P' \Rightarrow$

$t(QRx \text{ sont } P) \leq t(QRx \text{ sont } P')$.

Propriétés des modalités Trois propriétés sont requises pour les modalités et concernent le lien entre qualifieur et résumeur.

La première s'écrit $R \subseteq P \Rightarrow t(QRx \text{ sont } P) = 1$, illustrée par la phrase « Tous les très grands sont grands » qui doit être absolument vraie.

La seconde s'écrit $R \cap P = \varnothing \Rightarrow t(QRx \text{ sont } P) = 0$, illustrée par la phrase « Tous les petits sont grands » qui doit être absolument fausse.

Enfin, la dernière impose que la valeur de vérité d'une phrase pour laquelle R et P sont crisp doit être exprimable à l'aide d'opérateurs ensemblistes classiques par $t(QRx \text{ sont } P) = Q(|R \cap P|/|R|)$. Le lien entre le calcul de la valeur de vérité dans le contexte flou des RLF et l'expression utilisant les notations ensemblistes crisp est discuté dans la section 3.2.1 p. 51 concernant les quantificateurs généralisés.

Propriétés des quantificateurs La méthode proposée doit fonctionner sur tous les types de quantificateurs et pas sur certains d'entre eux seulement comme les quantificateurs dont la fonction d'appartenance est monotone par exemple.

De plus, les quantificateurs flous *Il existe* et *Tous* doivent avoir les mêmes propriétés que les quantificateurs classiques $\exists$ et $\forall$, i.e. $t(\exists Rx \text{ sont } P) = \bot(\top(R, P))$ et $t(\forall Rx \text{ sont } P) = \top(R \to P)$ où $\to$ est une implication floue.

Enfin, $\top(t(QRx \text{ sont } P), t(Q'Rx \text{ sont } P)) = t(\top(Q, Q')Rx \text{ sont } P)$ où Q et Q' sont non décroissants. Par exemple, la t-norme des valeurs de vérité de « Au moins 5 jeunes sont grands » et « Au plus 10 jeunes sont grands » doit être égale à la valeur de vérité de « Au moins 5 et au plus 10 jeunes sont grands ».

Propriétés de cohérence Les propriétés de cohérence sont les plus intéressantes en termes d'interprétabilité car faisant appel à des notions moins triviales d'un point de vue sémantique que celles présentées ci-dessus. Elles sont présentées par l'exemple ci-dessous et définies de manière formelle dans le chapitre 3.

La propriété d'*antonymie* est vérifiée si « Peu de jeunes ne sont pas grands » a la même valeur de vérité que « La plupart des jeunes sont grands ».

Celle de *négation externe* implique que la valeur de vérité de « Beaucoup de jeunes sont grands » est égale au complément de « Pas beaucoup de jeunes sont grands ».

Enfin, la *dualité* implique que la valeur de vérité de « Beaucoup de jeunes ne sont pas grands » est au complément de « Pas beaucoup jeunes sont grands ».

La vérification de deux de ces propriétés entraîne celle de la troisième. Ces propriétés ne sont pas vérifiées dans le cadre des RLF standards pour les protoformes du type « QRx sont P » avec la t-norme $\top_Z$ (Yager, 1982), ce qui a entraîné la création d'un ensemble d'extensions, soit de la méthode d'évaluation des valeurs de vérité, soit du paradigme flou, pour permettre des systèmes d'évaluation du degré de vérité qui les vérifient. Nous montrons cependant dans le chapitre 3 que ces dernières peuvent l'être dans le cadre de la logique floue standard.

Les méthodes basées sur une extension du calcul de la valeur de vérité et celles basées sur une extension du paradigme flou sont présentées dans les deux paragraphes suivants.

2.3.2 Extensions du système de RLF standard

Les extensions des méthodes de calcul de la valeur de vérité reposent sur l'utilisation de cardinalités différentes de σ-count et de t-normes spécifiques. Un état de l'art complet de ces différentes approches est donné par Delgado et al. (2014). Nous avons également mené une étude concernant certaines de ces extensions (Bouchon-Meunier & Moyse, 2012), détaillée dans l'annexe C p. 215.

Cardinalité entière

Une première proposition est celle de Ralescu (1995) qui utilise la cardinalité $nCard$ à valeurs dans $\mathbb{N}$ au lieu de σ-count à valeurs dans $\mathbb{R}$. L'intérêt de valeurs entières pour représenter le nombre d'éléments d'un sef réside dans leur interprétabilité a priori supérieure. En effet, la sémantique de « il y a 3,6 personnes grandes » calculée avec une cardinalité à valeurs dans $\mathbb{R}$ n'est pas simple à appréhender.

Comme nous le montrons dans l'annexe C p. 215 cependant, $nCard$ est peu robuste à certains changements légers dans des sef très similaires, ce qui en réduit l'intérêt dans le cadre des RLF.

Cardinalités floues

Le plus grand nombre de contributions réalisées pour l'extension du calcul de la valeur de vérité repose sur l'utilisation de cardinalités floues. Le nombre d'éléments d'un sef n'est alors plus représenté par un scalaire, entier ou réel, mais par un nombre flou, i.e. un sef normalisé convexe. L'annexe C p. 215 étudie ces cardinalités en détail.

Adéquation entre le quantificateur et la cardinalité Le calcul de la valeur de vérité avec une cardinalité floue est réalisé par une mesure de la similarité entre sa fonction d'appartenance et celle du quantificateur. Delgado et al. (2000) montrent que l'ensemble des méthodes proposées pour calculer $C(A,i)$, cardinalité floue en i du sef A défini sur X, sont réductibles au calcul de l'adéquation entre deux sef par :

$$t(Qx \text{ sont } P) = \perp_{i=0\ldots n} \top \left(Q(i), C\left(A,i\right)\right) \tag{2.1}$$

où $\perp$ représente une t-conorme appliquée à l'ensemble des t-normes calculées entre l'appartenance de i entre 0 et n aux sef du quantificateur Q et de la cardinalité de A. $C(A,i)$ peut également être interprétée comme la possibilité que A contienne i éléments (Dubois & Prade, 1985b).

Les méthodes proposées dans ce cadre sont possibilistes ou probabilistes en fonction du couple de t-norme et t-conorme utilisé (Delgado et al., 2014). Les premières utilisent $\top_Z(x,y) = \min(x,y)$ et $\perp_Z(x,y) = \max(x,y)$ et les secondes $\top_P(x,y) = xy$

et $\perp_P(x, y) = x + y - xy$. Dans le cadre possibiliste par exemple, l'éq. (2.1) peut s'interpréter comme la plus grande valeur de l'intersection entre le quantificateur et la cardinalité pour chaque i considéré.

Différentes extensions de l'éq. (2.1), originellement introduite pour le protoforme « Qx sont P », ont été proposées pour prendre en compte le protoforme « QRx sont P » ainsi que les quantificateurs absolus et relatifs.

Dans l'annexe C p. 215 nous proposons une méthode alternative basée sur la similarité de Jaccard entre le quantificateur et la cardinalité floue. Cette dernière cependant ne donne pas de résultat probant.

2.3.3 Extensions du paradigme flou pour le calcul du degré de vérité

En plus des alternatives au calcul de la valeur de vérité basé sur la cardinalité σ-count, d'autres approches construites comme des extensions de la logique floue ont été introduites pour permettre de vérifier les propriétés de cohérence (cf. section 2.3.1 p. 33).

Ces dernières découlent directement de celles d'idempotence $A \wedge A = A$, du tiers exclu $A \vee \neg A = X$ et de non contradiction $A \wedge \neg A = \varnothing$, où $\vee$ est l'opérateur de disjonction logique modélisée en logique floue par la t-conorme, $\wedge$ celui de conjonction logique modélisé par la t-norme et $\neg$ celui de négation logique. Or, comme rappelé par Delgado et al. (2014) et montré par Dubois & Prade (1985a), il n'existe pas de couple de t-norme et de t-conorme en logique floue standard permettant de les vérifier simultanément.

Les ensembles crisp néanmoins satisfont naturellement ces propriétés, ce qu'exploitent les méthodes proposées dans les deux paragraphes suivants. La première est basée sur l'usage d'une fonction inverse d'appartenance et la seconde sur un mécanisme de fuzzification des quantificateurs.

Fonction inverse d'appartenance

Le principe de la fonction inverse d'appartenance est introduit dans différents travaux : *fuzzy bag* (Rocacher & Bosc, 2005), nombre graduel (Dubois & Prade, 2008), représentation par niveaux (Sánchez et al., 2009) et représentation X-mu (Martin, 2013).

La fonction inverse d'appartenance associe aux degrés d'appartenance dans [0,1] des parties de l'univers de discours, à l'inverse de la fonction d'appartenance classique qui associe aux éléments de l'univers des degrés d'appartenance.

Les degrés d'appartenance auxquels sont associés des éléments de U peuvent être continus dans le cas de X-mu ou discrets pour les autres méthodes. En ce sens, cette approche reprend le principe des α-coupes du sef pour sa définition.

Les opérations réalisées sur ces ensembles crisp possèdent donc les caractéristiques de la logique classique et donc les propriétés de cohérence. L'interprétation en termes flous est réalisée par reconstruction de la fonction d'appartenance correspondante.

Le point faible de ces approches réside dans ce dernier point car la fonction ainsi reconstruite n'est pas toujours convexe et ne correspond plus alors à un sef (Delgado

et al., 2014). Dans l'exemple considéré par Martin (2013), l'évaluation de $A \cap \overline{B}$ n'est pas un sef mais une représentation pour laquelle l'élément 4 de l'univers de discours appartient entre 0,3 et 1 au résultat, en contradiction donc avec la convexité d'un sef qui implique que si un point appartient à un certain degré au sef alors il y appartient également pour tous les degrés positifs inférieurs.

Nous n'avons donc pas investigué plus avant ces méthodes qui, en dépit de leur intérêt pour les propriétés de cohérence, ne garantissent plus l'interprétabilité des sef classiques.

Fuzzification des quantificateurs

Cette approche, proposée par Glöckner (1997) puis détaillée par Glöckner & Knoll (2001), repose sur une définition axiomatique de la notion de quantificateur. Elle a l'intérêt majeur d'en permettre la définition d'un nombre important vérifiant l'ensemble des propriétés listées dans la section 2.3.1 p. 32. Ainsi, en plus des quantificateurs absolus ou relatifs (cf. section 1.2.3 p. 12) utilisés dans les RLF standards, cette méthode permet d'utiliser l'ensemble des quantificateurs généralisés de Barwise & Cooper (1981) présentés plus en détail au chapitre 3. Les degrés de vérité de phrases comme « Il y a environ deux R_1 de plus que de R_2 qui sont P » ou « Environ deux fois plus de R_1 sont P_1 que de R_2 sont P_2 » sont par exemple calculables dans ce cadre (Delgado et al., 2014).

La méthode de Glöckner utilise des quantificateurs semi-flous (*semi-fuzzy quantifiers* ou *SFQ*) et de mécanismes de fuzzification des quantificateurs (*quantifier fuzzification mechanism* ou *QFM*).

Les SFQ sont des quantificateurs dont les entrées sont crisp et le résultat flou, i.e. dans [0,1]. Leur intérêt réside dans leur définition directement basée sur celles des quantificateurs généralisés, permettant de ne pas avoir à les définir de façon subjective et non consensuelle (Glöckner & Knoll, 2001). De plus, l'utilisation d'ensembles crisp en entrée des SFQ leur permet de vérifier simplement les lois de la logique classique.

L'argument est néanmoins discutable dans la mesure où la propension de la logique floue à prendre en compte cette subjectivité dans un modèle, au plus proche des intentions de l'utilisateur, est habituellement présentée comme un des ses avantages.

Le mécanisme de fuzzification des SFQ, QFM, permet de les convertir en quantificateurs flous selon une axiomatique détaillée permettant d'en conserver les propriétés souhaitées. Bien que différentes méthodes aient été proposées pour fuzzifier les quantificateurs linguistiques (Liu & Kerre, 1998), l'approche QFM est la seule permettant de vérifier l'ensemble des propriétés souhaitables énoncées plus haut, notamment pour les protoformes du type « QRx sont P » (Glöckner & Knoll, 2001). En plus de ceux présentés par Glöckner & Knoll (2001), Díaz-Hermida & Bugarín (2010) définissent un nouveau QFM basé sur une approche probabiliste de la quantification.

Ainsi, la méthode de Glöckner suppose que les quantificateurs sont prédéfinis et acceptés et déplace donc la question de leur définition vers celle de la création d'un mécanisme pour les fuzzifier.

Si l'intérêt théorique des approches par QFM est évident, leur utilisation pratique

est moins claire du fait notamment de la multiplicité des quantificateurs qu'elle permet. En effet, le problème de la brièveté du résumé posé dans le cas des RLF standards et présenté dans la section suivante est ici plus critique encore du fait du nombre de phrases générables.

D'autre part, l'interprétabilité de certaines d'entre elles mérite d'être étudiée. La phrase d'exemple donnée plus haut en est un bon exemple : l'interprétation de « Environ deux fois plus de R_1 sont P_1 que de R_2 sont P_2 » est-elle triviale pour tous les utilisateurs ?

2.4 Résumé

Les propriétés du résumé peuvent s'appliquent soit à sa globalité soit à des sous-groupes de ses phrases, comme présenté dans les deux paragraphes suivants.

2.4.1 Propriétés sur l'ensemble du résumé

Brièveté

La mesure de brièveté du résumé est construite sur une mesure de sa longueur dans le but d'en limiter la taille. Castillo-Ortega et al. (2012) proposent de la calculer simplement comme l'inverse de son nombre de phrases. Comme indiqué dans la section 2.2.2 p. 31, Kacprzyk & Zadrozny (2005b); Wilbik (2010, p.86) proposent des méthodes de calcul de la longueur d'une phrase, sans toutefois en décrire l'agrégation pour la taille globale du résumé.

Une approche classique pour réduire la taille d'un résumé est de n'en retenir que les phrases dont une ou plusieurs mesures de qualité sont au-dessus d'un certain seuil, comme détaillé dans la section 2.5.2 p. 42.

Couverture

La mesure de la couverture est effectuée à trois niveaux dans les RLF : au niveau vocabulaire pour tester a priori la couverture d'une VL (cf. section 2.1.1 p. 27), au niveau de la phrase pour en tester la couverture a posteriori (cf. section 2.2.2 p. 30) et enfin au niveau résumé.

Castillo-Ortega et al. (2012) proposent, dans le cadre des RLF de séries temporelles, de mesurer la couverture des phrases comme l'intervalle temporel qu'elles recouvrent et d'arrêter la production de nouvelles phrases lorsque l'ensemble des données sont couvertes.

La question de la couverture est critique en termes d'interprétabilité, comme illustré dans les résultats de notre système de RLF sur les ventes de livre Amazon décrit dans l'annexe B p. 211. Dans cet exemple, la couverture du résumé généré n'est pas entièrement satisfaisante car seuls les livres ayant un mauvais classement sont pris en compte, et l'utilisateur peut se demander si toutes les données ont bien été prises en compte par exemple. Une discussion concernant la couverture des résumés est donnée en perspectives de cette thèse.

Précision

La précision définie par Castillo-Ortega et al. (2012) est simplement calculée comme la moyenne du degré de vérité de l'ensemble des phrases du résumé.

Spécificité

La mesure de spécificité développée pour les sef d'une VL, présentée dans la section 2.1.1 p. 28, est transposée au niveau résumé par le calcul de la moyenne de leur imprécision.

Cette mesure est calculée a posteriori car bien que l'imprécision soit calculable a priori puisque basée sur la définition des sef de Q et P, le nombre de phrases générées n'est connu qu'a posteriori.

Mesure globale

Castillo-Ortega et al. (2012) proposent également un opérateur de comparaison entre résumés. Précisément, un résumé est supérieur à un autre si toutes les mesures de couverture, brièveté, spécificité et précision sont supérieures à celles du résumé comparé. Si seul un sous-ensemble des mesures est supérieur à celles de l'autre résumé, ils sont déclarés inclassables.

Une approche plus simple et plus générale pourrait être simplement de calculer une moyenne pondérée de chacune de ces mesures. Aucune méthode à notre connaissance ne propose toutefois cette approche.

Propriétés linguistiques

Le système que nous avons développé pour tester les méthodes de RLF (cf. annexe B p. 211) montre que même lorsque les VL, les quantificateurs et les protoformes utilisés sont définis de manière interprétable, la phrase résultante ne l'est pas nécessairement. L'exemple de la méthode proposée par Štěpnička et al. (2010, 2011) illustre également ce constat. De la même manière, plusieurs phrases interprétables isolément ne le sont pas toujours ensemble.

Ainsi, l'utilisation de méthodes plus développées de génération de texte, comme celles portées par la communauté NLG, semble particulièrement pertinente à cet égard.

2.4.2 Propriétés des sous-groupes de phrases

D'autres propriétés sont définies non sur le résumé dans sa globalité mais sur des sous-ensembles de phrases qui le constituent. Les méthodes présentées dans ce paragraphe visent à améliorer l'interprétabilité du résumé généré par l'élimination de ses phrases redondantes. Nous présentons deux méthodes dans ce sens basées sur la détection d'inclusion et de similarité dans les phrases.

Inclusion

L'inclusion décrit une situation où le quantificateur ou le résumeur d'une phrase est inclus dans le quantificateur ou le résumeur d'une autre phrase. Par exemple, « La plupart des jeunes sont bien payés » est inclus dans « Plus de la moitié des jeunes sont bien payés » puisque le quantificateur *La plupart* est inclus dans *Environ la moitié*. De même, « La plupart des employés gagnent environ 3000€ » est inclus dans « La plupart des employés gagnent plus de 2000€ » par inclusion du résumeur *Environ 3000€* dans *Plus de 2000€*.

Afin de diminuer la redondance du résumé et en augmenter sa précision, Bodenhofer & Bauer (2005); Pilarski (2010) recommandent de ne conserver que les phrases incluses.

En cas d'inclusion du qualifieur, la phase contenant le plus large est retenue. Considérons les phrases « La plupart des employés entre 20 et 25 ans gagnent environ 3000€ » et « La plupart des employés de moins de 50 ans gagnent environ 3000€ » : le qualifieur *Employés entre 20 et 25 ans* est inclus dans *Employés de moins de 50 ans* et c'est la seconde phrase qui est alors retenue afin de maximiser la couverture de la phrase. Un compromis doit donc être trouvé entre la précision du quantificateur et le degré de vérité : plus le quantificateur est précis, plus la phrase générée l'est aussi, mais plus le degré de vérité est faible. A l'inverse, une phrase basée sur quantificateur plus général aura un degré de vérité plus élevé mais sera moins informative.

Díaz-Hermida & Bugarín (2010) soulignent que l'élimination des phrases les moins spécifiques permet d'augmenter l'interprétabilité du résumé. En effet, l'utilisateur peut trouver confus l'occurrence des deux phrases « La plupart des jours d'avril sont chauds » et « La plupart des jours d'avril sont très chauds » avec une valeur de vérité équivalente. La suppression de la première, moins précise que la seconde, permet d'éliminer cette imprécision.

Enfin, la détection des inclusions sert également à éviter les problèmes associés aux sens induits de certaines phrases (Lesot et al., 2016). Par exemple, la phrase « La plupart des gros livres se vendent mal » laisse à penser que les petits se vendent mieux. Or il est possible que tous les livres se vendent mal, indépendamment de leur taille. La suppression des phrases à moindre couverture permet donc d'éviter ce type de malentendu.

Dans l'approche SaintEtiQ à l'inverse (Raschia & Mouaddib, 2000), les phrases incluses sont conservée mais présentées sous forme arborescente pour permettre à l'utilisateur une navigation dans les données depuis les phrases les plus générales vers les phrases les plus spécifiques, comme détaillé dans la section 2.5.3 p. 43.

Similarité

La détection des phrases similaires permet également d'éliminer les phrases redondantes d'un résumé. Wilbik & Keller (2012) proposent une mesure de similarité entre phrases basée sur la plus petite similarité entre leurs quantificateurs, qualifieurs et résumeurs ainsi que leurs valeurs de vérité. En fonction de la mesure de similarité utilisée, la détection des phrases similaires peut englober celle des phrases incluses (Lesot et al.,

2016).

Lorsque deux phrases sont similaires au-delà d'un seuil fixé par l'utilisateur, l'une d'entre elles peut être supprimée. Dans ce cas également, le filtrage réalisé doit répondre à un compromis couverture / degré de vérité, i.e. plus la phrase couvre un grand nombre de données plus son degré de vérité est élevée mais moins elle est spécifique.

Inférence

Dans certains cas, la causalité entre plusieurs phrases d'un résumé peut être soulignée afin d'en faciliter la compréhension (Zadeh, 1985). Par exemple, si deux phrases issues des protoformes « $Q_1 Rx$ sont P » et « $Q_2(R$ et $P)x$ sont S » ont un degré de vérité élevé, alors la phrase « $(Q_1 * Q_2)x$ sont P et S » peut être déduite, où $*$ représente une multiplication floue. Considérons « La plupart des étudiants sont jeunes » et « La plupart des jeunes étudiants sont célibataires », alors « Presque tous les étudiants sont jeunes et célibataires » peut être déduite, en supposant que *LaPlupart∗LaPlupart=PresqueTous.*

La détection d'une relation de déduction entre deux phrases peut donner lieu à leur agrégation linguistique. Cette technique, issue des approches GAT décrites dans la section 1.1.2 p. 9, est utilisée par Portet et al. (2007) qui renvoient par exemple la phrase « FiO2 augmente *donc* la saturation augmente » si une causalité est détectée entre « FiO2 augmente » et « la saturation augmente ».

2.5 Système de RLF

Comme l'illustrent les deux méthodes de génération des RLF présentées dans cette section, la qualité d'un résumé est également dépendante des algorithmes de génération retenus. La première s'inscrit dans un paradigme de questions / réponses où l'utilisateur fournit les paramètres du protoforme à partir desquels le système génère la phrase et renvoie un ensemble de mesures de qualité, dont le degré de vérité.

La seconde vise à renvoyer l'ensemble des phrases possibles à partir des données et du vocabulaire. Différentes méthodes de filtrage sont alors exploitées pour rendre accessible le résultat.

Enfin, le troisième paragraphe décrit différents modes d'organisation des phrases permettant également de faciliter la compréhension globale du résumé.

2.5.1 Questions / réponses

L'implication de l'utilisateur dans la génération du résumé augmente son interprétabilité, notamment par les choix qu'il fait dans la modélisation du vocabulaire utilisé ainsi que par la question qu'il pose au système (Kacprzyk & Zadrozny, 2013a).

Les systèmes proposés selon ce fonctionnement exigent généralement qu'en plus des différentes VL et quantificateurs du système de RLF, l'utilisateur spécifie aussi tout ou partie des paramètres du ou des protoformes à instancier. Pour un protoforme « QRx sont P »

par exemple, l'utilisateur peut spécifier le quantificateur Q et/ou le qualifieur R et/ou le résumeur P. Si les trois paramètres sont spécifiés, le système n'a qu'à calculer la valeur de vérité de la phrase instanciée. Si un ou deux paramètres sont donnés, alors le système évalue la valeur de vérité des phrases instanciées à partir des différentes valeurs possibles du paramètre libre. Si aucun paramètre enfin n'est précisé, alors le système procède à la génération exhaustive des phrases, décrite au paragraphe suivant. Une classification des différentes phrases à extraire en fonction des variables données par l'utilisateur est donnée par Kacprzyk & Zadrozny (2005a).

Différents systèmes sont développés suivant ce principe, comme FQuery (Kacprzyk & Zadrozny, 1994), SummarySQL (Rasmussen & Yager, 1997) ou Quantirius (Pilarski, 2010). Les tableaux de l'annexe A p. 209 en donnent une liste plus complète.

Cette approche a le mérite d'être facile à comprendre, simple à implémenter et rapide à exécuter. Elle ne permet en revanche pas de découverte réelle d'informations dans les données et ne présente que des confirmations ou des infirmations à l'utilisateur.

2.5.2 Génération exhaustive

La seconde approche a précisément pour vocation de trouver de l'information inconnue, inattendue dans les données. Dans ce cadre, l'ensemble des phrases possibles pour un vocabulaire et un jeu de données est généré.

Dans la plupart des systèmes de ce type, l'utilisateur spécifie le vocabulaire, puis le système extrait l'ensemble des phrases à partir des données et calcule pour chacune d'elles un degré de vérité et dans certains cas d'autres mesures de qualité, comme présentées dans les sections 2.2 p. 29 et 2.4 p. 37.

Le système FFS détaillé dans l'annexe B p. 211 simplifie ce processus de spécification du vocabulaire en le générant automatiquement sous forme de partitions de Ruspini. Seuls certains paramètres comme le nombre de modalités par VL, la spécificité des modalités (cf. section 2.1.1 p. 28) et les quantificateurs doivent être précisés. Concernant ces derniers, Castillo-Ortega et al. (2011a) proposent de les créer automatiquement, à partir d'un modèle initial translaté sur l'axe des valeurs quantifiables.

Quelle que soit la méthode retenue, l'interprétabilité d'un résumé généré de la sorte n'est pas garantie. En effet, même sur la vingtaine de phrases générées par FFS sur l'exemple décrit en annexe B p. 211, un certain nombre de questions se posent : certaines modalités de VL semblent ne pas avoir été testées, certaines phrases sont peu compréhensibles ou portent un sens différent de celui évalué par le système, malgré l'usage de VL claires individuellement. Un ensemble de questions liées au résumé généré sont présentées en fin d'annexe B p. 211.

Nombre de phrases

Une des difficultés liées à la génération exhaustive des phrases est leur nombre potentiellement très important. En effet, pour v VL chacune composée de m modalités,

q quantificateurs, il est possible de générer qmv phrases pour le protoforme « Qx sont P » et $q(mv)^2$ pour « QRx sont P ».

La brièveté du résumé discutée dans la section 2.4.1 p. 37 est donc également un critère clé pour les performances. Deux stratégies détaillées dans les paragraphes suivants sont élaborées dans ce sens, l'une par l'usage de seuils sur des mesures de qualité et l'autre via des techniques d'abandon anticipé (*early abandon*).

Seuils

Seuil sur le degré de vérité Le seuil le plus couramment utilisé est basé sur le degré de vérité. En ce cas, les phrases dont le degré de vérité est inférieur au seuil ne sont pas intégrées dans le résumé.

En plus d'en diminuer la taille, ce seuil permet également d'éliminer des phrases dont l'interprétation n'est pas simple. En effet, que dire de « La plupart des jeunes sont grands (0,47) » ? D'une manière générale, les phrases dont le degré de vérité est autour de 0,5 sont peu interprétables. Concernant celles dont le degré de vérité est plus faible encore, la mesure d'informativité présentée dans la section 2.2.2 p. 31 en permet l'exploitation via une négation de la phrase.

De plus, l'utilisation d'un seuil sur le degré de vérité peut entraîner l'absence de phrases décrivant certains attributs dans le cas où la fonction de comptage sur les données à un degré d'appartenance aux quantificateurs utilisés inférieur au seuil.

Une solution est l'utilisation directe du résultat de la fonction de comptage pour générer « à la volée » le quantificateur correspondant. Si par exemple 72% des R sont P alors les phrases « 72% des R sont P » ou « Un peu plus de 70% des R sont P » peuvent être renvoyées. Une telle approche, discutée dans les perspectives de cette thèse, permet de renvoyer à la fois des phrases interprétables et de s'affranchir de la définition des quantificateurs et des problèmes de phrases absentes lorsqu'ils sont mal adaptés aux données.

Pilarski (2010) propose de calculer indirectement le seuil sur le degré de vérité en déterminant plutôt le nombre k de résumés à retenir ayant les plus hautes valeurs de vérité, k étant lui-même évalué à l'aide d'un RLF. Précisément, l'auteur propose de calculer les valeurs de vérité des protoformes « QRx sont P » et « QPx sont R » où P est l'ensemble crisp des k premiers résumés par ordre décroissant de valeur de vérité et R le sef des résumés dont la valeur est haute où *Haute* est une modalité définie par l'utilisateur. Les valeurs croissantes de k à partir de 1 sont testées jusqu'à retenir celle renvoyant la plus grande valeur du minimum des degrés de vérité des deux phrases.

Seuil sur le degré de focus Un seuil sur le degré de focus de la phrase (cf. section 2.2.2 p. 30) peut également être utilisé. Ce dernier permet notamment d'éviter de renvoyer une phrase comme « La plupart des jeunes sont grands » alors que les données ne contiennent qu'un seul individu « jeune ». Nous détaillons dans le paragraphe suivant une méthode d'abandon anticipé basée sur ce seuil.

Abandon anticipé

Afin de minimiser la taille du résumé, certaines approches préconisent l'abandon anticipé de l'évaluation des phrases dès qu'elles vérifient un critère donné, comme détaillé dans les paragraphes suivants.

Famille de quantificateurs et VL hiérarchiques La méthode de Castillo-Ortega et al. (2011a), également présentée dans la section 1.3.2 p. 17, propose d'utiliser un ensemble ordonné de q quantificateurs Q_i non décroissants. De plus, une hiérarchie de VL temporelles de la plus large, la décade, à la plus précise, l'année, est définie.

L'évaluation des degrés de vérité est effectuée en commençant par les intervalles temporels les plus larges, les décades, vers les plus précis, l'année. De même les quantificateurs sont évalués successivement du plus petit au plus grand. Dès que le degré de vérité d'une phrase est supérieur au seuil donné par l'utilisateur, l'évaluation des phrases est interrompue pour un résumeur donné. Par exemple, si « Au moins 80% des R sont P » a un degré de vérité suffisamment élevé, alors les phrases utilisant un quantificateur moins spécifique comme « Au moins 70% des R sont P » ou « Au moins 60% des R sont P » ont également un degré de vérité élevé. De la même manière que dans les cas d'inclusion (cf. section 2.4.2 p. 39), la phrase renvoyée a donc une couverture temporelle maximale et la plus grande précision en termes de quantification. Les autres phrases, soit moins précises soit couvrant moins de données, sont ignorées.

Degrés de vérité et de focus Dans la méthode proposée par Kacprzyk & Wilbik (2009), les degrés de vérité des phrases du type « Qx sont P » sont calculés dans un premier temps. Lors de l'évaluation des phrases basées sur le protoforme « QRx sont P », les résultats des fonctions de comptage obtenus à l'étape précédente sont réutilisés directement pour calculer le degré de focus, égal à ce résultat divisé par n, permettant ainsi de ne pas évaluer les phrases du type « QRx sont P » si le degré de focus de « Qx sont R » est inférieur au seuil spécifié par l'utilisateur.

De plus, pour les quantificateurs croissants, l'usage de t-normes garantit que $|R| \geq |\top(R, P)|$ et donc que $t(Qx \text{ sont } P) \geq t(QRx \text{ sont } P)$ (cf. éq. (1.2) p. 14), si bien que lorsque la valeur de vérité d'une phrase issue de « Qx sont P » est inférieure au seuil défini, il n'est pas utile de calculer le degré de vérité des phrases issues de « QRx sont P ».

A l'aide de ces optimisations, Kacprzyk & Wilbik (2009) parviennent à écarter entre 75% et 99% des phrases générables.

2.5.3 Organisation

L'organisation des phrases renvoyées par l'algorithme de génération permet également d'en faciliter l'interprétation. Ainsi, le système SaintEtiQ (Raschia & Mouaddib, 2002) calcule l'ensemble des phrases du résumé et les présente de manière hiérarchique, la phrase la plus générale, i.e. ayant la couverture la plus importante, étant placée à la racine de

l'arborescence. L'utilisateur peut décider de plonger dans telle ou telle branche de l'arbre en fonction des aspects des données qui l'intéressent.

Si l'approche ne supprime pas de phrases, le mode de présentation retenu facilite toutefois la lecture du résumé. De plus, comme l'utilisateur est sollicité pour sa recherche d'information, le résumé bénéficie des avantages de l'implication de l'utilisateur dans la compréhension, décrite plus haut dans le cadre des systèmes de question / réponse.

Pilarski (2010) propose également d'organiser les phrases du résumé de la plus générale à la plus spécifique. Pour ce faire, l'utilisateur choisit un protoforme général puis sélectionne un ou plusieurs des protoformes plus spécifiques proposés par le système. Un protoforme général porte par exemple sur une modalité d'une VL et ceux plus spécifiques utilisent cette modalité comme qualifieur et une autre modalité d'une autre VL comme résumeur. L'auteur donne cet exemple :

> Environ 50% des compagnies prises en compte sont petites *(général)*
> Environ 50% des petites compagnies ont un CA quotidien moyen *(spécifique)*
> La plupart des petites compagnies ont un bénéfice mensuel faible *(spécifique)*

2.6 Conclusion

Ce chapitre présente différentes mesures et propriétés associées aux différents éléments d'un RLF, à savoir son vocabulaire, constitué des variables linguistiques et des quantificateurs, ses phrases, le résumé les contenant, le degré de vérité et le système de génération à proprement parler.

Parmi ces propriétés et mesures, certaines sont évaluées de manière indépendante des données quand les autres les utilisent, certaines visent à améliorer l'interprétabilité du résumé, de la phrase, du vocabulaire tandis que d'autres enfin ont pour objet la validation des phrases au regard des données en entrée.

Enfin, si la majorité de ces mesures et propriétés s'appliquent aux RLF, certains systèmes étendant le paradigme de la logique floue classique sont également présentés, leur objectif étant notamment de vérifier les propriétés de cohérence des phrases fondées sur les lois de non contradiction et du tiers exclu. Ces propriétés font l'objet du prochain chapitre, qui discute notamment de leur validation dans le cadre de la logique floue classique.

Chapitre 3

Cohérence d'un résumé : analyses et modèle des oppositions

La tautologie et la contradiction sont vides de sens.

—Ludwig Wittgenstein, *Tractatus logico-philosophicus*

Comme discuté dans le chapitre précédent, une des composantes de la qualité, et plus précisément de l'interprétabilité des résumés linguistiques flous considérés globalement comme un ensemble de phrases, est l'absence d'opposition entre les phrases qui le constituent : elle doit par exemple empêcher un résumé contenant à la fois « La plupart des jeunes sont petits » et « La plupart des jeunes sont grands ». Pour éviter des telles contradictions, il est d'abord nécessaire de les identifier, ce qui constitue une tâche complexe du fait des nombreux degrés de libertés permis par les RLF.

Nous proposons de les organiser selon une vue hiérarchique dépendant de la complexité des phrases considérés, comme illustré sur la figure 3.1 : la section 3.1 décrit les formalismes permettant de représenter des oppositions entre phrases simples ou quantifiées avec les seuls quantificateurs classiques, $\forall$ et $\exists$. La section 3.2 p. 51 considère le cas des phrases construites avec des quantificateurs généralisés. La section 3.3 p. 53 discute le troisième degré de liberté, qui provient des différentes négations qui peuvent être définies dans le formalisme des prédicats flous.

Nous introduisons ensuite dans la section 3.4 p. 56 un modèle général de l'opposition entre phrases d'un résumé linguistique flou, représenté sous la forme d'une nouvelle structure, un cube en 4 dimensions, qui représente toutes les relations que l'on peut établir entre toutes les variantes de négation pouvant être considérées en utilisant les degrés de liberté existants dans les RLF.

Sur la base de ce modèle général, la section 3.5 considère la propriété de dualité pour la fonction de comptage, permettant de satisfaire les propriétés de cohérence dans le cadre de la logique floue standard.

Ces travaux ont été publiés dans (Moyse et al., 2015) et récompensés d'un prix du meilleur papier étudiant.

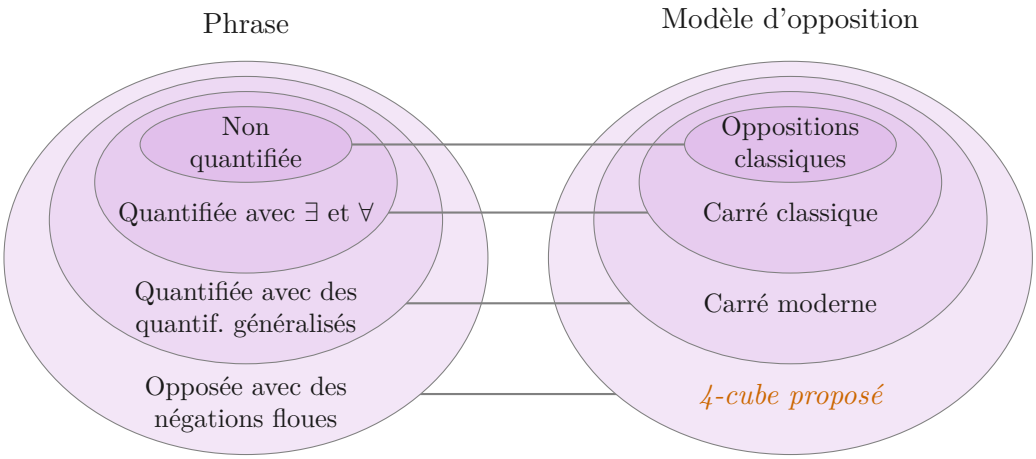


FIGURE 3.1 – Modèles d'oppositions entre phrases de complexité croissante

3.1 1^{er} niveau d'opposition : phrases simples et quantificateurs classiques

Le premier type d'opposition étudié dans cette section oppose deux phrases simples, sans quantificateur ni qualifieur. Celles-ci sont représentées par « S est P » et « S est P' » où S est un sujet et P, P' deux prédicats crisps, par exemple « Jean est grand » et « Jean est petit ». Le cas des prédicats flous est présenté dans la section 3.3 p. 53.

Ces phrases peuvent être rendues plus complexes par l'utilisation des deux quantificateurs classiques ∃ et ∀, i.e. « Tous les S sont P » et « Aucun S n'est P ». Ces phrases sont dites quantifiées et sont étudiées dans le deuxième paragraphe de cette section. Leurs oppositions sont modélisées au travers des carrés classiques et modernes, étudiés dans les deux paragraphes suivants.

Enfin, l'étude des oppositions entre plus de deux phrases simples ou quantifiées est présentée dans le dernier paragraphe de la section.

3.1.1 Opposition de phrases simples

Aristote a été le premier à étudier l'opposition entre phrases simple à l'aide des loi du tiers exclu (LTE) et de non contradiction (LNC) (Horn, 2002, p. 62). La LTE impose que « S est P ou non P » soit une tautologie et la LNC que « Aucun S n'est P et non P » en soit une également.

A l'aide de ces deux lois, Aristote propose les trois relations d'opposition classiques de contradiction, contraire et subcontraire pour les phrases simples du type « S est P ».

Contradiction Étant donné deux prédicats P et P', « S est P » et « S est P' » sont en relation de contradiction si S peut être soit P soit P' mais pas les deux. P et P' vérifient en ce cas la LNC et la LTE. Par exemple, « S est froid » et « S est non froid » sont en contradiction car S est soit froid soit non froid, mais ne peut vérifier les deux prédicats simultanément.

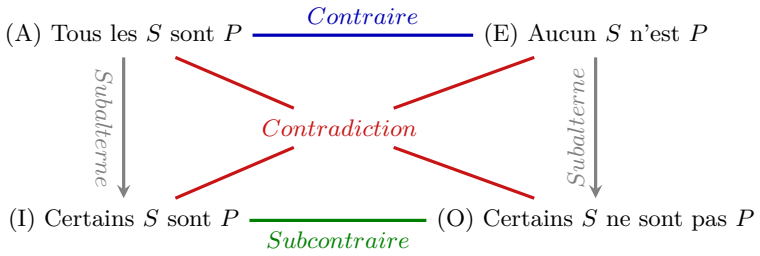


FIGURE 3.2 – Carré classique des oppositions

Contraire P et P' sont en relation de contraire si S peut n'être ni P ni P' mais ne peut être les deux à la fois. P et P' vérifient ici la LNC mais pas la LTE. « S est froid » et « S est chaud » sont en relation de contraire car S peut n'être ni chaud ni froid mais ne peut être chaud et froid.

Subcontraire P et P' sont en relation de subcontraire si S peut être P et/ou P' mais doit être au moins l'un des deux. P et P' vérifient ici la LTE mais pas la LNC. « S est non froid » et « S est non chaud » sont en relation de subcontraire car S peut être non chaud et non froid mais doit être au moins l'un des deux.

3.1.2 Carré classique des oppositions

Sur la base des trois relations définies pour les phrases simples, le carré classique ou aristotélicien présente graphiquement les oppositions entre des phrases quantifiées à l'aide des quantificateurs *Tous* et *Certains*, i.e. les phrases « Tous les S sont P », « Aucun S n'est P », « Certains S sont P » et « Certains S ne sont pas P ». Il convient de noter que S ici ne désigne plus un individu seulement, comme dans le cas précédent des phrases simples, mais un prédicat qu'un individu x peut ou non satisfaire. Ces phrases s'inscrivent donc dans le cadre des protoformes présentés dans la section 1.2.4 p. 12 avec les quantificateurs *Tous*, *Aucun* et *Certains*, le qualifieur S et les résumeurs P et *non* P.

Le carré aristotélicien est représenté sur la figure 3.2. Les sommets A/E sont en relation de contraire car les phrases « Tous les S sont P » et « Aucun S n'est P » ne peuvent être vraies simultanément mais peuvent être fausses tous les deux. De la même manière, A/O et E/I sont en relation de contradiction et I/O en relation de subcontraire.

Une relation additionnelle, dite de subalterne, est également présente dans le carré. Cette dernière n'est pas une opposition mais une implication. En effet, « Tous les S sont P » implique que « Certains S sont P » ; de même, « Aucun S n'est P » implique que « Certains S ne sont pas P ».

Par exemple, si S est une personne et P sa taille, le sommet A « Toutes les personnes sont grandes » est en relation de contradiction avec O « Certaines personnes ne sont pas grandes », de contraire avec E « Aucune personne n'est grande » et de subalterne avec I « Certaines personnes sont grandes ».

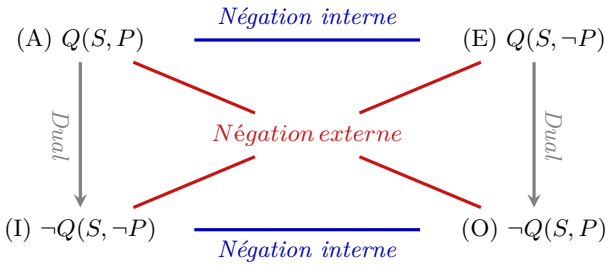


FIGURE 3.3 – Carré moderne des oppositions

3.1.3 Carré moderne des oppositions

Le carré moderne des oppositions illustré sur la figure 3.3 est proposé par George Boole et représente également de manière graphique les oppositions entre phrases quantifiées, mais diffère en plusieurs points du carré classique (Westerstahl, 2012).

Il utilise tout d'abord un formalisme générique avec la notation $Q(S,P)$ qui représente la phrase « Q S sont P ». « Tous les S sont P » est donc un cas particulier dans le carré moderne, avec $Q = Tous$, mais ce carré peut être défini avec un quantificateur général Q différent de *Tous* ou *Certains*. De plus, il n'est pas défini par les relations entre ses sommets comme le carré classique, mais par l'application de l'opérateur de négation à différentes parties de la phrase : la négation interne est la négation du prédicat P, la négation externe est celle de l'expression dans son ensemble, et le dual est la composition des négations interne et externe.

D'une manière générale, du fait de la diversité des quantificateurs qu'il permet, le carré moderne est plus expressif que le carré classique. Il lui est cependant équivalent dans le cas ou $Q = \forall$, illustré sur la figure 3.4. En ce cas en effet, le sommet A est défini comme $\forall x, Sx \rightarrow Px$. Le sommet E est ensuite obtenu par application de la négation sur P dans l'expression de A, soit $\forall x, Sx \rightarrow \neg Px$. Le sommet O est construit par application de la négation sur l'ensemble de l'expression de A, soit :

$$\begin{aligned} \neg\left(\forall x, Sx \rightarrow Px\right) &\Leftrightarrow \exists x, \neg\left(\neg Sx \vee Px\right) \\ &\Leftrightarrow \exists x, Sx \wedge \neg Px \end{aligned}$$

Le sommet I est déduit du sommet O par négation de P, ou de manière équivalente par négation de l'expression du sommet E, menant à $\exists x, Sx \wedge Px$.

Pour ce carré moderne défini avec $Q = \forall$, les relations du carré classique sont vérifiées. En effet, les sommets A et E sont bien en relation de contraire car ils vérifient la LNC mais pas la LTE. La LNC est vérifiée car $(\forall x, Sx \rightarrow Px) \wedge (\forall x, Sx \rightarrow \neg Px)$ est équivalente à $\forall x, (\neg Sx \vee Px) \wedge (\neg Sx \vee \neg Px)$ qui se simplifie en $\forall x, \neg Sx$, qui est toujours fausse dès lors que l'on considère qu'il existe au moins un x vérifiant S. Cette hypothèse, appelée import existentiel, est supposée vérifiée dans l'étude que nous présentons ici. Ce point a néanmoins été largement discuté, comme détaillé par Horn (2002, p.67).

Les sommets A et E ne vérifient pas en revanche la LTE car $(\forall x, Sx \rightarrow Px) \vee (\forall x, Sx \rightarrow$

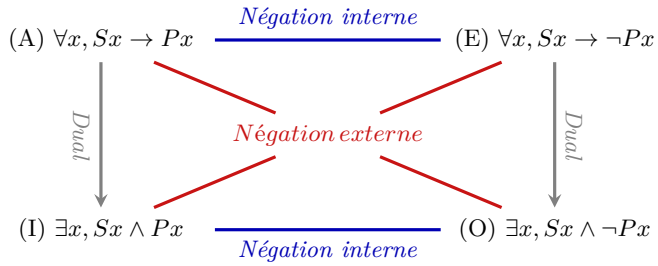


FIGURE 3.4 – Carré moderne des oppositions pour $Q = \forall$

$\neg Px$) est aussi équivalente à $\forall x, \neg Sx$, qui n'est pas une tautologie.

L'interprétation linguistique de A en « Tous les S sont P » et de E en « Aucun S n'est P » est en accord avec ces résultats car les deux phrases ne peuvent être vraies simultanément, vérifiant ainsi la LNC, mais peuvent être fausses simultanément, ne vérifiant pas alors la LTE.

Par des raisonnements similaires, il est simple de montrer que A et O vérifient la relation de contradiction, i.e. la LTE et la LNC, que I et O vérifient la relation de subcontraire, i.e. la LTE mais pas la LNC et enfin que A implique I et E implique O, vérifiant ainsi la relation de subalterne.

Toutefois, comme expliqué dans la section 3.2 p. 51, ces relations ne sont pas vérifiées pour tous les quantificateurs Q dans le carré moderne et plusieurs propositions sont réalisées afin de combiner les propriétés sémantiques des relations du carré classique avec la généralité et l'efficacité formelle du carré moderne.

3.1.4 Autres structures d'opposition

Cette sous-section présente deux autres structures d'opposition qui ont été développées pour permettre, respectivement, une interprétation différente du quantificateur *Certains* et l'utilisation de termes plus nombreux dans les phrases quantifiées.

Interprétation de « Certains » D'autres interprétations de $\forall$ et $\exists$ ont été proposées, donnant lieu à des représentations géométriques plus complexes que le carré : dans celles présentées plus haut, *Certains* est interprété comme « Certains, éventuellement tous » ; il peut néanmoins être vu comme « Certains, mais pas tous ».

Cette seconde interprétation est utilisée dans l'hexagone des oppositions de Blanché (1966) illustré sur la figure 3.5. Ce dernier contient deux nouveaux sommets, Y interprété comme « Ni tous ni aucun » et U « Tous ou aucun ». Il est intéressant de noter que les relations classiques (contradiction, contraire, subcontraire et subalterne) sont également vérifiées pour l'hexagone.

Dans le cadre d'une étude théorique générale de ce type de structures, Dubois & Prade (2012) montrent que la représentation en hexagone est possible dès que les cas A, E et Y sont mutuellement exclusifs et donc que $A \wedge E$, $A \wedge Y$ et $E \wedge Y$ sont faux et que $A \vee E \vee Y$ est une tautologie.

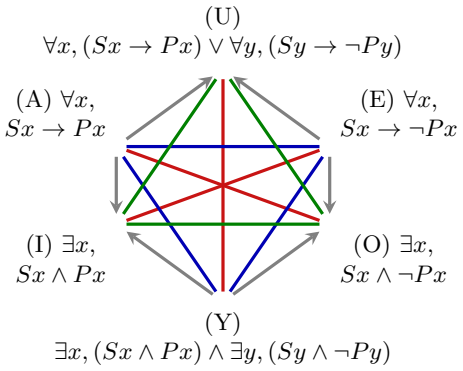


FIGURE 3.5 – Hexagone des oppositions de Blanché (1966). La relation de contradiction est représentée en rouge, celle de contraire en bleu, celle de subcontraire en vert et celle de subalterne en gris.

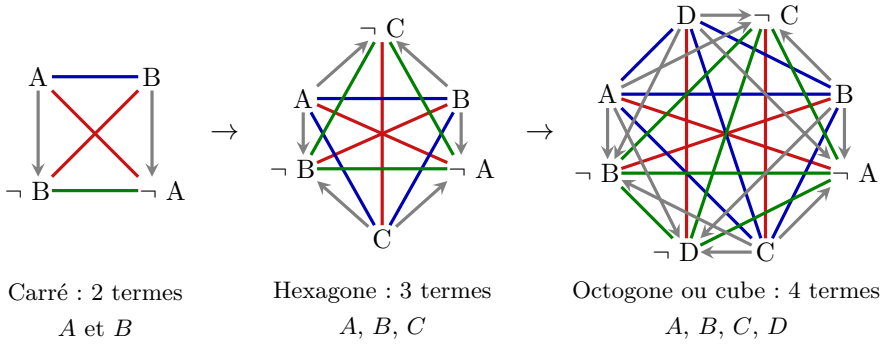


FIGURE 3.6 – Partitions de taille croissante vérifiant les relations d'opposition classiques (Moretti, 2011). Les couleurs utilisées sont identiques à celles de l'hexagone.

Phrases simples avec plus de deux termes Une autre extension des carrés logiques repose sur l'utilisation de termes autres que P et $non\,P$ pour des phrases simples. Une méthode de construction géométrique permettant de vérifier les relations classiques dans ce cadre est proposée par Moretti (2011) pour tout ensemble de termes $P_1, ..., P_n$ constituant une partition de l'univers de discours, i.e. , $\forall x$, $P_1(x) \vee ... \vee P_n(x)$ est vrai et $\forall i, j$, $P_i(x) \wedge P_j(x)$ est faux.

Dans cette approche, chaque élément P_i de la partition est lié aux autres au travers d'une relation de contraire. Le terme $\neg P_i$ en contradiction avec P_i est déterminé par symétrie centrale. Les relations de subcontraire sont les contradictions des contraires. Enfin, les relations de subalterne sont les sommets adjacents des autres sommets.

La figure 3.6 illustre la construction de structures de ce type avec un nombre croissant d'éléments mis en opposition. La figure permet de vérifier que les relations classiques sont en effet vérifiées. Dans le cube par exemple, représenté ici en deux dimensions sous la forme d'un octogone (figure de droite), $\{A, B, C, D\}$ est une partition de l'univers. Ainsi, en prenant le sommet A, les relations de contraire sont bien satisfaites avec B, C et D. La relation de contradiction est satisfaite par rapport à $\neg A$ et celle de subcontraire avec $\neg A$, $\neg B$, $\neg C$ et $\neg D$.

En considérant par exemple que S est la couleur d'un objet qui peut prendre deux valeurs, que A représente le *rouge* et B le *vert* dans le carré (figure de gauche), la phrase « si l'objet est *rouge* alors il n'est *pas vert* » est bien illustrée par la relation de subalterne de A vers B. Il ne peut pas non plus être *non rouge*, comme rappelé par le lien de contradiction entre A et $\neg A$. En ajoutant la couleur C pour le *bleu* dans l'hexagone (figure centrale), si un objet est *rouge* alors il est *non vert* et *non bleu* comme indiqué par les relations de subalterne partant de A vers B et C, et ainsi de suite pour les autres relations.

3.2 $2^{ème}$ niveau d'opposition : quantificateurs généralisés

Les quantificateurs généralisés constituent le deuxième niveau de la hiérarchie des structures d'opposition. En effet, bien que définis dans un formalisme de logique classique et donc moins souples les quantificateurs flous présentés dans la section 1.2.3 p. 12, ils ajoutent un degré de liberté dans les modes de quantification des phrases, comme détaillé dans la première sous-section. Dans la seconde, les liens qu'entretiennent ces quantificateurs avec les carrés logiques présentés dans la section précédente sont étudiés en détail.

3.2.1 Quantificateurs généralisés

Les quantificateurs généralisés, définis par (Mostowski, 1957), peuvent être vus comme les intermédiaires entre $\exists$ et $\forall$. Dans le cadre linguistique, le quantificateur généralisé $MoinsDe5(S, P)$ par exemple s'interprète « Moins de 5 S sont P », et sa valeur de vérité est évaluée par des opérateurs ensemblistes standards : $MoinsDe5(S, P) \Leftrightarrow |S \cap P| < 5$, où P désigne à la fois le prédicat et le nombre d'éléments qui le vérifient. D'autres exemples de l'interprétation ensemblistes des quantificateurs généralisés, absolus ou relatifs, sont (Barwise & Cooper, 1981) :

$$\begin{array}{rlrl} Tous\,(S,P) & \Leftrightarrow S \subseteq P & PlusDe20\%\,(S,P) & \Leftrightarrow |S \cap P| \,/\, |S| > 0.2 \\ Aucun\,(S,P) & \Leftrightarrow S \cap P = \varnothing & LaPlupart\,(S,P) & \Leftrightarrow |S \cap P| > |S - P| \end{array}$$

Le mécanisme de fuzzification des quantificateurs proposé par Glöckner (1997) et présenté dans la section 2.3.3 p. 36 utilise ces définitions pour créer des quantificateurs flous.

3.2.2 Liens avec les carrés logiques

Les quantificateurs généralisés peuvent être utilisés directement dans le carré moderne des oppositions puisqu'il est justement conçu sur la base d'un quantificateur générique, contrairement au carré classique construit sur les quantificateurs *Aucun*, *Certains* et *Tous* (Westerstahl, 2012). Le carré moderne de *MoinsDe5* illustré sur la figure 3.7 est ainsi construit sur le base de l'expression $|S \cap P| < 5$ pour le sommet A puis sur les négations interne et externe définies dans la section 3.1.3 p. 48. Ainsi, le sommet E est déterminé en appliquant sur le sommet A l'opérateur de négation sur P et par la relation $|S \cap \neg P| < 5 \Leftrightarrow |S - P| < 5$, le sommet O par application de la négation sur l'ensemble

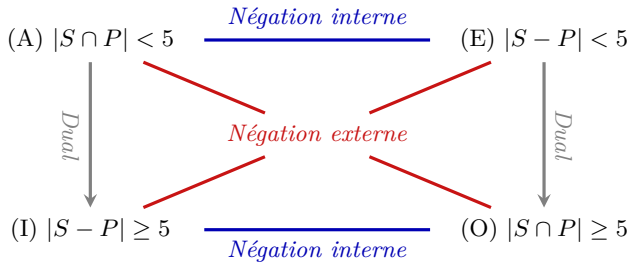


FIGURE 3.7 – Carré moderne de $MoinsDe5(S, P)$

de l'expression et par la relation $\neg(|S \cap \neg P| < 5) \Leftrightarrow |S \cap \neg P| \geq 5$ et le sommet I par application sur le sommet O de la négation sur P.

Si le carré moderne permet de construire les oppositions de toute phrase construite avec un quantificateur généralisé, il ne permet cependant pas de vérifier les propriétés des oppositions classiques, i.e. contradiction, contraire et subcontraire (Brown, 1984). Sur le carré de la figure 3.7, les sommets I et O ne sont pas en relation de subcontraire car ils peuvent être faux tous les deux, si 4 S sont P et 4 S sont non P par exemple.

La non vérification des relations classiques concerne également les quantificateurs relatifs. $MoinsDeDeuxTiers(S, P)$ par exemple ne vérifie pas la relation de contraire entre les sommets $A = |S \cap P| < 2\,|S|\,/3$ et $E = |S - P| < 2\,|S|\,/3$ puisque les deux peuvent être vrais simultanément, si 50% des S sont P.

Comme annoncé plus haut, différentes tentatives ont été développées visant à retrouver les relations classiques, sémantiquement riches, dans le carré moderne, plus générique et mieux défini formellement. Peterson (1979) propose l'utilisation de deux carrés vérifiant ces relations, l'un avec *Peu* et *Beaucoup* et l'autre avec *Beaucoup* et *La Plupart*. Dans cette approche cependant les deux carrés ne peuvent être utilisés simultanément car *Beaucoup* est dans le premier en contradiction avec *La Plupart* mais avec *Peu* dans le second. De plus, *Beaucoup* doit être défini comme n'incluant pas plus de 50% des individus, ce qui peut sembler contre-intuitif.

La majorité des autres propositions pour vérifier les relations classiques dans le carré moderne ont été définies dans des contextes graduels, i.e. où la valeur de vérité des propositions est comprise dans [0,1].

En plus de la proposition est celle de Glöckner (1997) déjà évoquée dans la section 2.3.3 p. 36, les quantificateurs intermédiaires développés dans la théorie des types flou (Novák, 2008) sont utilisés pour modéliser plusieurs des quantificateurs généralisés en satisfaisant les relations d'opposition classiques. Murinová & Novák (2014) détaillent comment ces quantificateurs peuvent être calculés à partir de modificateurs utilisés avec $\forall$ et $\exists$.

Dubois et al. (2015) proposent l'extension des carrés avec conservation des propriétés classiques pour différents modèles graduels comme les théories de l'évidence ou des possibilités. Ces modèles généraux ne sont cependant pas applicables au contexte des RLF car des expressions du degré de vérité des phrases ne sont pas disponibles dans ce cadre.

Ainsi, les quantificateurs généralisés, s'ils permettent une plus grande expressivité

rendent la modélisation des oppositions plus complexe en ne vérifiant pas toujours les relations classique d'opposition.

3.3 $3^{ème}$ niveau d'opposition : négations floues

Les deux premiers niveaux de la hiérarchie précédents utilisent les opérateurs ensemblistes et la logique classique pour modéliser l'opposition entre phrases, avec un opérateur unique de négation. L'un des attraits de la logique floue est l'utilisation de différentes négations, permettant de prendre en compte des cas d'opposition plus variés.

Dans la suite de cette section, les prédicats sont flous, donc représentés par leur fonction d'appartenance. En notant un sef défini sur l'univers numérique $[a^-; a^+]$, $A(x)$ représente le degré auquel « x est A », à rapprocher de « S est P » dans la section 3.1.1 p. 46 avec $S = x$ et $P = A$.

Après un rappel de l'opérateur de négation floue standard, nous présentons dans cette section les trois négations définies pour les prédicats en logique floue, le complément, l'anonyme et l'antonyme complément ainsi que leurs liens avec les relations classiques d'opposition.

3.3.1 Opérateur de négation

Étant donné un intervalle $I = [i^-, i^+]$, l'opérateur standard de négation floue n est défini par :

$$\forall x \in I,\ n\,(I, x) = i^+ - i^- - x \tag{3.1}$$

La négation correspond donc une involution de I dans I. Dans le cadre d'une représentation graphique, elle est équivalente à la symétrie d'axe $x = (i^+ + i^-)/2$. Nous l'écrirons par la suite $n(x)$ ou $\overline{x}$ lorsque I n'est pas ambigu.

3.3.2 Complément

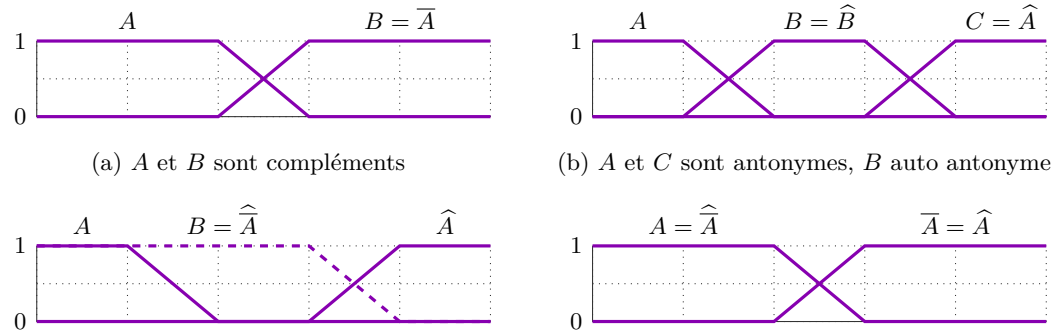
Le complément $\overline{A}$ d'un sef A est défini par :

$$\overline{A}(x) = n([0, 1], A(x)) = 1 - A(x) \tag{3.2}$$

i.e. la négation de la fonction d'appartenance $A(x)$ définie dans $[0, 1]$.

Le complément flou est le correspondant de la négation en logique classique, égal à cette dernière pour un ensemble crisp. Il est également involutif, i.e. $\overline{\overline{A}} = A$. Enfin, comme illustré sur le graphique `(a)` de la figure 3.8, $\overline{A}$ est le symétrique de A par rapport à l'horizontale $y = 1/2$.

Il faut noter que, bien que le plus souvent, $A \neq \overline{A}$, il existe un sef tel que $A = \overline{A}$, ce qui rend cette forme de négation contre-intuitive. Toutefois, ce sef a pour fonction d'appartenance $A(x) = 1/2$ pour tout $x \in [a^-, a^+]$: il n'est jamais utilisé en pratique, car dénué d'intérêt dans un cadre linguistique et donc en particulier dans le cadre des RLF.



(a) A et B sont compléments

(b) A et C sont antonymes, B auto antonyme

(c) A et B (pointillés) sont antonymes compléments

(d) A est auto antonyme complément

FIGURE 3.8 – Compléments (a), antonymes (b) et antonymes compléments (c et d)

3.3.3 Antonyme

L'antonyme d'un sef A est défini par :

$$\widehat{A}(x) = A(n([a^-, a^+], x)) = A(a^+ - a^- - x) = A(\overline{x}) \tag{3.3}$$

donc comme la négation du paramètre x défini sur $[a^-, a^+]$. Il est involutif, i.e. $\hat{\widehat{A}} = A$.

Comme illustré en `(b)` sur la figure 3.8, $\widehat{A}$ est le symétrique de A par rapport à la verticale $x = (a^+ - a^-)/2$.

On peut noter, à nouveau, que bien que le plus souvent $B \neq \widehat{B}$, il existe des sef tels que $B = \widehat{B}$. C'est par exemple le cas des modalités centrales dans des partitions de Ruspini comportant un nombre impair de modalités de même taille (Mesiar & Stupnanová, 2015). Nous appelons ces sef auto-antonymes.

3.3.4 Antonyme complément

L'antonyme complément (a.c.) $\widehat{\overline{A}}$ est la composition de l'anonyme et du complément :

$$\widehat{\overline{A}}(x) = 1 - A(a^+ - a^- - x) = \overline{A}(\overline{x}) \tag{3.4}$$

Il est commutatif, donc $\widehat{\overline{A}} = \overline{\widehat{A}}$ (De Soto & Trillas, 1999). Comme illustré en `(c)` sur la figure 3.8, $\widehat{\overline{A}}$ est le symétrique de A par rapport au point $((a^+ - a^-)/2, 1/2)$. De plus, si $A = \widehat{\overline{B}}$, alors $\widehat{A} = \overline{B}$, $\overline{A} = \widehat{B}$, $\widehat{\overline{A}} = B$ et si $A = \widehat{\overline{A}}$ alors A est un auto antonyme complément, comme illustré sur le graphe `(d)`.

3.3.5 Liens entre les relations classiques et les négations floues

Afin d'étudier les relations entre les trois négations floues présentées ci-dessus (antonyme, complément et antonyme complément) et les relations classiques d'opposition (contradiction, contraire, subcontraire et subalterne), nous définissons ici la LNC et la LTE en termes flous à l'aide de l'opérateur de négation classique $\neg$, qui peut correspondre à l'une des trois négations floues selon que nous cherchons à vérifier la LNC ou la LTE.

Comme la t-norme et la t-conorme sont utilisées en logique floue pour représenter respectivement la conjonction et la disjonction logique, une première définition stricte pour la LNC est $\top(A, \neg A) = 0$ et pour la LTE $\bot(A, \neg A) = 1$ (Delgado et al., 2014).

Nous proposons deux définitions plus souples, $\top(A, \neg A) \leq 0.5$ pour la LNC et $\bot(A, \neg A) \geq 0.5$ pour le LTE. La valeur $0,5$ est ici prise comme valeur seuil, en dessous duquel une proposition peut être considérée comme fausse et au-dessus duquel elle peut être vue comme vraie. L'intérêt de ces dernières définitions et leur capacité à vérifier les relations classiques d'opposition dans le cadre de partitions de Ruspini avec un couple dual de t-norme et t-conorme.

D'abord, le complément flou est l'équivalent de la contradiction classique car il vérifie la LTE et la LNC. En effet, du fait des propriétés des partitions de Ruspini, $\min(A, \overline{A}) \leq 0.5$ ce qui implique que $\top(A, \overline{A}) \leq 0,5$ donc que le complément vérifie la LTE. De plus, comme $\top(A, \overline{A}) = 1 - \bot(\overline{A}, A)$ par dualité du couple t-norme/t-conorme, on a $\bot(A, \overline{A}) \geq 0.5$, ce qui satisfait la LTE. D'un point de vue linguistique, les couples d'étiquettes en relation de complément flou sont par exemple *chaud / non chaud*, *beaucoup / pas beaucoup*.

L'antonyme flou de A vérifie la relation de contraire puisqu'il satisfait la LNC à condition que A et $\widehat{A}$ sont des modalités d'une VL. Cette condition est recommandée dans la définition d'une variable linguistique (Díaz-Hermida & Bugarín, 2010) comme indiqué dans la section 2.1.1 p. 27. En effet, deux modalités d'une partition de Ruspini vérifient $\top(A, \widehat{A}) = 0$ si elles ne sont pas adjacentes et $\top(A, \widehat{A}) \leq 0.5$ sinon, vérifiant donc la LNC dans les deux cas. Cette loi n'est évidemment pas vérifiée si $A = \widehat{A}$, i.e. pour les modalités centrales de partitions de Ruspini uniformes. En termes linguistiques, des couples antonymes sont par exemple *chaud / froid* ou *peu / la plupart*.

En outre, $\overline{A}$ et $\widehat{\overline{A}}$ sont dans une relation de subcontraire classique : comme $\top(A, \widehat{A}) \leq 0.5$, on a $\bot(\overline{A}, \widehat{\overline{A}}) \geq 0.5$ donc $\overline{A}$ et $\widehat{\overline{A}}$ vérifie la LTE. D'un point de vue linguistique, des étiquettes de complément et d'antonymes compléments de ce type sont par exemple *non chaud / non froid* et *pas la plupart / pas peu*.

Enfin, l'antonyme complément vérifie une relation d'implication de manière équivalente à celle de subalterne dans le carré classique et de dual dans le carré moderne. Elles sont obtenues de manière identique pour ce dernier, puisque toutes deux sont définies comme la composition des deux autres types de négation. Pour montrer que l'antonyme complément vérifie bien une relation d'implication floue, il suffit de remarquer que $\widehat{A}(x) \leq \overline{A}(x)$ (De Soto & Trillas, 1999) et donc que $A(x) \leq \widehat{\overline{A}}(x)$. Ainsi, en notant $I_g(x, y)$ l'implication floue de Gödel et et $I_\Delta(x, y)$ celle de Goguen, $I_g(A, \widehat{\overline{A}}) = I_\Delta(A, \widehat{\overline{A}}) = 1$. Par exemple, le couple *chaud / non froid* peut être vu comme une implication car si un élément est chaud, alors il n'est pas froid, ou, avec des quantificateurs, si « La plupart des S sont P » alors « pas Peu de S sont P ». Il convient de noter que, dans le cas particulier où l'antonyme et le complément sont égaux, i.e. $A = \widehat{\overline{A}}$, alors A est un auto antonyme complément et la partition ne contient que ces deux éléments et couvre l'univers de discours (cf. graphe `(d)` sur la figure 3.8).

Les négations en logique floue permettent donc de vérifier simplement des caractéri-

sations souples de la LNC et de la LTE. La section suivante présente un modèle général permettant d'inclure ces négations dans le cadre des phrases quantifiées.

3.4 Présentation d'un modèle général d'opposition

Cette section considère le problème général de l'opposition des phrases quantifiées telles qu'elles apparaissent dans les RLF : nous proposons un modèle permettant d'utiliser la richesse des quantificateurs généralisés avec celle des négations floues. Les différents types de protoformes avec leurs négations sont présentés dans un premier temps. Leur représentation sous forme de quadruplet est ensuite détaillée et finalement le 4-cube des oppositions est introduit.

3.4.1 Protoformes de négation

Comme rappelé dans la section 1.2.4 p. 12, un protoforme « QRx sont P » est composé d'un quantificateur Q, d'un qualifieur R et d'un résumeur P. Il faut noter que le problème de l'opposition ne se pose que pour des protoformes liés au même univers de discours, i.e. au même qualifieur R. Ainsi, « La plupart des jeunes sont sont petits » est en opposition avec « La plupart des jeunes sont sont grands » car les deux phrases se rapportent au même univers de discours, les jeunes. En revanche, « La plupart des adultes sont sont petits » n'est pas en opposition avec l'une des deux phrases précédentes car cette phrase porte sur un autre univers de discours, les adultes.

Ainsi, les deux éléments variables d'un protoforme sur lesquels une opposition peut être établie sont le quantificateur Q et le résumeur P. Nous proposons donc dans le cadre de l'étude de leur opposition de noter les protoformes QP sous forme abrégée.

L'absence de négation ainsi que les trois négations floues permettent donc de définir 16 protoformes de négation : QP, $Q\widehat{P}$, $Q\overline{P}$, $Q\widehat{\overline{P}}$, $\widehat{Q}P$, $\widehat{Q}\widehat{P}$, $\widehat{Q}\,\overline{P}$, $\widehat{Q}\,\widehat{\overline{P}}$, $\overline{Q}\,P$, $\overline{Q}\,\widehat{P}$, $\overline{Q}\,\overline{P}$, $\overline{Q}\,\widehat{\overline{P}}$, $\widehat{\overline{Q}}\,P$, $\widehat{\overline{Q}}\,\widehat{P}$, $\widehat{\overline{Q}}\,\overline{P}$, $\widehat{\overline{Q}}\,\widehat{\overline{P}}$.

3.4.2 Représentation des protoformes de négation

Les 16 protoformes de négation peuvent être obtenus par application de l'antonyme sur Q ou P, opérations notées a_1 et a_2 respectivement, ou du complément sur Q ou P, notées c_1 et c_2 respectivement. On note ainsi par exemple $a_1(a_2(c_2(\overline{Q}\,\overline{P}))) = \widehat{\overline{Q}}\,\widehat{P}$. Du fait de leur commutativité, le résultat ne dépend pas de l'ordre d'application des opérations. Comme elles sont également involutives, les 16 protoformes de négation peuvent être obtenus à partir de n'importe quel autre en quatre opérations.

Les compositions spécifiques $a_1 \circ c_1$ et $a_2 \circ c_2$ sont notées ac_1 et ac_2 et représentent l'antonyme complément de Q et P respectivement. La composition $a_1 \circ c_2$ est appelée dualité et fait l'objet de la section 3.5 p. 58.

Nous proposons de représenter les 16 protoformes de négation sous forme de quadruplets (a, b, c, d) dans $\{0, 1\}^4$ où chaque composante représente une opération, égale à 1 si

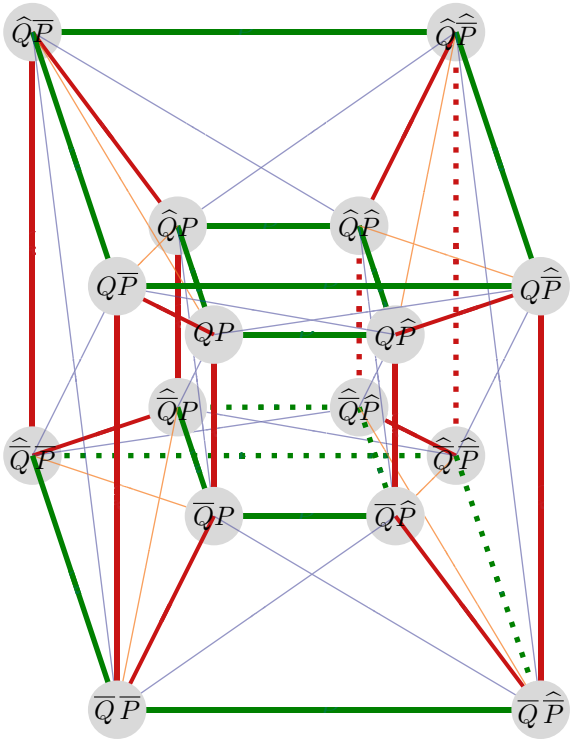


FIGURE 3.9 – Le 4-cube des oppositions

l'opération est appliquée, en partant du protoforme sans négation QP. a indique l'application de a_1, b celle de c_1, c celle de a_2 et d celle de c_2. Par exemple, $(1, 0, 0, 1)$ correspond à $a_1(c_2(QP))$, donc $\widehat{Q}\,\overline{P}$.

L'éq. (1.2) p. 14 permettant l'évaluation du degré de vérité du protoforme « QRx sont P » s'écrit alors :

$$t\,(p) = |b - Q(|a - \rho\,(R, Q) \times \nu(R\,(x)\,, |d - P\,(|c - x|)|)|)| \tag{3.5}$$

x est mentionné explicitement dans les paramètres de la fonction ν afin de représenter l'antonyme de P et donc la composante c. Cette expression permet de calculer de manière similaire le degré de vérité de tous les protoformes d'opposition et donc de réduire considérablement la complexité d'un algorithme de génération exhaustive de toutes les phrases (cf. section 2.5.2 p. 41).

3.4.3 Le 4-cube des oppositions

La représentation des 16 protoformes de négation sous forme de quadruplet indique que leur représentation graphique doit être réalisée dans un espace à quatre dimensions, entraînant une représentation plus complexe que les carrés présentés dans la section 3.1 p. 46. Nous proposons de figurer ces protoformes dans un cube en 4 dimensions, également appelé 4-cube ou tesseract, qui peut être représenté en trois dimensions sous forme de deux cubes imbriqués, comme illustré sur la figure 3.9. Il est commenté ci-dessous.

Des liens de couleurs relient les 16 sommets de cette figure. Ceux en trait épais représentent le complément en rouge et l'antonyme en vert. Les traits verticaux correspondent à c_1, les horizontaux à a_2, ceux en profondeur à a_1 et ceux d'un cube à l'autre à c_2. Enfin, les traits fins représentent les négations composées, l'antonyme complément en bleu et la dualité d présentée plus bas en orange.

Le 4-cube comporte 56 arêtes : 7 opérations (a_1, a_2, c_1, c_2, ac_1, ac_2, d) et 16 sommets donnent $7 \times 16 = 112$ liens directs. Ces derniers sont symétriques car les opérations sont involutives, ce permet de retrouver les $56 = 112/2$ arrêtes.

3.4.4 Relations avec le carré moderne des oppositions

Les liens entre les relations classiques d'opposition et les phrases non quantifiées utilisant des négations floues ont été détaillés dans la section 3.3.5 p. 54. Nous précisons ici celles existant entre le 4-cube et le carré moderne illustré sur la figure 3.3 p. 48.

Dans le carré moderne, seules deux opérations d'opposition sont prises en compte, la négation interne et la négation externe : ainsi, les protoformes d'opposition pour ce carré peuvent se représenter par un couple (α, β), où α et β représente respectivement la négation interne et externe. Par exemple, $(0, 1)$ correspond à $Q(S, \neg P)$. Cette notation en deux dimensions est à comparer à celle en quatre dimensions du 4-cube, ce dernier étant plus complexe du fait de sa prise en charge des négations floues, plus nombreuses que la simple négation classique intégrée dans le carré moderne.

Puisque les 3 sommets d'un carré moderne sont obtenus à partir d'un sommet initial auquel sont appliquées la négation interne, la négation externe et leur composition, tout ensemble de 4 sommets du 4-cube pour lesquels un couple de négations floues est utilisé pour représenter les négations internes et externes peut décrire un carré moderne.

Les 9 couples possibles sont ceux issus du produit cartésien $\{a_1, c_1, ac_1\} \times \{a_2, c_2, ac_2\}$. En partant des 16 sommets du 4-cube et en considérant que les 4 rotations du carré moderne sont équivalentes, le 4-cube contient donc $(16 \times 9)/4 = 36$ carrés modernes.

Par exemple, en utilisant le couple (a_1, a_2) pour les négations interne et externe, le carré moderne QP, $\hat{Q}P$, $Q\hat{P}$, $\hat{Q}\,\hat{P}$ sur la face supérieure du cube interne est retrouvé. Un autre carré, moins visible, est celui déterminé par les sommets QP, $\widehat{Q}P$, $Q\overline{P}$, $\widehat{Q}\,\overline{P}$. Il convient de remarquer que l'ordre d'application des opérations est différent entre le 4-cube et le carré moderne, i.e. les sommets opposés dans le premier sont définis par la composition des opérations tandis qu'ils le sont par la négation externe dans le second. Les deux cependant décrivent bien des relations d'opposition construites par composition de négations élémentaires.

3.5 Propriétés de cohérence des RLF

Nous montrons dans cette section comment le formalisme du 4-cube ainsi que celui introduit dans la section 1.2.5 p. 13 pour le calcul de la valeur de vérité permettent de

satisfaire les propriétés de cohérence énoncées dans la section 2.3.1 p. 33 dans le cadre de la logique floue standard.

3.5.1 Négation de la fonction de comptage

La fonction de comptage est définie dans l'éq. (1.1) p. 13 par $\nu(R,P) = |\top(R,P)|$. Sa négation est basée sur la définition du qualifieur R utilisé. Plus précisément, R est une restriction de l'univers de discours au sein duquel les x vérifiant P sont comptés. Les valeurs de $\nu(R,P)$ s'étendent donc de 0 dans le cas où aucun x ne vérifie P à $|R| = \sum_x R(x)$ si tous les x vérifient P.

L'intervalle de définition de ν est donc $[0,|R|]$ et ne dépend pas de P. Par conséquent, la négation de la fonction de comptage au sens de l'éq. (3.1) p. 53 est :

$$\overline{\nu}(R,P) = n([0,|R|],\nu(R,P)) = |R| - \nu(R,P) \tag{3.6}$$

Par exemple, si 8 éléments parmi 10 vérifient P, alors la négation de ce compte est $10-8=2$ ce qui signifie que 2 éléments ne le vérifient pas.

3.5.2 Propriété de dualité pour une fonction de comptage

Une fonction de comptage vérifiant la propriété de dualité est telle que :

$$\nu(R,\overline{P}) = \overline{\nu}(R,P) = |R| - \nu(R,P) \tag{3.7}$$

La fonction de comptage étant basée sur une t-norme, il convient de déterminer celles vérifiant la propriété de dualité. En pratique, parmi les t-normes classiques rappelées dans la section 1.2.5 p. 13, seule la t-norme probabiliste $\top_P$ la vérifie :

$$\begin{aligned}\nu(R,\overline{P}) &= \sum \top_P(R,1-P) = \sum R(1-P)\\ &= \sum R - \sum \top_P(R,P)\\ &= |R| - \nu(R,P) = \overline{\nu}(R,P)\end{aligned}$$

3.5.3 Exploitation de la propriété de dualité

Nous montrons comment le formalisme présenté dans la section 1.2.5 p. 13 pour le calcul de la valeur de vérité et la propriété de dualité de la fonction de comptage développée ci-dessus permettent de vérifier les propriétés de cohérence décrites dans la section 2.3.1 p. 33, à savoir l'antonymie, la négation externe et la dualité même pour les protoformes du type « QRx sont P ».

Antonymie

La propriété d'antonymie, à ne pas confondre avec l'antonyme comme négation floue décrit dans la section 3.3.3 p. 54, stipule que :

$$t(QP) = t(\widehat{Q}\,\overline{P}) \tag{3.8}$$

Outre le cas $Q = \widehat{Q}$ et $P = \overline{P}$ qui la vérifie trivialement, nous montrons que l'antonymie est également vérifiée pour tout Q et tout P si la fonction de comptage utilisée vérifie la propriété de dualité, donc si la t-norme utilisée est $\top_P$.

En effet, si Q est un quantificateur absolu défini sur $[0, |R|]$, alors $\rho(R, Q) = 1$ (cf. éq. (1.3) p. 14) et :

$$t(QP) = Q(\nu(R,P)) = Q(\overline{\nu}(R,\overline{P})) = \widehat{Q}(\nu(R,\overline{P})) = t(\widehat{Q}\,\overline{P})$$

Si Q est une quantificateur relatif défini sur $[0, 1]$, alors $\rho(R, Q) = 1/|R|$ et :

$$\begin{aligned} t(QP) &= Q(\nu(R,P)/|R|) = Q(1 - \nu(R,\overline{P})/|R|) \\ &= \widehat{Q}(\nu(R,\overline{P})/|R|) = t(\widehat{Q}\,\overline{P}) \end{aligned}$$

Négation externe

La propriété de négation externe est définie par la relation :

$$t(QP) = 1 - t(\overline{Q}\,P) \tag{3.9}$$

Cette propriété est immédiatement vérifiée à l'aide de la définition du complément d'un sef par l'éq. (3.2) p. 53, du calcul de la valeur de vérité par l'éq. (1.2) p. 14 et en notant que $\rho(Q, R) = 1 - \rho(\overline{Q}, R)$.

Dualité d'une phrase

La propriété de dualité est ici définie pour une phrase et ne doit pas être confondue avec celle de la fonction de comptage discutée plus haut. La propriété de dualité pour une phrase implique que (cf. section 1.2.5 p. 13) :

$$t(QP) = 1 - t(\widehat{Q}\overline{P}) \tag{3.10}$$

Cette relation est découle directement des propriétés de négation externe et d'antonymie.

3.6 Conclusion

Nous avons présenté dans ce chapitre une analyse détaillée des différents niveaux d'opposition entre les phrases en fonction de leur complexité, depuis les phrases simples non

quantifiées, celles quantifiées avec des quantificateurs classiques $\forall$ et $\exists$, puis celles quantifiées avec des quantificateurs généralisés et enfin celles utilisant ces mêmes quantificateurs en plus de négations floues. Ce dernier types de phrases est d'un intérêt particulier puisqu'il constitue celles utilisées dans le cadre des RLF standards.

Nous avons par la suite présenté une vision unifiée de ces différentes oppositions au travers d'une structure originale, le 4-cube des oppositions, permettant de représenter les 16 négations de protoformes possibles dans le cadre des RLF.

Enfin, à l'aide des formalisations proposées pour construire le 4-cube, nous montrons que les propriétés de cohérence, i.e. d'antonymie, de négation externe et de dualité sont vérifiables dans le cadre de la logique floue standard même pour les protoformes du type « QRx sont P ». Ce résultat est particulièrement intéressant car il permet de s'affranchir de l'utilisation d'autres approches plus complexes développées pour vérifier ces propriétés. En pratique, l'utilisation de la t-norme probabiliste vérifie la propriété de dualité de la fonction de comptage qui, avec sa négation, satisfont les propriétés de cohérence.

Deuxième partie

Résumés linguistiques de périodicité

Introduction

Nous nous intéressons dans cette seconde partie à des résumés de données d'un type particulier, les séries temporelles, qui associent des valeurs à des dates. Les résumés produits étudient leur caractère périodique, i.e. la répétition approchée de motifs en leur sein. Plus précisément, les protoformes utilisés intègrent la *périodicité*, vue comme un degré mesurant à quel point la série est périodique, et la *période*, mesurant l'intervalle temporel entre deux occurrences du motif périodique.

Nous présentons dans le chapitre 4 un état de l'art détaillé des différentes approches dédiées au calcul de la période d'une série temporelle, présentées en fonction du domaine de représentation des données qu'elles utilisent. Trois points essentiels de l'ensemble de ces méthodes apparaissent à l'issue de ce chapitre. D'abord, à l'exception de la proposition de Lloyd et al. (2014), aucune approche ne prend en charge le rendu linguistique de la période déterminée. De plus, la plupart ignorent la question de la périodicité et ne renvoient au mieux qu'un intervalle de confiance. Enfin, elles nécessitent un certain nombre de paramètres et sont basées sur des hypothèses plus ou moins fortes concernant le modèle de génération des données.

C'est pour répondre à cette triple problématique que nous présentons dans le chapitre 5 une nouvelle méthode, DPE pour *Detection of Periodic Events*, obtenue par la mise en œuvre du principe simple suivant : la série est *périodique si elle alterne de manière régulière des groupes de valeurs hautes et basses, où la régularité est fonction de leurs tailles respectives*. Comme détaillé dans la description des trois étapes de la méthode, DPE permet la génération de phrases comme « Environ toutes les semaines, les valeurs sont élevées », calcule un degré de périodicité et n'est basée sur aucun paramètre ni supposition quant au modèle des données.

Afin d'implémenter le principe présenté ci-dessus et de séparer automatiquement les groupes de valeurs hautes et basses, DPE utilise un nouvel outil, le score d'érosion. Ce dernier peut néanmoins se révéler coûteux car de complexité quadratique dans son implémentation naïve. Nous proposons donc au chapitre 6 différentes alternatives permettant son calcul par niveaux, incrémental et incrémental par niveaux. Ces approches sont supportées par un ensemble de théorèmes montrant que le résultat qu'elles renvoient est équivalent au score d'érosion calculé de manière naïve. De plus, les approches incrémentales permettent également l'analyse de données en flux, comme détaillé dans la seconde moitié du chapitre.

En vue de valider la méthode DPE et ses différentes variantes présentées dans les deux chapitres précédents, nous détaillons au chapitre 7 les expériences menées pour mesurer sa pertinence et sa performance. Concernant la première, un grand nombre de tests sont réalisés sur des données artificielles crées à l'aide d'un générateur spécifique, visant à valider des critères comme la décroissance régulière du degré de périodicité avec le bruit dans les données générées, la robustesse de la méthode pour des séries de bruits équivalents ou encore l'estimation correcte de la période des données. Des données réelles confirmant la pertinence de DPE sont également utilisées. Concernant la performance, des séries de données de types variés et de tailles croissantes sont générées permettant de démontrer la rapidité de l'approche incrémentale par niveaux qui permet le calcul du score d'érosion d'un million de points en 1,5 seconde.

Nous présentons finalement dans le chapitre 8 une extension de la méthode DPE, nommée LDPE pour *Local Detection of Periodic Events*. Cette méthode permet la contextualisation dans le temps des résultats obtenus avec DPE, donc la détection de périodicités locales, ainsi que la génération de phrases comme « Les deux premiers mois, la série est très périodique de période environ 1 semaine ». Des tests réalisés sur des données réelles et artificielles permettent de valider la méthode LDPE.

Chapitre 4

Caractérisation de séries temporelles périodiques : un état de l'art

> Moi, ce que je voudrais bien trouver dans chaque homme, c'est une pulsation, un mouvement régulier et souple qui l'accorde au temps et au monde.
>
> —Jean-Marie Gustave Le Clézio, *L'Extase matérielle*

Nous présentons dans ce chapitre un état de l'art des différentes méthodes permettant de mesurer la *période* et la *périodicité* d'une série temporelle. De manière générale, comme formalisé dans la section 4.1, la période mesure le temps séparant deux occurrences du même événement répété dans la série et la périodicité indique à quel point la série est périodique.

La période est une question centrale dans un grand nombre de domaines parmi lesquels on peut citer par exemple la biologie (Baier, 2005), la musique (De Cheveigné & Kawahara, 2002), l'astronomie (Heck et al., 1985), l'électricité (Chicharo, 1996), la géophysique (Stopa & Cheung, 2014), la physique nucléaire (Ryan et al., 1988), la mécanique (Goldblum et al., 1988), la météorologie (Jones & Brelsford, 1967), les réseaux informatiques (Argon et al., 2013), la zoologie (Li, 2013)...

Dans la première section de ce chapitre, nous définissons les notions de séries temporelles, de période et de périodicité. Nous présentons également les différents domaines dans lesquels les séries peuvent être représentées pour en réaliser l'étude. Les sections suivantes regroupent par domaine les méthodes permettant le calcul de la période des séries étudiées et, le cas échéant, de leur périodicité.

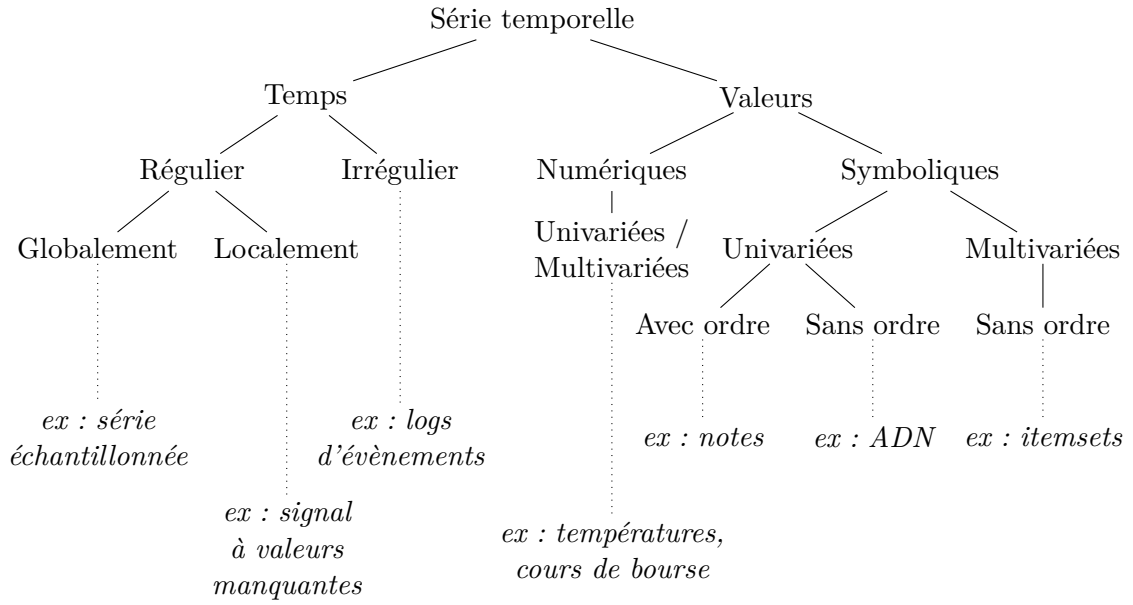


FIGURE 4.1 – Taxonomie des séries temporelles

4.1 Définitions

Les nombreuses méthodes détaillées dans les sections suivantes portent sur différents types de séries temporelles présentés dans la première sous-section. Elles permettent l'analyse de plusieurs formes de la période, détaillées dans la seconde sous-section. Dans la troisième enfin, les domaines de représentation des séries temporelles qui structurent les sections suivantes de ce chapitre sont présentées.

4.1.1 Séries temporelles

Une *série temporelle,* ou *signal*, X est un ensemble ordonné de valeurs x_i associées à des dates t_i. La série temporelle peut être reçue au fur et à mesure et traitée de manière incrémentale, comme détaillé au chapitre 6. Dans la suite de ce chapitre cependant, nous considérons que les n points de la série sont connus a priori. X est alors défini comme :

$$X = (t_i, x_i)_{i=1...n} \tag{4.1}$$

Nous proposons dans les paragraphes suivants une taxonomie des séries temporelles illustrée sur la figure 4.1, structurée selon fonction des domaines auxquels appartiennent les t_i et les x_i.

Dates t_i

Les dates t_i sont ordonnées, i.e. $\forall i = 1, ...n-1$, $t_i < t_{i+1}$. Pour une série temporelle, les t_i peuvent faire référence à une date réelle, t_5 correspondant par exemple à la $5^{ème}$ valeur de la série, mesurée à 15h42. Pour une série non temporelle, les t_i sont des indices ordonnés

non liés à une échelle de temps, comme dans une séquence d'ADN par exemple. Le terme de *série* peut être employé seul pour des raisons de simplicité, le contexte permettant alors de déterminer son caractère temporel ou non.

Si l'écart Δ entre deux dates successives d'une série temporelle est constant, la série est à temps *régulier*. La série est à temps *globalement régulier* si Δ est constant pour toutes les dates de la série et à temps *localement régulier* s'il ne l'est que pour certaines d'entre elles. Sur les parties où Δ est constant, chaque date peut s'écrire $t_i = i\Delta t + t_0$.

Dans le cas où la série est obtenue par la mesure régulière d'un phénomène temporel, $1/\Delta$ représente sa fréquence d'échantillonnage. La détermination de cette fréquence n'est pas triviale et la plupart du temps définie comme la fréquence de Nyquist égale à la moitié de la plus grande fréquence du phénomène étudié.

Dans certains cas, la mesure ne peut être effectuée sur toute la durée de l'enregistrement (cas de la luminosité des étoiles par exemple (Huijse Heise et al., 2012)) et la série est alors à temps localement régulier ou à temps globalement régulier avec des valeurs manquantes (Fahlman & Ulrych, 1982).

Enfin, les séries à *temps irrégulier* associent aux valeurs des dates appartenant à un ensemble continu, habituellement $\mathbb{R}^+$, et qui ne peuvent pas s'exprimer en fonction d'une fréquence d'échantillonnage, même localement. L'enregistrement des heures d'entrée dans un bâtiment en sont un exemple. Ces séries ne sont pas réductibles simplement à des séries à temps régulier et les méthodes dédiées à l'analyse de leur période les prennent en compte spécifiquement.

Valeurs x_i

Les spécificités des différentes séries temporelles, univariées ou multivariées, à valeurs numériques ou symboliques, sont décrites ci-dessous.

Les séries qui associent une seule valeur chaque instant sont *univariées*, celles qui leur associent plusieurs valeurs sont *multivariées*. Ces dernières peuvent également être vues comme la superposition de plusieurs séries univariées.

Les séries *numériques* sont à valeurs dans des ensembles continus, souvent $\mathbb{R}$, $\mathbb{C}$, ou un de leurs sous-ensembles. Une série numérique univariée peut par exemple contenir des relevés de températures tandis qu'une série multivariée pourrait en plus contenir des valeurs de pression et d'humidité à chaque date.

Les séries *symboliques* sont à valeurs dans des ensembles discrets, le plus souvent $\mathbb{N}$ ou un alphabet fini de caractères. Elles peuvent être univariées comme dans le cas d'une séquence d'ADN ou multivariées comme dans le cas de paniers clients (*itemsets*).

Le domaine de définition des valeurs d'une série numérique peut être ordonné ou non. Il ne l'est pas pour les 4 caractères d'une séquence d'ADN mais il l'est pour une série dont les caractères sont associés à des intervalles de valeurs ordonnés, e. g. $[0; 1[\rightarrow a$ et$[1; 2] \rightarrow b$.

Modèle de série

L'étude des séries temporelles peut supposer que les données sont de nature *déterministe* ou *stochastique*. La supposition déterministe est retenue dans deux cas : soit les équations du phénomène mesuré sont connues et les erreurs de mesure et de bruit sont ignorées (Shin & Hammond, 2008, p.8), soit à l'inverse rien n'est connu du phénomène et l'on ne souhaite introduire aucune information a priori (Mallat, 1999, p.724).

Entre ces deux extrêmes, des modèles intermédiaires ont également été proposés, reposant sur différentes approches de modélisation de l'incertitude comme la théorie des probabilités (Stoica & Moses, 2005, p.2) ou du flou (Mendel, 2000). En ce cas, le modèle couramment considéré comme la somme d'un signal déterministe et d'un bruit.

Lorsque le signal est considéré comme *stochastique*, la série est vue comme la réalisation d'un processus stochastique, soit d'un ensemble de variables aléatoires à valeurs dans un même univers et indexées par une variable temporelle. Un processus stochastique dont les caractéristiques d'espérance et de covariance sont constantes est dit *stationnaire*.

4.1.2 Définition des séries périodiques et de leurs variantes

Cette sous-section introduit les concepts de période et de périodicité avec différents cas de pseudo-périodicité.

Période et périodicité La *période* p d'une série temporelle mesure le temps entre deux occurrences du même événement. Elle est telle que :

$$\forall i,\ x_i = x_{i+p} \tag{4.2}$$

La fréquence $f = 1/p$ est son inverse et mesure le nombre d'occurrences d'un événement par unité de temps.

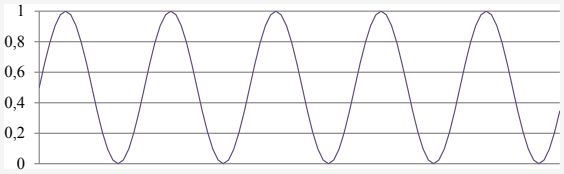
Dans le cadre de série temporelles finies, la période p est définie dans $[1, n/2]$. Le cas $p = 1$ correspond à une série constante et le cas $p = n/2$ à la période maximale car au moins deux occurrences de chaque valeur doivent être présentes dans la série pour vérifier l'éq. (4.2).

La *périodicité* désigne le caractère périodique d'une série temporelle. Si une série vérifie l'éq. (4.2) pour $i = 1...n - p$ comme celle illustrée sur la figure 4.2 série `(a)`, alors sa périodicité vaut 1.

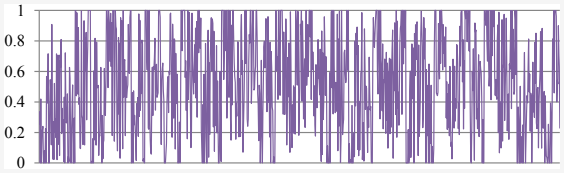
Dans le cas où l'éq. (4.2) n'est pas vérifiée, une valeur faible ou nulle de périodicité peut être retenue, comme dans le cas de la série `(b)` par exemple. En ce cas, la période de la série n'est pas définie.

Pseudo-périodicité Il semble néanmoins limitatif de n'associer que les valeurs 0 ou 1 à la périodicité car de nombreux cas de comportements *pseudo-périodiques* peuvent être rencontrés pour lesquels une valeur de périodicité *entre* 0 et 1 est plus adaptée.

Période et périodicité

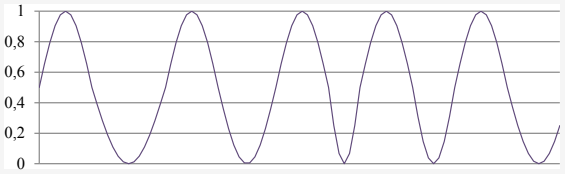


(a) Périodicité stricte - sinusoïde

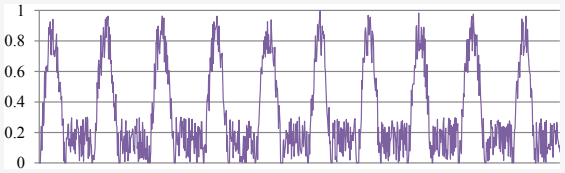


(b) Périodicité nulle

Pseudo-périodicité

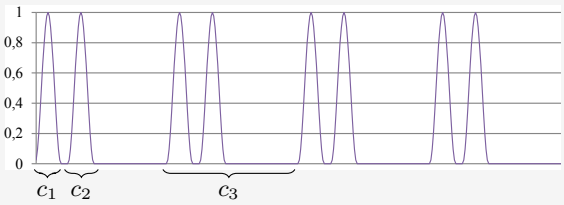


(c) Pseudo-périodicité en temps

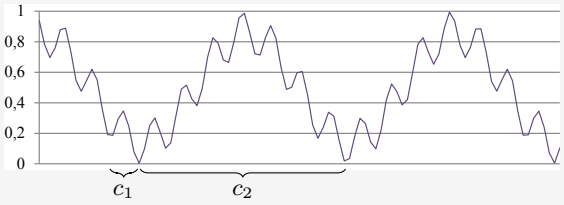


(d) Pseudo-périodicité en valeurs

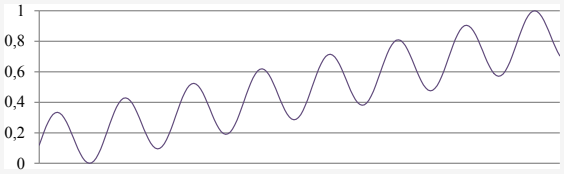
Composantes multiples



(e) Produit de composantes

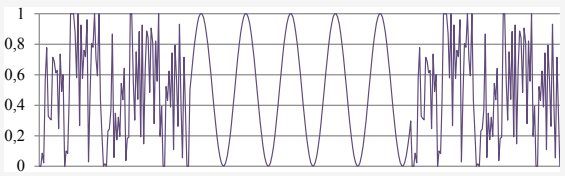


(f) Somme de composantes



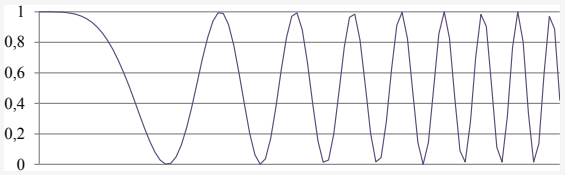
(g) Composante tendancielle

Période non constante



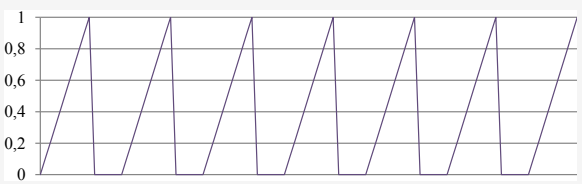
(h) Périodicité locale - 1 période

(i) Périodicité locale - 2 périodes



(j) Période variable

Motifs périodiques



(k) Périodicité stricte - triangles

d		d	d	d		d	
	c				c	c	c
b		b		b	b	b	
a		a	a	a		a	
t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8

(l) Série multivariée symbolique

FIGURE 4.2 – Illustrations des différents cas de périodes

Les séries `(c)` et `(d)` par exemple illustrent deux types de pseudo-périodicité, en temps et en valeurs respectivement. La première correspond à des valeurs égales approximativement tous les p points, tandis que la seconde correspond à des valeurs approximativement égales tous les p points.

Composantes multiples D'autres cas de pseudo-périodicité peuvent venir de la présence de plusieurs composantes dans le signal. La série `(e)` par exemple résulte du produit entre deux composantes périodiques, l'une de période c_1 ou c_2 faite de pics fins et l'autre rectangulaire de période c_3 égale à 1 sur la première moitié et zéro sur la seconde. La série `(f)` quant à elle résulte de la somme de composantes périodiques sinusoïdales, la première de période c_1, d'amplitude moindre que la seconde de période c_2.

Enfin, la série `(g)` est également une somme de composantes, l'une périodique et l'autre linéaire. La périodicité d'une telle série est discutable et l'étude de la période pour ce type de séries débute généralement par le retrait de la composante linéaire.

Période non constante La période et la périodicité peuvent également évoluer dans le temps. Dans la série `(h)` par exemple, la périodicité est élevée au milieu du signal et faible ou nulle aux extrémités. Dans la série `(i)` la périodicité est globalement élevée mais les périodes sont différentes. Enfin, la période de la série `(j)` augmente de manière linéaire avec le temps. Ce type de série, appelé *chirp*, est utilisé pour démontrer l'intérêt des méthodes temps-fréquence décrites dans la section 4.4 p. 83. Nous ne définissons en revanche pas de périodicité en ce cas.

Motifs périodiques En plus des valeurs de périodes et de périodicité présentées jusqu'ici, la recherche du motif répété est également intéressante. Les séries `(a)` et `(k)` ont par exemple des valeurs identiques de période et de périodicité mais deux motifs différents, sinusoïdaux et triangulaires respectivement.

L'identification de motifs est par ailleurs centrale pour l'analyse de séries symboliques, comme détaillé dans la section 4.5 p. 87. En ce cas, la suite de caractères constitutifs du motif est renvoyée. Là aussi les répétitions peuvent être approximatives en valeurs, i.e. seule une partie du motif est répétée, ou en temps, auquel cas l'écart entre deux motifs successifs n'est pas tout à fait constant. Cette analyse peut également être menée dans un contexte multivarié comme par exemple la série `(l)` qui exhibe un comportement périodique pour les symboles a et b qui apparaissent tous les deux pas de temps et c ou d présents à chaque pas de temps.

4.1.3 Principes de représentations des séries temporelles

Dans la suite d'autres états de l'art sur le sujet (Gerhard, 2003; Costa et al., 2013), les méthodes étudiées ici sont classées selon le domaine de représentation qu'elles font des données pour en extraire la période, la périodicité et d'autres informations le cas échéant.

Ces domaines peuvent être temporel, fréquentiel, temporo-fréquentiel, symbolique ou autre. Dans les trois premiers, les séries étudiées sont numériques, la plupart du temps issues de l'enregistrement de phénomènes physiques. Dans le domaine symbolique, les séries sont à valeurs dans des ensembles finis et discrets de caractères, comme par exemple une séquence d'ADN ou un log d'événements. Enfin, dans le domaine « Autre » sont rassemblées des méthodes plus marginales, présentées pour des raisons d'exhaustivité.

Les sections suivantes détaillent les méthodes développées pour calculer la période et la périodicité des différents types de séries présentés ci-dessus dans chacun de ces domaines. La plupart des méthodes ignorent toutefois le calcul de la périodicité et la majorité d'entre elles se concentrent sur celui de la période.

A l'exception du domaine temporel où la représentation est directe, les sections sont réparties en deux sous-sections, la première détaillant l'étape de représentation des données et la seconde présentant les traitements effectués sur ces représentations permettant le calcul de la période et des différents aspects de la périodicité pris en charge.

4.2 Représentations temporelles

La représentation d'une série de données dans le domaine temporel est celle donnée par l'éq. (4.1) p. 68. C'est la représentation la plus simple et la plus intuitive d'une série temporelle : elle met en relation une date et une valeur.

Différentes approches présentées dans les paragraphes suivants ont été développées pour permettre le calcul de la période et de la périodicité dans ce cadre. Les premières sont basées sur le calcul de statistiques simples dans le domaine temporel, comme le nombre de croisements avec l'axe des abscisses ou la mesure de corrélation des données, les secondes sur des techniques de segmentation du signal et les dernières sur des approches par régression.

4.2.1 Croisement avec l'axe des abscisses ou *zero-crossing*

Le principe de la méthode de *zero-crossing* est de dénombrer le nombre de croisements de la série avec l'axe des abscisses, $c = |\{j \in 1...n-1 | (x_j > 0 \wedge x_{j+1} < 0)\}|$ (Kedem, 1986). Dans le cas d'une série stationnaire à moyenne nulle additionnée d'un bruit gaussien, $c\pi/(n-1)$ tend vers la fréquence dominante du signal.

La méthode est cependant sensible au bruit dans les données (Tsuji & Yamada, 2001). Elle est utilisée dans le cas où les données sont peu bruitées, par exemple dans le domaine de la production électrique (Backmutsky et al., 2000; Ratanamahatana et al., 2005) ou pour confirmer une fréquence calculée avec une autre méthode (Kedem, 1986).

4.2.2 Mesures de corrélation

La mesure de corrélation d'une série est l'évaluation du lien γ_k entre les points x_i et x_{i+k}, où k désigne le décalage ou *lag*. k est une période candidate si γ_k est un extremum

pour $k = 0, ..., n-1$. Deux méthodes de calcul de γ_k sont détaillées dans les paragraphes suivants.

Autocorrélation En supposant une série de moyenne nulle et de variance unitaire, l'autocorrélation évalue le lien entre x_i et x_{i+k} par le produit $x_i x_{i+k}$ divisé par les n points de la série. C'est une mesure statistique définie pour des séries infinies et son calcul pour des séries finies est réalisé par deux estimateurs respectivement, biaisés et non biaisés, définis par $\gamma_k^b = \sum_{i=1}^{n}(x_i x_{i+k})/n$ et $\gamma_k^{nb} = \sum_{i=1}^{n}(x_i x_{i+k})/(n-k)$. Le biais du premier est compensé par sa faible variance (Stoica & Moses, 2005, p.23). Le choix de l'estimateur dépend de l'utilisation qui en est faite. L'estimateur biaisé notamment est utilisé dans le calcul des représentations fréquentielles, comme détaillé dans la section 4.3.1 p. 79.

Les valeurs d'autocorrélation pour $k = 0...n-1$ constituent la séquence d'autocorrélation (SA), ou corrélogramme, et ses maxima déterminent les périodes candidates de la série. Pour une série périodique de période p, la SA est maximale aux indices kp avec $k \in \mathbb{N}$ et $kp < n$.

Dans le cas des séries pseudo-périodiques, les pics de la SA peuvent être liés au bruit ou à une période. Afin de les distinguer, un seuil peut être défini, soit par l'utilisateur (De Cheveigné & Kawahara, 2002), soit par l'expression $2/\sqrt{n}$ qui permet de les distinguer à 95% de confiance. Cette dernière n'est robuste que pour n grand (Chatfield, 1996, p. 21).

L'autocorrélation est couramment utilisée pour le calcul de la période et donne de bons résultats avec des signaux sinusoïdaux et stationnaires principalement (Gerhard, 2003). Ses résultats sont moins bons avec d'autres signaux et la méthode est notamment susceptible « d'erreurs d'octaves », i.e. de sélection d'une fréquence autre que la fréquence de base (De Cheveigné & Kawahara, 2002).

Aussi, différentes améliorations ont été proposées afin de rendre la méthode plus robuste, en particulier dans le domaine de l'analyse de la voix : De Cheveigné & Kawahara (2002) proposent la méthode YIN qui intègre un certain nombre d'opérations postérieures au calcul de l'autocorrélation, Talkin (1995) introduit l'autocorrélation normalisée et la méthode RAPT, améliorée plus tard par Zahorian & Hu (2008) ou Rashidul Hasan & Shimamura (2012).

Fonction de fluctuation moyenne, AMDF La fonction de fluctuation moyenne (Rosenblum & Kurths, 1995) étudiée par ailleurs sous le nom d'*Average Magnitude Difference Function* (Ross et al., 1974), met en œuvre une approche fenêtrée dépendant d'un paramètre w qui définit la taille de la fenêtre : la corrélation est ensuite calculée comme la moyenne des différences sur la fenêtre, plus précisément $\gamma_k = \left(\sum_{i=1...w} |x_i - x_{i+k}|\right)/w$.

A l'inverse de l'autocorrélation, les données sont d'autant plus corrélées que γ_k est faible et les indicateurs dans ce contexte sont liés à la recherche des minima de cette séquence. Aucun seuil statistique n'est en revanche donné pour cette approche.

4.2.3 Segmentation

Les méthodes par segmentation consistent à découper la série en blocs de p points consécutifs où p est la période candidate. Formellement, la série est décomposée segments s_j pour $j = 1...\lfloor n/p \rfloor$ dont le $i^{ème}$ point $s_j[i]$ est défini comme $s_j[i] = x_i$ avec $(j-1)p < i \leq jp$. Si p ne divise pas n, les points restants sont ignorés.

Ces méthodes sont dédiées à l'analyse de séries discrètes à temps régulier. Elles détectent une composante périodique dans le signal et ne sont pas adaptées aux périodicités locales, variables, tendancielles : elles déterminent une unique composante périodique. Elles renvoient en revanche un motif périodique, qui peut être le segment moyen dans le cas des méthodes par actogramme ou le segment le plus représentatif, comme détaillé ci-dessous.

Actogramme La représentation par actogramme ou tableau de Buys-Ballot est couramment utilisée en biologie dans la recherche de rythmes circadiens (Enright, 1965; Refinetti et al., 2007). Elle consiste à étudier les propriétés d'un segment S de p points définis par $S_i = \sum_{j=1}^{\lfloor n/p \rfloor} s_j[i]$ pour $i = 1...p$.

Si la distribution de S est à peu près uniforme, alors p n'est pas la période de la série puisqu'en ce cas ses valeurs sont régulièrement réparties dans les segments. Si au contraire la distribution de S contient un ou plusieurs pics, alors p peut être la période de la série car en ce cas la majorité des valeurs similaires apparaît tous les p points. De plus, les différents pics correspondent aux phases des différentes composantes périodiques du signal (cf. série `(e)` de la figure 4.2 p. 71).

Différents tests statistiques sont proposés afin d'exploiter S, directement ou au travers de formes dérivées, comme la moyenne ou le carré de ses valeurs selon le test considéré : analyse de la variance (Schwarzenberg-Czerny, 1989), efficace même pour les séries contenant peu de points, test du χ^2 avec des variantes pour les signaux non sinusoïdaux (Larsson, 1996), test de Rayleigh (Brazier, 1994), statistiques Z_m^2 (Buccheri, 1988) et Q_P (Enright, 1965), également mentionnées par Refinetti et al. (2007) et Zielinski et al. (2014). Zucker (2015) propose une comparaison de ces méthodes dans le cas de séries discrètes localement régulières.

D'une manière générale, l'intérêt de ces méthodes est leur simplicité, mais elles requièrent une connaissance a priori de la période recherchée pour ne pas avoir à tester toutes les valeurs possibles.

Segments représentatifs A l'inverse des méthodes par actogramme, celles présentées dans ce paragraphe n'agrègent pas les segments s_j mais cherchent à identifier le plus significatif d'entre eux, i.e. permettant la reconstruction de la série la plus similaire à la série d'origine. La reconstruction est ici entendue comme la génération d'une série de taille n par répétition du segment candidat.

La recherche naïve du meilleur motif est computationnellement trop complexe puisqu'elle implique, pour chaque période p, le calcul de tous les segments de longueur p qu'il

est possible d'extraire de la série d'origine et pour chacun d'entre eux la distance entre la reconstruction et la série.

Indyk et al. (2000) proposent une méthode probabiliste permettant de déterminer le segment le plus représentatif en construisant des segments candidats de longueur croissante en partant du premier point de la série. Si l'algorithme fonctionne en $O(n \log n)$, il ne peut toutefois pas identifier de segment dont le premier point ne coïncide pas avec celui de la série. Ces seuls segments sont également pris en compte par la méthode APT d'Otunba & Lin (2014) qui font référence à d'autres approches de comparaison rapide de segments via des techniques de hachage ou de structures de données spécifiques (Indyk & Motwani, 1998). D'autres solutions détaillées dans la section 4.5 p. 87 utilisent une représentation symbolique pour accélérer le traitement.

4.2.4 Régression

Le principe des méthodes par régression est de déterminer à partir des données les paramètres d'un modèle périodique défini analytiquement et fixé a priori. En fonction du modèle utilisé, différents types de périodicité peuvent être détectés.

Trois formules de régression sont présentées ici par complexité croissante : locales, au sens des moindres carrés et par processus gaussiens.

Régression simple Les méthodes de régression les plus simples consistent en l'extraction des paramètres d'un modèle déterministe, donc sans bruit, à partir de points consécutifs de la série. Comme indiqué dans les références ci-dessous, ces modèles sont utilisés en particulier dans le cadre du suivi de production électrique où les données sont régulières.

Mahmood et al. (1985) estiment à l'aide de trois points consécutifs les paramètres d'un modèle de la forme $x_t = V \sin(wt)$ où V est l'amplitude de la sinusoïde et ω sa fréquence radiale, i.e. $\omega = 2\pi f$. Moore et al. (1994) développent le modèle en y intégrant la phase et Zayezdny et al. (1992) utilisent de plus les dérivées premières et secondes du modèle sur 3 ou 7 points consécutifs.

Ces méthodes sont simples et donnent un aperçu de la période locale du signal, comme illustré sur les séries `(h)` et `(i)` de la figure 4.2 p. 71. Elles ne permettent pas en revanche de déterminer d'autres types de période et ne renvoient pas de périodicité.

Régression au sens des moindres carrés Les méthodes de régression au sens des moindres carrés permettent de déterminer les paramètres d'un modèle minimisant l'erreur quadratique entre le modèle et les données réelles. Par rapport aux méthodes de régression simple, ces deniers permettent de prendre en compte le bruit sur les données. Formellement, en notant θ les paramètres d'un modèle $f(x_i; \theta)$, les meilleurs paramètres au sens des moindres carrés sont définis comme $\arg\min_\theta \sum_{i=1}^{n} (x_i - f(x_i; \theta))^2$.

La méthode COSOPT (Straume, 2004) détermine de la sorte les paramètres d'un modèle trigonométrique additionné d'un bruit gaussien et d'une composante linéaire. ARSER (Yang & Su, 2010) ajoute un degré de complexité en prenant en compte la superpo-

sition de plusieurs composantes périodiques. Des tests statistiques sont proposés pour ces deux méthodes.

Ces méthodes permettent d'exploiter des données à temps discret localement régulier et d'analyser des séries multi-composantes pour ARSER. Elles nécessitent cependant la connaissance a priori d'un modèle. Il faut noter que lorsqu'un modèle construit sur une somme de sinusoïdes est utilisé, ces méthodes sont équivalentes à la transformée de Fourier présentée dans la section suivante, puisque cette dernière est précisément la solution du problème de minimisation exposé ci-dessus (Scargle, 1982).

Régression par processus gaussiens Le principe de ces méthodes est d'effectuer une régression des données à l'aide d'un processus gaussien, i.e. telle que la distribution jointe d'un sous-ensemble consécutifs des variables aléatoires qui le composent est gaussienne. L'intérêt de cette modélisation est sa relative simplicité puisqu'un tel processus est entièrement défini par sa moyenne et sa matrice de covariance, cette dernière pouvant être vue comme un noyau (Rasmussen & Williams, 2006, p. 13).

L'utilisation d'un noyau périodique (Rasmussen & Williams, 2006, p. 92) permet la détection de périodicité dans les données. Preotiuc-Pietro & Cohn (2013) en proposent une utilisation pour évaluer la périodicité de hashtags Twitter.

Différents noyaux peuvent être combinés pour déterminer des modèles plus complexes (Duvenaud et al., 2013). Par exemple, Durrande et al. (2013) utilisent un noyau défini comme la somme d'un noyau périodique et d'un noyau apériodique, leur permettant de définir la périodicité comme la part de la variance du signal portée par la régression avec le noyau périodique rapportée à celle donnée par le noyau complet. L'article présente également une comparaison avec les méthodes COSOPT et ARSER présentées plus haut.

Sur le même principe, Duvenaud et al. (2013) utilisent d'autres noyaux, chacun décrivant une caractéristique particulière du signal, comme sa linéarité, sa périodicité ou l'évolution de ses variations, et recherchent une combinaison de noyaux définissant le processus gaussien optimal au regard des données, i.e. maximisant sa vraisemblance et minimisant sa complexité calculée par le critère d'information de Bayes (BIC).

Lloyd et al. (2014), dans le projet Automated Statistician, développent cette méthode et y ajoutent un module linguistique générant une description des données en fonction des noyaux utilisés et des paramètres déterminés. Cette approche de rendu textuel est également utilisée par les méthodes DPE et LDPE présentées dans les chapitres suivants de cette thèse.

Les méthodes de régression par processus gaussiens ont l'avantage de permettre l'analyse d'un grand nombre de cas de périodicité, comme la périodicité tendancielle, approximative ou locale. De plus, les séries temporelles traitées peuvent être à temps irrégulier et multivariées. Cependant, elles optimisent un nombre important de paramètres sur un espace de fonctions important et ont une complexité en $O(n^3)$ (Barber, 2012, p.396).

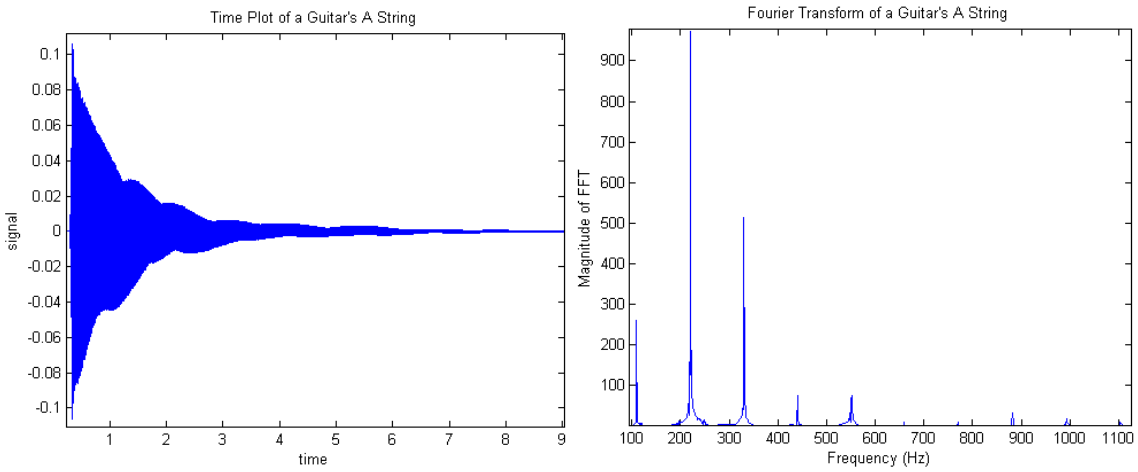


FIGURE 4.3 – Signal correspondant à la note La d'une guitare : représentations temporelle (*gauche*) et fréquentielle (*droite*) (Nelson Lee - `http://bit.ly/1MxXp7c`)

4.3 Représentations fréquentielles

La représentation fréquentielle d'une série est construite à partir de la répartition de sa puissance sur un ensemble de fréquences. La puissance mesure l'énergie du signal par unité de temps, l'énergie étant définie comme la somme des carrés des valeurs de la série (Stoica & Moses, 2005, chap. 1).

Les séries considérées dans ce cadre sont stochastiques, i.e. issues de la réalisation d'un processus stochastique, et stationnaires, i.e. leur moyenne et leur séquence d'autocorrélation sont constantes. Des tests de la stationnarité d'un signal sont proposés par Nason (2013).

La représentation d'une série de données dans le domaine fréquentiel est appelée Densité Spectrale de Puissance, ou DSP, et met en rapport la puissance du signal pour chaque fréquence considérée. Les fréquences de puissance élevées sont les fréquences candidates du signal.

La figure 4.3 illustre l'intérêt de la DSP par la comparaison des représentations temporelle et fréquentielle d'un signal audio : aucune période n'est directement visible sur la première tandis que la seconde montre clairement cinq fréquences de puissances importantes qui sont les fréquences du signal.

Le calcul de la DSP, aussi appelé estimation spectrale, ainsi que son exploitation pour la détermination des différents cas de pseudo-périodicité présentés dans la section 4.1.2 p. 70 sont respectivement décrits dans les deux sous-sections suivantes.

4.3.1 Représentation par estimation spectrale

Le calcul de la DSP a suscité un très grand nombre de recherches : Stoica (1993) en donne une liste de 312 références.

Parmi les nombreux travaux publiés à ce sujet, la présente sous-section cite régulièrement les ouvrages de Stoica & Moses (2005) (« SM »), Percival & Walden (1998) (« PW »),

Shin & Hammond (2008) (« SH ») et Mallat (1999) (« Ma ») ainsi que l'article de Kay & Marple (1981) (« KM »).

Cette sous-section présente les deux familles principales d'estimation de la DSP, non-paramétriques et paramétriques (PW, p. 18; SM, p. 2).

Approches non paramétriques

Les approches non paramétriques permettent de calculer la DSP du signal sans modèle présupposé pour les données.

Transformée de Fourier et approches équivalentes Le calcul de la DSP peut être réalisé par calcul de la transformée de Fourier (TF) ou via d'autres méthodes équivalentes comme celles basées sur les banques de filtres ou la régression sur une somme de sinusoïdes.

Principe Le calcul de la DSP basé sur la TF est le plus classique. Pour les séries discrètes et finies utilisées dans le cadre du traitement du signal informatique, la $k^{ème}$ composante de la TF discrète (TFD) est un nombre complexe donnée par (SH, p. 153) :

$$X(k) = \sum_{j=0}^{n-1} x_j e^{-2i\pi jk/(n\Delta t)} \tag{4.3}$$

$X(k)$ est une fonction sinusoïdale complexe dont la fréquence dite « de Fourier » est donnée par l'expression $k/(n\Delta t)$ et la puissance par son module au carré. La TFD peut être calculée rapidement grâce à l'algorithme de *Fast Fourier Transform* (Cooley & Tukey, 1965). La puissance d'une fréquence peut également être obtenue par TFD de la SA du signal (SH, p. 334).

Une décomposition équivalente peut être obtenue par l'usage de banques de filtres. En ce cas, le signal est passé successivement au travers de filtres passe-bande dont la bande passante est centrée autour de la fréquence $k/n\Delta t$. La puissance du signal en sortie est celle de la DSP pour la fréquence correspondante. Les avantages de cette approche sont détaillé dans (PW, p. 332).

Enfin, la puissance des fréquences de la TFD peut également être déterminée par le calcul de la régression au sens des moindres carrés sur un modèle constitué d'une somme de sinusoïdes complexes (Lomb, 1976; Scargle, 1982). L'intérêt de cette approche est qu'elle permet le calcul de la DSP dans le cas de données à temps irrégulier.

Problèmes d'estimation Pour l'ensemble de ces méthodes, le calcul d'un ensemble *fini* de coefficients sur des données *finies* et *discrètes* entraîne d'une part des fuites spectrales, i.e. l'ajout de puissances inexistantes dans le voisinage d'une fréquence donnée, une discontinuité dans l'évaluation des puissance correspondant uniquement aux fréquences de Fourier ainsi qu'une variance importante pour ces estimateurs (PW, p. 90; SM, p. 30).

Différentes fenêtres sont proposées afin de réduire leur variance au détriment de leur résolution : celles de Barlett et Welch sont calculées sur la représentation temporelle du

signal, celles de Blackman-Tukey, Barlett, Parzen sur sa séquence d'autocorrélation et celle de Daniell sur sa représentation fréquentielle. En pratique, ces fenêtres sont réductibles à celle de Blackman-Tukey (SM, p. 55).

Enfin, l'augmentation artificielle de la taille de la série par l'ajout de valeurs nulles à la fin du signal (*zero padding*) permet d'augmenter le nombre de points auxquels la puissance est évaluée puisque ces dernières sont de la forme $k/n\Delta t$. Il ne modifie pas la résolution fréquentielle mais permet le calcul de la DSP sur une échelle plus fine (Oppenheim et al., 1999, p. 712; PW, p. 114; SM, p. 27).

Autres méthodes Si les approches présentées ci-dessus permettent le calcul d'une DSP théoriquement exacte, les problèmes liés à son estimation en pratique ont entraîné le développement de méthodes permettant l'évaluation de DSP approchées, i.e. dont la convergence théorique vers la DSP n'est pas garantie mais dont les résultats sont dans certains cas mieux adaptés. Les valeurs associées aux fréquences pour ces dernières ne sont pas des puissances *stricto sensu* mais des évaluations de l'importance de ladite fréquence dans le signal.

Parmi ces nombreuses approches, on compte la méthode de Prony qui effectue une régression sur un modèle de sinusoïdes amorties (Kay & Marple, 1981, p. 1404), le calcul de la correntropie qui redéfinit une mesure de corrélation pour le calcul de la DSP (Santamaria et al., 2006; Huijse Heise et al., 2012), le périodogramme χ^2 (Enright, 1965) dont le principe est similaire à celui des approches par segmentation présentées dans la section 4.2.3 p. 75, l'analyse cepstrale, orientée vers l'analyse de signaux audio (Gerhard, 2003; Oppenheim & Schafer, 2004, p. 11), la définition de bases spécifiques non sinusoïdales comme les sommes de Ramanujan (Vaidyanathan, 2014), les suites de Farey (Vaidyanathan & Pal, 2014) ou celles dédiées aux transformées périodiques (Sethares & Staley, 1999).

Approches paramétriques

Les approches paramétriques fonctionnent par estimation des paramètres d'un modèle fixé a priori (SM, chap. 3 et 4). Les modèles utilisés, AR, MA et ARMA sont suffisamment génériques pour être applicables à un grand nombre de cas concrets (KM, p. 1387). L'intérêt de ces méthodes est de produire une estimation continue, i.e. non restreinte aux fréquences de Fourier, et statistiquement cohérente, i.e. dont la variance de l'estimateur est asymptotiquement nulle (SM, p. 106). Cette amélioration de la précision est la contrepartie de l'information sur le modèle qui doit être fourni a priori.

Ces modèles sont construits en deux temps : leur complexité, ou ordre, est d'abord calculée, puis leurs paramètres sont déterminés.

Ordre du modèle Différentes approches du calcul de l'ordre du modèle sont décrits dans (KM, p. 1395; SM, p. 377; PW, p. 434). Elles ont pour objectif de déterminer le meilleur compromis entre une précision maximale et une complexité minimale.

Deux types de modèles sont utilisés en fonction du type de DSP ou *spectre* du signal. Un spectre *continu* correspond à un signal général tandis qu'un spectre *discret* correspond à un modèle de signal composé d'une somme de sinusoïdes et d'une composante de bruit.

Spectres continus Pour les spectres continus, les modèles AR, MA et ARMA peuvent être utilisés. En pratique, les modèles AR sont les plus utiles car correspondant à plus de cas réels que les modèles MA et moins complexes à calculer que les modèles ARMA. Les équations de Yule-Walker et la méthode de Burg permettent notamment le calcul des paramètres d'un modèle AR (KM, p. 1387; SM, chap. 3; PW, chap. 9).

Spectres discrets Les méthodes d'évaluation du spectre discret peuvent être basées sur des techniques de régression non linéaire, précises mais très sensibles aux valeurs d'initialisation, sur la matrice de covariance du signal avec les approches MUSIC, de Pisarenko, Min-Norm ou ESPRIT ou enfin sur des approches bayésiennes détaillés par Bretthorst (1997). Stoica & Moses (2005, p. 167) rapportent que la méthode ESPRIT est la plupart du temps celle présentant le meilleur compromis précision / complexité.

4.3.2 Exploitation des représentations fréquentielles

Les méthodes permettant le calcul de la période des différentes composantes du signal à l'aide de la DSP sont présentées dans le premier paragraphe ci-dessous.

La périodicité, définie comme un degré dans [0,1] déterminant à quel point une série est périodique, n'est pas traitée par ces méthodes. Des tests statistiques présentés dans le second paragraphe sont en revanche développés pour permettre la détection de périodes significatives dans les séries pseudo-périodiques. La *p-value* utilisée dans ce contexte pourrait par exemple être utilisée pour construire une mesure du périodicité mais cette dernière n'est pas développée dans les méthodes présentées ici.

Estimation des composantes périodiques à partir de la DSP

D'une manière générale, les composantes périodiques de la série sont les valeurs maximales de la DSP. La détermination de ces valeurs est réalisée différemment selon que la DSP est estimée de manière non-paramétrique ou paramétrique.

Pour les approches non-paramétriques, la puissance est estimée aux fréquences de la forme $k/n\Delta t$. Certaines méthodes de pondération de la puissance entre fréquences adjacentes ont été proposées pour évaluer la puissance des fréquences autres (Quinn, 1994; Kootsookos, 1999; Bernd et al., 2009).

Les méthodes paramétriques quant à elles permettent un calcul haute résolution des fréquences constitutives du signal car le modèle de la DSP est supposé connu. Une fois ses paramètres estimés, les valeurs précises des fréquences peuvent être déterminées à l'aide d'algorithmes de recherche d'extrema comme la méthode du gradient (SM, p. 182).

Tests

Des tests généraux indépendants de la méthode d'estimation de la DSP sont présentés dans un premier temps. Par la suite, ceux dédiés aux DSP calculées avec une méthode non-paramétrique sont détaillés. Les tests dédiés aux DSP calculées avec une méthode paramétrique ne sont pas présentés du fait de leur petit nombre et de leur intérêt limité lié aux bonnes propriétés statistiques de ces méthodes (SM, p. 86). Le troisième paragraphe enfin présente les tests statistiques spécifiques à certaines approches du calcul de la DSP.

Tests généraux Les tests généraux présentés ici servent soit à détecter une série complètement aléatoire et donc non périodique, soit à tester la validité statistique d'une DSP sans a priori sur la méthode utilisée.

Différents tests permettent de vérifier que les valeurs d'une série sont aléatoires (Brockwell & Davis, 2002, p. 35). En ce cas, la série ne peut être périodique et ce type de test peut donc être réalisé avant toute analyse de périodicité.

Le premier test est basé sur la séquence d'autocorrélation de la série qui suit une distribution $\mathcal{N}(0, 1/n)$ si les données sont aléatoires. Le test « du portemanteau » est similaire mais considère le carré de la séquence d'autocorrélation, qui suit alors une loi du χ^2. Enfin, le test du *turning point* étudie les groupes de trois points successifs et compte ceux où le point central est inférieur ou supérieur aux deux autres : cette statistique est distribuée selon une $\mathcal{N}(2n/3, 8n/45)$ pour un signal aléatoire.

Une autre famille de tests statistiques est basée sur le calcul des DSP de séries construites par permutations des points de la série originale. Le nombre de fois où la valeur d'un pic de la DSP sur la série d'origine est supérieure aux pics des DSP des séries permutées donne la statistique de significativité de ce pic. L'intérêt de cette approche réside dans sa capacité à donner de bons résultats sur des séries courtes et à fonctionner avec toutes les méthodes d'estimation spectrale. Elle est en revanche très coûteuse en temps de calcul : Pardo-Igúzquizaa & Rodriguez-Tovarb (2000); Krawczyk & Krapiec (2010) réalisent ces tests avec au moins 1000 séries permutées.

Tests pour les méthodes non-paramétriques Les tests présentés pour les méthodes non-paramétriques permettent de traiter deux types de bruits, blanc ou coloré, dont la différence provient de leur représentation fréquentielle, constante pour le premier et non pour le second (PW, p. 489).

Bruit blanc Dans le cas d'une série composée d'un bruit blanc et de sinusoïdes possédant un spectre discret (cf. section 4.3.1), chaque valeur du périodogramme non fenêtré suit une loi du χ^2 à deux degrés de liberté (SM, p. 174; PW, p. 489).

Les tests construits sur cette statistique sont peu robustes car basés sur la variance de la série qui ne peut être qu'estimée. Fischer propose donc un test exact basé sur la statistique du pic de plus grande puissance rapporté à la somme des puissances du périodogramme (PW, p. 491). Percival & Walden (1998, p. 492) introduisent un test présenté

par Siegel et basé sur celui de Fischer pour tester la présence significative de plusieurs pics.

Durnerin (1999, p. 116) propose un test basé sur l'écart moyen entre les pics significatifs de la séquence d'autocorrélation. Si la variabilité de cet écart moyen est faible, le premier pic significatif après les premiers lags de la séquence est une période de la série. L'approche retenue par DPE proposée au chapitre 5 utilise un principe équivalent.

Bruit coloré Vaughan (2005, 2010) propose différents tests statistiques pour la validation d'un pic de la DSP d'un signal additionné d'un bruit coloré. Thomson (1982, p. 1082) en décrit également un basé sur le test de Fisher d'adéquation, comme rappelé dans (PW, p. 496).

Tests propres aux méthodes Certaines des méthodes spécifiques présentées dans la section 4.3.1 p. 80 proposent des tests particuliers : Huijse Heise et al. (2012) dans le cadre du calcul de la DSP basée sur la correntropie, Scargle (1982) pour sa méthode de calcul du périodogramme sur des données irrégulièrement échantillonnées, Sokolove & Bushell (1978) pour le périodogramme χ^2.

4.4 Représentations temporo-fréquentielles

Les représentations temporo-fréquentielles permettent d'avoir la vision en temps et en fréquence d'un signal et de s'affranchir de l'hypothèse de stationnarité associée aux représentations fréquentielles traitées dans la section précédente. Dans ce cadre, la période d'une série est ici associée à chacune des dates de la série étudiée.

La représentation temps-fréquence (T-F) du signal permet d'associer la puissance du signal (ou une mesure équivalente) à un point du plan temps-fréquence. Cette représentation 3D permet donc la combinaison des représentations 2D $temps \times amplitude$ et $fréquence \times puissance$.

Comme pour les sections précédentes, les différentes méthodes de représentation sont introduites dans un premier temps, suivies de celles liées à leur exploitation.

4.4.1 Représentations temps-fréquence

Un très grand nombre d'approches a été proposé pour représenter le signal dans le domaine T-F, réparties en non-paramétriques et paramétriques à l'instar de celles du domaine fréquentiel. Demars (2005) et Bradford (2007) donnent des listes très complètes des premières tandis que les secondes, moins développées, sont traitées par Leonowicz (2006) sous un angle théorique et comparées par Poulimenos & Fassois (2006).

Les quatre premiers paragraphes de cette sous-section introduisent les méthodes non-paramétriques basées sur le calcul de la fréquence instantanée, sur les transformations linéaires, quadratiques puis les approches plus récentes par décompositions successives. Le dernier paragraphe introduit les méthodes paramétriques.

Fréquence instantanée

La fréquence instantanée est la dérivée première de la phase $\phi'(t)$ pour un signal de la forme $a(t)cos(\phi(t))$ avec $a(t)$ l'amplitude, $\phi(t) = 2\pi ft + \phi_0$ la phase et ϕ_0 le décalage initial en phase (MA, p. 137; Flandrin, 1998, p. 27; Gonçalves et al., 1997, p. 38).

La fréquence instantanée est utilisée en modulation de fréquence. Bien que simple et précise, elle ne permet pas d'étudier les signaux avec plusieurs composantes périodiques, puisqu'elle renvoie en ce cas la moyenne de leurs fréquences.

Transformées linéaires / pavage du plan T-F

Gabor (1946) propose une famille d'approches basée sur la projection du signal sur des blocs ou *atomes* du plan T-F translatés en temps et en fréquence. Le résultat de la projection mesure l'importance de l'atome considéré dans le signal pour un temps et une fréquence donnés, tout comme la projection du signal sur une sinusoïde dans la transformée de Fourier donne l'importance de sa fréquence dans l'ensemble du signal.

Cependant, ces atomes ne peuvent être aussi précis que possible : le produit de la largeur en temps et de la hauteur en fréquence est supérieur ou égal à 1/2 (Auger et al., 2013, p. 32; MA, p. 107).

La TF à temps court ou *Short Time Fourier Transform* (STFT) proposée par Gabor (1946) utilise un atome rectangulaire. Cette approche est équivalente au calcul de la TF du signal appliqué à une fenêtre temporelle glissante (Gonçalves et al., 1997). D'autres approches fenêtrées sont utilisées pour déterminer des représentations temps-fréquence avec des méthodes autres que la TF, comme la méthode MFCC basée sur le calcul du cepstre (Ganchev et al., 2005).

La transformée en ondelettes (MA, p. 30) utilise un atome translaté dans le temps et *dilaté* en fréquence construit à partir d'une ondelette mère (MA, p. 119). Le pavage est toujours rectangulaire mais n'est plus constant : lorsque l'intervalle temporel est petit, la bande de fréquence est large et située dans les hautes fréquence. A l'inverse, lorsque l'intervalle temporel est large, la bande de fréquence est étroite et située dans les basses fréquences (MA, p. 27). Ce fonctionnement permet notamment l'étude de pics de haute fréquence précisément localisés dans le temps. De plus, un algorithme de calcul rapide de la transformée en ondelette est disponible (MA, p. 134)

Un certain nombres d'atomes T-F généralisant les transformées précédentes ont également été développés : la transformée de Fourier fractionnaire à temps court qui permet une rotation du pavage obtenu par STFT (Sejdić et al., 2011) ou la transformation canonique linéaire fenêtrée (Kou & Xu, 2012) et les chirplets (Mann & Haykin, 1992) qui permettent de plus un changement d'échelle des atomes après rotation.

Distributions quadratiques / répartition de l'énergie dans le plan T-F

Les distributions quadratiques ou bilinéaires du signal ont pour objectif d'estimer directement la distribution de l'énergie du signal dans le plan T-F (Gonçalves et al., 1997;

Flandrin, 1998, p. 103). Elles permettent de s'affranchir des limites de résolution des atomes T-F ainsi que de la nécessité de leur choix (Daubechies et al. 2011; MA, p. 156).

Les distributions quadratiques invariantes par translation en temps et en fréquence définissent la classe de Cohen tandis que celles invariantes par translation en temps et dilatation en fréquence sont dites affines. Le *spectrogramme*, calculé via la STFT, appartient à la classe de Cohen et le *scalogramme*, calculé via la transformée en ondelettes, appartient à celle des distributions affines (Gonçalves et al., 1997).

Parmi ces transformations, la transformée de Wigner-Ville (TWV) occupe une place particulière puisque les autres distributions quadratiques peuvent s'écrire comme une TWV multipliée par une fenêtre particulière (MA, p. 156). Cependant, toutes ces distributions engendrent des interférences dans le plan temps-fréquence et la TWV peut avoir des valeurs négatives, en contradiction avec la notion de distribution.

Certaines distributions ont ainsi été proposées afin de réduire les termes d'interférences, comme les transformées de Choï-Williams, de Born-Jordan, de Wigner-Ville lissée ou de Zhao-Atlas-Marks entre autres (Bradford, 2007, pp. 28-31).

Ainsi, le compromis temps / fréquence lié au choix de l'atome pour les transformations linéaires devient un compromis résolution / interférences pour les distributions quadratiques (Flandrin et al., 2002, p. 182).

Transformées par décomposition

Le principe des transformées par décomposition est l'identification des composantes périodiques du signal (*Intrisinc Mode Functions* ou *IMF*) par décompositions successives basées sur les données uniquement. Elles ne nécessitent donc pas de choisir un atome T-F ou une transformée spécifique (Daubechies et al., 2011).

La transformée de Hilbert-Huang (HHT), proposée par Huang et al. (1998), est la première méthode de décomposition en IMF. Elle est basée sur la décomposition modale empirique qui permet l'extraction récursive des IMF. Une extension *Ensemble* (Wu & Huang, 2009) permettant d'en réduire la variance et une version pour séries multivariées (Rehman & Mandic, 2010) ont été proposées proposées.

La HTT est efficace (Ke et al., 2014) mais manque de fondements théoriques (Flandrin et al., 2004; Huang & Wu, 2008; Daubechies et al., 2011). Afin d'en améliorer la formalisation, Gilles (2013) propose une méthode hybride basée sur des ondelettes. Dragomiretskiy & Zosso (2014) introduisent une méthode non récursive présentée comme un problème d'optimisation. Enfin, Frei & Osorio (2007) proposent une alternative à la HHT qui fonctionne également de manière incrémentale.

Méthodes paramétriques

Au même titre que la STFT correspond à une application de la méthode non paramétrique de la transformée de Fourier sur des sous-parties du signal, certaines des méthodes paramétriques d'estimation de la DSP présentées dans la section 4.3.1 p. 80 peuvent égale-

ment être appliquées de manière fenêtrée sur le signal. Ces approches reposent sur l'hypothèse qu'un signal non-stationnaire peut être considéré localement stationnaire pour des fenêtres suffisamment petites.

Avec une fenêtre glissante dans le temps sur le signal, Leonowicz et al. (2002) proposent ainsi d'appliquer la méthode Min-Norm et Poulimenos & Fassois (2006) présentent un état de l'art de l'utilisation des approches AR et ARMA dans ce cadre.

4.4.2 Exploitation des représentations T-F

La représentation T-F d'un signal permet d'une part d'en analyser les différentes composantes périodiques grâce à la décomposition fréquentielle et d'autre part de les localiser dans le temps grâce à l'information temporelle qui leur est associée. Cette représentation est donc adaptée au calcul de périodes non constantes ainsi qu'aux composantes multiples illustrées sur la figure 4.2 p. 71.

Comme dans le domaine fréquentiel, les maxima de ces représentations donnent la ou les périodes du signal. Elles sont de plus attachées à une information temporelle, ce qui rend leur analyse plus complète mais également plus complexe.

Les méthodes d'exploitation sont présentées dans les paragraphes ci-dessous, d'abord pour les représentations obtenues par transformées linéaires ou par distributions quadratiques puis pour celles construites par décomposition.

Exploitation des représentations linéaires et quadratiques

La résolution de ces représentations doit être améliorée dans un premier temps avant d'être analysées à l'aide de tests statistiques.

Amélioration de la résolution Les transformations linéaires sont bruitées du fait de l'incertitude liée à la projection du signal sur des atomes T-F et les transformées quadratiques le sont à cause de la présence de termes d'interférences.

Afin de limiter ces effets d'étalement, différentes approches d'amélioration de la résolution sont introduites. Le calcul des crêtes (*ridges*) (MA, pp. 139 et 149) et les méthodes de réallocation (Auger et al., 2013) et de synchrosqueezing (Maes, 1994) permettent de ne retenir que la valeur maximale de la projection du signal sur un atome T-F. Les opérateurs de morphologie mathématique, plus largement détaillés au chapitre 5, sont également utilisés pour ne retenir que les valeurs maximales de la représentation (Evans et al., 2002; Borda et al., 2005).

Propriétés statistiques Les contributions suivantes détaillent les distributions des coefficients calculés pour les représentations T-F, permettant la construction de tests destinés à distinguer du bruit les périodes réellement présentes dans le signal. Différentes approches prenant en compte des signaux bruités vus comme la somme du signal d'origine et d'un bruit, blanc ou coloré (cf. section 4.3.2 p. 82) sont proposées.

Millioz et al. (2006) montrent que les coefficients de la STFT suivent une distribution gaussienne et Torrence & Compo (1998) que les coefficients d'ondelettes dans du bruit blanc et coloré suivent une distribution du χ^2. Chassande-Motin et al. (1998) détaillent les statistiques associées aux méthodes de réallocation. Sayeed & Jones (1995) proposent un ensemble de tests statistiques basés sur des représentations T-F pour distinguer les cas signal vs. signal + bruit selon que le bruit est blanc, coloré ou que le rapport signal / bruit est faible. Enfin, Foster (1996) donne des statistiques sur les coefficients d'une transformée en ondelettes pour des données échantillonnées irrégulièrement.

Exploitation des représentations par décomposition

Huang & Wu (2008) présentent un certain nombre d'approches statistiques pour interpréter les représentations obtenues par HHT. Deux méthodes de calcul des intervalles de confiance pour les IMF sont également proposées.

L'étude de la HHT de différents bruits gaussiens permet d'établir des tests statistiques pour différencier les données composés de bruit uniquement de celles comportant également un signal (Flandrin et al., 2004).

De manière similaire aux tests par permutation présentés dans la section 4.3.2 p. 82, une autre approche basée sur la génération de signaux aléatoires du même type que le signal en entrée permet également de distinguer les IMF issues du bruit des IMF significatives (Huang & Wu, 2008).

Dans tous les cas, ces méthodes ne présentent pas directement une technique d'extraction de la période associée à un intervalle temporel, cette dernière étant laissée à l'utilisateur de ces statistiques.

4.5 Représentations symboliques

Les séries symboliques associent une ou plusieurs valeurs symboliques à une date donnée. Comme détaillé dans la section 4.1.1 p. 69, certaines sont obtenues à partir de séries numériques après symbolisation, d'autres le sont d'emblée, comme les séquences d'ADN ou les logs d'événements.

A l'instar des deux sections précédentes, celle-ci contient une première sous-section dédiée à la représentation des séries par symbolisation et une seconde liée à leur exploitation.

4.5.1 Représentation par symbolisation

Le processus de symbolisation permet de convertir une série numérique en série symbolique. Il a l'intérêt de réduire la complexité de la série et d'en retirer le bruit (Daw et al., 2003; Sant'Anna & Wickstrom, 2011).

La symbolisation agit en discrétisant la série numérique en temps et/ou en valeurs dans un alphabet $\Sigma = \{a_1, ..., a_\alpha\}$ contenant α caractères. La discrétisation en temps est appelé *segmentation* temporelle et celle en valeur *quantification* (*quantization*). Les w segments de taille $l = n/w$ issus de la segmentation sont notés $s_1, ..., s_w$ avec $w \leq n$.

L'évaluation de la période et de la périodicité dans le domaine symbolique utilise la comparaison des segments entre eux et nécessite donc la définition de distances particulières, présentées dans le premier paragraphe. Par la suite, les différentes méthodes de symbolisation sont détaillées, en valeurs uniquement, en temps et en valeurs et enfin en temps seulement. Sant'Anna & Wickstrom (2011) proposent une comparaison de de certaines d'entre elles.

Distances

En supposant deux segments a et b, la distance la plus simple dans le domaine symbolique est la distance de Hamming d_H, égale au nombre de caractères différant entre les séquences, i.e. $d_H = \sum[a_i \neq b_i]$ où [.] est le crochet d'Iverson qui renvoie 1 si l'expression évaluée est vraie et 0 sinon (Knuth, 1992).

Cette distance est cependant très sensible aux variations même légères en temps et en valeurs. Afin de remédier à ce problème, différentes adaptations ont été proposées.

La plus courante est la distance DTW (*Dynamic Time Warping*) qui est une distance d'édition entre segments permettant de rendre la distance de Hamming plus souple aux décalages temporels au prix d'un coût de calcul plus important (Keogh & Ratanamahatana, 2005). Elle est définie récursivement par $d_D(a,b) = d_H(a_1,b_1) + \min(d_D(a_{2:},b), d_D(a,b_{2:}), d_D(a_{2:},b_{2:}))$ (Elfeky et al., 2005b), où $a_{2:}$ et $b_{2:}$ désignent les séquences a et b privées de leur premier caractère.

D'autres adaptations plus spécifiques ont également été proposées. Han et al. (1998) proposent l'utilisation de jokers représentés par le caractère * égal par convention à tous les caractères. La distance de Hamming avec joker est définie par $d_{H*}(a,b) = \sum[a_i \neq b_i \wedge a_i \neq * \wedge b_i = *]$.

Mannila et al. (1997) proposent également de rendre la distance de Hamming insensible aux permutations. Les auteurs ne définissent pas de distance à proprement parler, mais cette dernière peut être définie comme $d_M(a,b) = \sum \notin (a_i, b)$ où la fonction $\notin (x,y)$ renvoie 1 si le symbole x n'est pas dans la séquence y et 0 sinon.

Enfin, Lin et al. (2002) proposent une distance exploitant l'ordre des symboles d'une séquence symbolisée à l'aide de seuils (cf. paragraphe suivant). Par exemple, la quantification $s_i = a$ si $x_i < 0$ et b sinon définit un ordre sur l'alphabet $\Sigma = \{a, b\}$, ici $a < b$. La distance définie permet alors de considérer comme égaux deux symboles dont l'un est successeur de l'autre, i.e. $d_L(a,b) = \sum[a_i \neq b_i \wedge a_i \neq succ(b_i) \wedge b_i \neq succ(a_i)]$ où $succ(x)$ désigne le successeur de x dans l'ensemble ordonné des symboles Σ.

Plus généralement, les noyaux de séquence (Lodhi et al., 2002) permettent la comparaison entre deux segments. Ces derniers ne sont pas détaillés ici car les méthodes présentées ci-dessous n'en font pas usage.

Quantification seule

Les méthodes de quantification seule symbolisent la série numérique en associant un symbole à chacune de ses dates, générant une série symbolique de même taille ($w = n$).

Les méthodes les plus simples sont basées sur une échelle a priori associant un symbole à une valeur dans un intervalle, par exemple a si $x_i < 0,5$ et b sinon (Daw et al., 2003).

D'autres méthodes utilisent une échelle construite sur la moyenne empirique μ de la série numérique. Par exemple, $s_i = a$ si $x_i \leq \mu$ et b sinon (Bagnall et al., 2006).

La symbolisation des tendances plutôt que des valeurs de la série est aussi envisagée (Andre-Jonsson & Badal, 1997).

Enfin, Mörchen & Ultsch (2005) proposent la méthode Persist qui prend en entrée la taille de l'alphabet $\alpha > 1$ et dont l'objectif est de renvoyer une série composée d'un maximum d'états persistants, i.e. dont les valeurs successives sont globalement constantes.

Quantification avec segmentation

Les méthodes de quantification avec segmentation discrétisent la série sur les deux dimensions temps / valeurs ($w < n$).

La méthode SAX (Lin et al., 2002) réalise la segmentation d'une série supposée gaussienne avec un nombre de symboles α et une taille des segments l fournis par l'utilisateur. Le principe de SAX est de générer une série symbolique contenant un nombre d'occurrence à peu près égal de chaque symbole. Cette méthode, quoique couramment utilisée (Androulakis, 2005; Tanaka et al., 2005; Minnen et al., 2007) et rapide d'exécution (Renard et al., 2015), est fortement dépendante de ses paramètres.

Une variante proposée par Li et al. (2012) pour s'affranchir du paramètre l repose sur le découpage de la série en fenêtres recouvrantes.

Qu et al. (1998) proposent une symbolisation en deux temps. D'abord, la série est divisée en segments de taille l fournie par l'utilisateur, puis la pente des valeurs sur chaque segment est évaluée par régression linéaire. Si l'erreur quadratique entre la droite de régression et les données est supérieure à un seuil, le segment est ignoré.

La méthode PLA (*Piecewise Linear Agregation*) de linéarisation par morceaux (Keogh et al., 2001; Morinaka et al., 2001) permet de représenter une série temporelle par une suite de droites affines dont la taille en nombre de points n'a pas à être spécifiée en amont.

Symbolisation par clustering Le principe de ces approches est de diviser la série en segments de taille l et de les regrouper par similarité en α groupes, l et α étant spécifiés par l'utilisateur. A la différence de SAX, ces méthodes peuvent renvoyer des segments de tailles différentes.

Wang & Megalooikonomou (2008) utilisent un algorithme de type k-moyennes pour réaliser cette opération. Zhou et al. (2008) proposent la méthode ACA, également basée sur les k-moyennes mais utilisant DTW au lieu de la distance euclidienne et permettant que les symboles codent pour des segments de tailles potentiellement différentes. Hugueney (2006) propose la méthode SBSR-L0 qui fonctionne également par clustering mais sans contrainte sur la taille des segments.

Segmentation seule

Les méthodes de segmentation seule permettent de découper la série dans le temps sans en modifier les valeurs. Ces approches, également dites de discrétisation, sont décrites par Liu et al. (2002); Ramírez-Gallego et al. (2016). Elles peuvent être utilisées pour des séries symboliques, l'objectif étant d'en diminuer la taille.

Tanaka et al. (2005) proposent d'associer un symbole aux groupes de symboles récurrents d'une série symbolique, par exemple obtenus avec SAX. Otunba et al. (2014) utilisent une approche similaire basée sur l'extraction automatique d'une grammaire à partir de la série en entrée : ainsi la série *abcdbcabcd* est représentée par CAC à l'aide d'une grammaire possédant les deux règles $A = bc$ et $C = aAd$.

4.5.2 Exploitation des séries symboliques

L'intérêt principal des méthodes symboliques réside dans leur capacité à décrire un motif périodique et pas uniquement à renvoyer une période comme les méthodes précédentes. Elles permettent aussi de prendre en charge l'analyse de séries multivariées, souvent plus complexe dans les domaines présentés précédemment.

Elles sont en revanche moins souples que ces dernières concernant les approximations en temps et en valeurs des éléments périodiques (cf. section 4.1.2 p. 70) du fait de leur représentation dans un espace discret.

Les paragraphes suivants présentent successivement les méthodes traitant les séries univariées puis celles multivariées.

Séries univariées

Li et al. (2015) proposent une méthode de détection de la périodicité pour des séries symboliques univariées à deux symboles et à temps irrégulier. Le principe utilisé est similaire à celui des méthodes par actogramme présentées dans la section 4.2.3 p. 75 et permet donc de détecter les composantes de même période et de phases différentes. La méthode est justifiée théoriquement et robuste au bruit ainsi qu'aux valeurs manquantes. Elle est cependant de complexité quadratique car elle nécessite d'être exécutée pour chaque période candidate.

Ergün et al. (2010) proposent une approche en flux permettant de traiter des données symboliques univariées à temps régulier basée sur l'algorithme de hachage de Rabin-Karp qui permet de détecter rapidement l'occurrence ou non d'un motif dans une série. L'algorithme proposé a une complexité en $O(n \log n)$ et renvoie la période de la série si elle existe, sinon ne renvoie rien. Une mesure de périodicité basée sur le nombre d'opérations à effectuer sur la série pour la rendre périodique est également proposée. La méthode ne gère pas en revanche les approximations en temps et en valeur.

Otunba et al. (2014) introduisent un algorithme de détection de la périodicité de séries numériques dans un espace symbolique après symbolisation par SAX. Les règles de réécriture issues d'une grammaire (cf. ci-dessus) sont mises à profit pour identifier les motifs

fréquents. L'écart moyen entre deux instances successives d'un motif en donne la période et la variabilité de cet écart sa périodicité. La méthode a l'avantage d'être simple et de permettre une fonctionnement incrémental. Elle dépend en revanche des paramètres de symbolisation, notamment la taille de l'alphabet.

Arora et al. (2008) analysent des séquences d'ADN en supposant un modèle cyclostationnaire, i.e. où les propriétés statistiques du signal sont périodiques de période k dans le temps et non constantes comme dans le cas de la stationnarité. L'intérêt de cette méthode est qu'elle permet d'identifier des motifs disjonctifs, par exemple *AG(C/T)A* qui correspond à *AGCA ou AGTA*.

Adalbjornsson et al. (2015) traitent le problème des répétitions de symboles et non de motifs et le résolvent en estimant la distribution de chaque symbole pour des ensembles d'indices périodiques. Ainsi, la période estimée ne correspond qu'à des caractères simples.

Séries multivariées

Les méthodes d'analyse de la périodicité pour les séries symboliques multivariées sont majoritairement issues du domaine de l'extraction des règles d'association temporelles. Dans ce contexte, un k-itemset est un ensemble de k symboles associés à une date.

Un l-k-itemset est un motif composé de l k-itemsets associés à des dates successives. Le terme générique d'itemset peut être utilisé pour désigner un l-k-itemset ou un k-itemset lorsque le contexte est suffisant.

La recherche de motifs périodiques passe par celle des itemsets fréquents. Un itemset est fréquent si son support, calculé comme son nombre d'occurrences rapporté à la longueur n de la série, est supérieur à un seuil utilisateur *minsup*. Le support est à la base de diverses optimisations dans la recherche d'itemsets fréquents (Agrawal et al., 1993, 1995).

Les méthodes dédiées au calcul de la période des k-itemsets sont présentées dans un premier temps, suivies par celles traitant des l-k-itemsets.

Périodicité des k-itemsets Ozden et al. (1998) proposent une méthode de détection de la périodicité par *augmentation* des k-itemsets, i.e. en calculant d'abord la période des caractères (1-itemsets), puis celle des 2-itemsets, des 3-itemsets etc. Le calcul de la périodicité des $(k+1)$-itemsets est basé sur une optimisation similaire à celle d'Apriori et basée sur le constat qu'un $(k+1)$-itemset périodique est nécessairement constitué de k-itemsets périodiques. Les $(k+1)$-itemsets candidats sont donc construits à partir des k-itemsets fréquents et périodiques et non à partir des combinaisons possibles de $k+1$ caractères tirés de l'alphabet Σ.

De plus, les périodes potentielles des $(k+1)$-itemsets sont celles des k-itemsets ou de leurs multiples. Si par exemple les périodes des 1-itemsets A et B sont 2 et 3 respectivement et que leur première occurrence est à t_1, les seules dates à étudier pour le 2-itemset AB sont les multiples de 6, i.e. t_1, t_7, t_{13} etc. Cette méthode est peu robuste au bruit en temps et en valeurs.

Ma & Hellerstein (2001) proposent une méthode exploitant des données à temps irrégulier et basée sur le principe d'augmentation mais plus robuste au bruit. Concernant le bruit en valeur, un seuil sur le support de l'itemset est utilisé afin de retenir ceux dont quelques occurrences seulement sont manquantes. Pour le bruit en temps, le méthode exploite un paramètre de tolérance δ fourni par l'utilisateur destiné à identifier comme périodiques une occurrence de l'itemset et sa suivante si leur écart dans le temps est compris dans $[p-\delta; p+\delta]$ où p est une période candidate. Le nombre d'occurrences d'un symbole vérifiant cette propriété rapporté à celui qui serait obtenu si les symboles étaient distribués de manière aléatoire selon une loi du χ^2, qui permet la définition d'un test statistique pour retenir les périodes candidates.

La méthode fonctionne rapidement mais est sensible au bruit car la robustesse de l'algorithme est dépendante du paramètre utilisateur δ.

Périodicité des l-k-itemsets D'autres méthodes utilisant des concepts similaires ont été proposées pour étudier la périodicité de l-k-itemsets, i.e. de motifs de taille l composés d'itemsets de taille inférieure ou égale à k.

Han et al. (1999) introduisent une approche permettant de rechercher des motifs disjonctifs utilisant des jokers, comme par exemple $a\{b,c\}d{*}{*}f$, qui peut correspondre à $abd{*}{*}f$ ou $acd{*}{*}f$ et où les jokers * peuvent prendre une valeur quelconque de l'alphabet Σ.

Comme pour les méthodes précédentes, la périodicité des symboles (motifs constitués d'un seul caractère) est étudiée pour une période candidate donnée. Les motifs extraits sont ensuite combinés pour définir le motif potentiel le plus précis, i.e. contenant tous les symboles périodiques trouvés. Si par exemple les motifs $a{*}{*}{*}$, $*b{*}{*}$, $**c*$ sont déterminés lors de la première passe, le motif le plus précis est $abc*$.

La série est ensuite découpée en segments consécutifs de taille p et les motifs les plus précis sont comptés et les plus fréquents d'entre eux sont renvoyés.

La méthode est intéressante en ce qu'elle permet l'identification de motifs périodiques complexes. Elle est néanmoins coûteuse en temps de calcul puisque le processus doit être répété pour chaque période candidate.

Aref et al. (2004) proposent une version incrémentale plus rapide de cette méthode et Elfeky et al. (2005a) une version accélérée à l'aide d'un calcul des périodes pour les symboles basé sur une convolution évaluée par transformée de Fourier rapide.

Néanmoins, la méthode de Han et al. (1999) suppose que les motifs se répètent parfaitement tout au long de la série, i.e. que la série n'est composée que de leur répétition, sans espace ni recouvrement. Yang et al. (2000) détaillent une approche permettant d'assouplir cette contrainte en ajoutant deux paramètres, l'un pour le nombre minimal de répétitions du motif, l'autre pour déterminer l'écart maximal entre deux occurrences successives. Le nombre minimal de répétitions permet l'identification de la périodicité locale d'un motif, par opposition au support qui est calculé sur l'ensemble des données. Le paramètre d'écart maximal permet lui d'introduire de la souplesse dans la répétition des motifs. L'expressivité qu'offre la méthode est contrebalancée par sa complexité. D'autre part, elle doit

également être exécutée pour un intervalle de périodes candidates.

Elfeky et al. (2005b) proposent une méthode équivalente en utilisant la distance DTW. Le seuil d'écart maximal disparaît mais un autre paramètre équivalent lui est substitué pour définir cette distance. Elfeky et al. (2006) accélèrent cette approche en permettant son calcul de manière fenêtrée.

Huang & Chang (2005) étendent ces méthodes afin de permettre l'identification de la périodicité de l-k-itemsets. Leur algorithme bénéficie des différents atouts des propositions précédentes mais sa complexité est importante. Elle est de plus très dépendante des différents paramètres fournis par l'utilisateur.

4.6 Autres représentations

D'autres représentations des séries temporelles ont été proposées pour calculer leur période. Ces dernières, nettement moins courantes que celles présentées dans les sections précédentes, sont brièvement décrites dans les paragraphes suivants.

4.6.1 Approches par graphes

Ferreira & Zhao (2014) proposent de convertir une série temporelle en graphe en divisant la série initiale en plusieurs segments correspondants chacun à un nœud du graphe. Un lien est établi entre deux nœuds si deux valeurs égales de la série sont présentes dans les deux segments correspondants. Un algorithme de détection de communautés est ensuite exécuté et une période est renvoyée lorsque l'appartenance aux communautés des valeurs de la série prises séquentiellement est cyclique. La méthode est décrite comme robuste aux valeurs manquantes mais est dépendante de sa discrétisation initiale.

4.6.2 Espace de phases

La représentation d'un signal dans un espace de phases en deux dimensions est une représentation paramétrique telle que $\forall t = 1, ..., n-1$, $x(t) = x_t$ et $y(t) = x_{t+k}$ où k est un retard ou lag. Ces représentations sont également appelées *lag scatter plot* (Percival & Walden, 1998, p.4).

Elles possèdent des propriétés topologiques particulières et sont notamment fermées lorsqu'elles correspondent à un signal périodique (Gerhard, 2003). Dans le cas de signaux réels toutefois, cette propriété n'est pas exactement vérifiée, entraînant l'utilisation de techniques variées comme la triangulation de Delaunay pour la retrouver (Emrani et al., 2014).

4.6.3 Approches floue

Règles Novák et al. (2008) proposent d'évaluer la périodicité d'une série temporelle à l'aide de règles floues définies par l'utilisateur. Ces dernières sont basées sur une métrique Q calculée comme la différence entre deux valeurs séparées de k points. Pour celles

correspondant à la période du signal, $Q(k)$ doit être faible. La périodicité du signal est donc évaluée à l'aide de la règle « Si $Q(k)$ est Très Faible et $Q(2k)$ est Très Faible alors la périodicité est Très Élevée », où *Faible* et *Élevé* sont des variables linguistiques définies par l'utilisateur et *Très* est un modificateur.

Machine à états finis Dans le cadre de l'étude de la marche humaine, Sanchez-Valdes & Triviño (2013) proposent d'associer l'une des modalités d'une variable linguistique à chaque valeur de la série temporelle puis à rapprocher cette modalité d'un état d'une machine floue à états finis.

Du fait des paramètres nécessaires à la définition de la variable linguistique et des transitions d'un état au suivant dans la machine à états finis, cette approche permet de reconnaître un type de signal donné. Elle n'est pas bien adaptée cependant au calcul d'une période inconnue a priori.

4.6.4 Méthodes hybrides

Différentes représentations du signal peuvent être utilisées afin de consolider l'analyse produite par différents points de vue ou bien pour initialiser certains paramètres.

Kedem (1986) propose par exemple d'utiliser le nombre de croisements du signal avec l'axe des abscisses, rapidement calculé et dont le comportement statistique est connu, pour confirmer ou infirmer une période détectée avec une autre méthode.

Un certain nombre de méthodes utilisent également la FFT en complément d'autres approches. Berberidis et al. (2002) par exemple l'exploitent pour détecter rapidement les périodes candidates avant d'identifier plus précisément les motifs périodiques d'une série symbolique avec la méthode de Han et al. (1999).

Vlachos et al. (2005) introduisent l'algorithme AUTOPERIOD qui utilise les pics de la DSP pour identifier les périodes candidates et les valident en s'assurant que ces dernières correspondent également à des pics dans la séquence d'autocorrélation.

Plautz et al. (1997) initialisent le calcul des paramètres d'un modèle construit comme une somme de sinusoïdes à l'aide d'une FFT dont les valeurs les plus importantes donnent les périodes. Les paramètres d'amplitude et de phase sont ensuite déterminés par régression au sens des moindres carrés. Yang & Su (2010) utilisent une méthode similaire avec une DSP calculée à l'aide d'un processus AR.

Papadimitriou et al. (2003) proposent la méthode incrémentale AWSOM de mise à jour d'un modèle AR basé sur les coefficients d'ondelettes calculés sur les données déjà reçues.

Enfin, Leise et al. (2013) proposent d'analyser automatiquement les segments utilisés pour les actogrammes en détectant leur phase à l'aide des coefficients d'ondelettes associés aux fréquences basses.

Tableau 4.1 – Comparaison des méthodes en fonction de leur domaine de représentation

Domaine	*Avantages*	*Inconvénients*
Temporel	- Diversité des méthodes adaptées à de nombreux cas de figure	- Signaux stationnaires
Fréquentiel	- Standard - Rapide - Multi composantes - Tests statistiques	- Signaux stationnaires
Temporo-fréquentiel	- Rapide - Multi composantes - Périodicité locale - Périodicité évolutive	- Exploitation complexe - Tests statistiques variables - Choix de l'ondelette
Symbolique	- Motifs - Nombreuses méthodes en multivarié - Périodicité locale	- Complexité - Paramètres nombreux

4.7 Bilan

Cet état de l'art propose une vision large des différentes approches existantes pour le calcul de la période et de la périodicité dans une série temporelle. Il est à notre connaissance le seul recensant les propositions faites dans différents domaines pour résoudre ces questions, du fait peut être du nombre très important de travaux proposés dans des champs scientifiques variés.

Deux taxonomies sont également introduites, l'une pour classer les séries temporelles, représentée sur la figure 4.1 p. 68, et l'autre pour désigner les différents cas de figures pris en compte dans ce chapitre, représentée sur la figure 4.2 p. 71.

Les avantages et les inconvénients des différentes représentations utilisées pour la détermination de la période et de la périodicité présentées dans ce chapitre sont synthétisés dans le tableau 4.1.

Enfin, cet état de l'art permet de situer la méthode DPE que nous présentons au chapitre suivant qui permet de calculer la période d'une série temporelle sans paramètre a priori, de manière rapide et qui fournit de plus une estimation de sa périodicité, contrairement à la majorité des méthodes présentées précédemment.

Chapitre 5

Détection d'évènements périodiques : la méthode DPE

> Je doute qu'il arrive jamais à cette simplification, cette « puissante érosion des contours » dont parle Nietzsche, et sans laquelle il n'y a pas de parfaite œuvre d'art.
>
> —André Gide, *Journal 1889-1939*

Ce chapitre décrit la méthode DPE (*Detection of Periodic Events*) que nous avons proposé pour calculer la période et la périodicité d'une série temporelle ainsi que pour produire une phrase la décrivant, de la forme « M toutes les p unités, les valeurs sont élevées », où M est un adverbe et p *unités* une mesure de période. DPE repose sur le principe que la série est *périodique si elle alterne de manière régulière des groupes de valeurs hautes et basses, où la régularité est fonction de leurs tailles respectives.*

Par rapport aux méthodes présentées au chapitre précédente, DPE calcule la périodicité de la série temporelle, en propose un rendu linguistique et fonctionne sans paramètre ni modèle a priori sur les données.

La première section de ce chapitre introduit le principe de fonctionnement de la méthode. Les trois sections suivantes détaillent ses trois étapes, à savoir le clustering des données en groupes de valeurs hautes et groupes de valeurs basses, le calcul de statistiques liées à ces groupes et permettant la détermination de la périodicité et de la période candidate, et enfin le rendu linguistique des éléments calculés.

Les travaux présentés dans ce chapitre ont fait l'objet des deux publications (Moyse et al., 2013a) et (Moyse et al., 2013b).

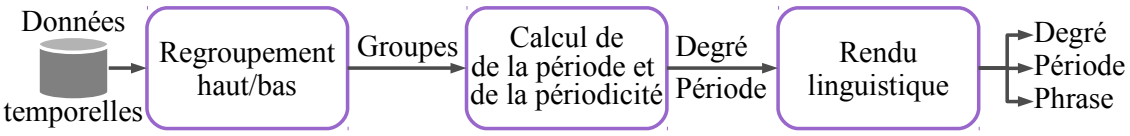


FIGURE 5.1 – Architecture de la méthode DPE

5.1 Architecture

Entrées X est une série à temps régulier de fréquence d'échantillonnage Δt (cf. section 4.1.1 p. 68) à valeurs dans [0,1], telles que ces bornes sont atteintes :

$$X = \{x_i, i = 1, ..., n\} \text{ tel que } \forall i\, x_i \in [0, 1] \text{ et } \exists i, j \text{ tels que } x_i = 0 \text{ et } x_j = 1 \qquad (5.1)$$

X peut en particulier représenter une série temporelle de degrés d'appartenance à des modalités floues.

Sorties Les résultats produits par DPE sont une période candidate p_c et une périodicité π décrites dans la section 4.1.2 p. 70, ainsi qu'une phrase descriptive de la forme « M toutes les p unités, les valeurs sont élevées ». Lorsque les données en entrée sont des degrés d'appartenance à la modalité P, cette phrase peut être interprétée comme « M toutes les p unités, les x sont P ».

Dans la phrase renvoyée, M représente un adverbe comme « exactement », « environ » ou « grossièrement » et p *unités* est la représentation textuelle de p_c avec *unités* représentant une unité de temps comme « heures », « jours » ou « secondes ».

Les trois étapes de la méthode La méthode DPE repose sur le postulat intuitif qu'une série est *périodique si elle alterne de manière régulière des groupes de valeurs hautes et basses, où la régularité est fonction de leurs tailles respectives*. DPE fonctionne donc par identification des groupes de valeurs hautes et basses puis par estimation de la régularité de leur alternance.

Plus précisément, la méthode est composée des trois étapes illustrées sur la figure 5.1. La première étape, décrite dans la section 5.2, réalise un regroupement ou *clustering* des données en groupes de valeurs hautes et groupes de valeurs basses. Dans un second temps, détaillé dans la section 5.3 p. 105, des statistiques visant à estimer la régularité de la taille de ces groupes sont calculées afin de renvoyer le degré de périodicité π et la période candidate p_c. Enfin, ces valeurs sont rendues textuellement lors de la troisième étape de rendu linguistique présentée dans la section 5.4 p. 109.

5.2 Regroupement

La première étape de DPE a pour vocation le clustering des données en groupes de valeurs hautes et groupes de valeurs basses. La figure 5.2 illustre le résultat de ce regrou-

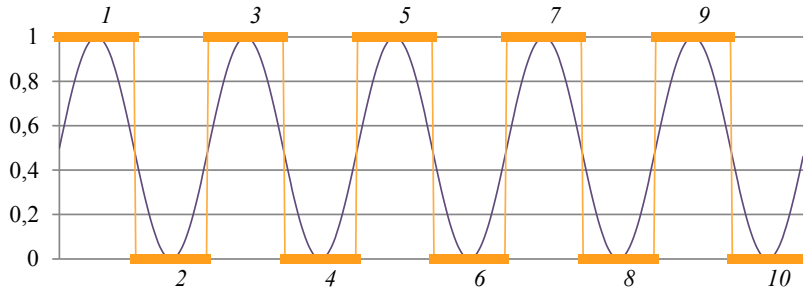


FIGURE 5.2 – Regroupement en groupes de valeurs hautes (impairs) et groupes de valeurs basses (pairs)

pement sur des données sinusoïdales.

La répartition d'une série entre valeurs hautes et basses peut être réalisée par *seuillage*, comme les méthodes SAX et Persist présentées dans la section 4.5.1 p. 89, celle de Silverman (1998) à base d'ondelettes ou celle de Castro et al. (2007) utilisant des statistiques simples. Dans tous les cas, ces méthodes réalisent le calcul d'une valeur globale de seuil. La méthode que nous proposons et mise en œuvre dans DPE a l'avantage d'être locale et donc de s'adapter automatiquement aux données. Nous montrons dans les expériences réalisées au chapitre 7 que ces seuils locaux sont plus appropriés que des valeurs globales. De plus, DPE fonctionne sans paramètres contrairement aux méthodes mentionnées ci-dessus.

Dans la suite de cette section, une formalisation de la méthode de regroupement proposée est présentée, suivie de la description du score d'érosion, une nouvelle transformation basée sur la morphologie mathématique utilisée pour définir un seuil adaptatif local.

5.2.1 Formalisation

L'étape de regroupement des données temporelles repose sur la définition d'une fonction de prédiction :

$$\gamma : X \to \{H, L\} \tag{5.2}$$

qui associe à chaque donnée x_i de X un type, H ou L, selon qu'elle appartient à un groupe de valeurs hautes ou à un groupe de valeurs basses. Les points consécutifs de même type sont regroupés en g groupes rassemblés en une liste ordonnée $G = (G_k)_{k=1...g}$.

L'attribution d'une étiquette par γ à chacun des points n'est pas une tâche de classification puisque γ n'est pas apprise. La méthode est plus proche d'une tâche de clustering avec la particularité que la fonction de similarité utilisée ne dépend pas seulement des valeurs x_i mais aussi de leur position, par le biais de leur voisinage. C'est pourquoi le terme de regroupement est retenu pour qualifier cette méthode.

5.2.2 Le score d'érosion

Nous rappelons dans un premier temps les éléments de morphologie mathématique nécessaires à la formalisation du score d'érosion présenté à leur suite. Nous définissons ensuite γ_{es}, une méthode de regroupement basée sur ce score.

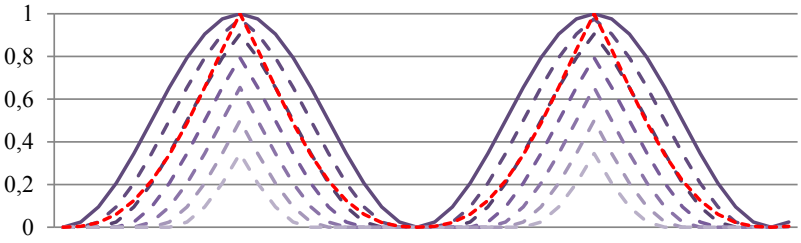


FIGURE 5.3 – En trait plein violet, les données en entrée, en pointillés de plus en plus clair, leurs érosions successives, et en pointillés rouges, le score d'érosion

Morphologie mathématique pour l'analyse de données La morphologie mathématique (MM) propose un ensemble d'opérateurs pour l'analyse de structures spatiales, comme la forme ou la taille des objets. Elle est couramment utilisée pour le traitement, l'analyse, la segmentation ou la compression d'images (Serra, 1986; Najman & Talbot, 2013).

La MM fonctionnelle, ou 1D, ne s'applique pas à des images mais à des fonctions. Elle est utilisée à des fins de débruitage dans différentes applications. Par exemple, ces opérateurs peuvent être utilisés pour simplifier des sous-ensembles flous appris sur des données afin de générer des arbres de décision (Marsala & Bouchon-Meunier, 2003), des clusters (Turpin-Dhilly & Botte-Lecoq, 1998) ou des règles graduelles (Oudni et al., 2013). Lefèvre & Claveau (2011) insistent sur le fait que la MM 1D peut être appliquée à des domaines autres que le traitement d'image et en proposent une extension dans le cadre de l'analyse textuelle. La MM 1D peut aussi être utilisée dans le cadre du traitement du signal (Bangham & Marshall, 1998), avec des applications liées par exemple à la reconnaissance de la parole (Wang et al., 2005) ou à l'analyse d'ECG (Sun et al., 2005).

L'érosion est l'une des deux opérations élémentaires de la MM. Formellement, soit une fonction $f : E \to F$ et un élément structurant B défini comme un sous-ensemble de E, l'érosion est la fonction $\epsilon_B(f) : E \to F$ définie comme (Serra, 1983) :

$$[\epsilon_B(f)](x) = \inf_{b \in B} f(x + b) \tag{5.3}$$

La dilatation, définie identiquement avec un opérateur sup, est l'opération duale de l'érosion. L'érosion, comme la dilatation, peuvent être utilisées de manière répétée et/ou alternée, permettant la création d'opérateurs composés plus complexes, comme l'ouverture, la fermeture ou les filtres alternés (Serra, 1986).

Principe La capacité des outils de morphologie mathématique à retirer le bruit du signal nous a amené à proposer le score d'érosion. Afin de permettre une identification robuste au bruit des groupes de valeurs hautes, nous proposons d'appliquer l'opérateur d'érosion de façon répétée afin d'extraire le « squelette » du groupe, ou dans la même ordre d'idée son centre de gravité. Cette approche est assimilable à celles du feu de forêt (Blum, 1967) ou d'extraction du squelette (Lantuejoul & Maisonneuve, 1984) en morphologie mathématique.

L'originalité de notre approche tient en son application sur des données 1D ainsi qu'en la reconstruction par addition des érosions successives. Elle est illustrée sur la figure 5.3 avec un jeu de données initial en trait plein violet, ses érosions successives en pointillés violet du plus foncé pour la première au plus clair pour la dernière, et enfin la reconstruction par addition en trait plein rouge.

Score d'érosion L'érosion que nous proposons d'utiliser pour le score d'érosion repose sur le plus petit élément structurant symétrique non trivial $B = \{-1, 0, 1\}$. L'érosion de la $i^{ème}$ valeur de X s'écrit alors (cf. éq. (5.3)) :

$$\epsilon_i = \min(x_{i-1}, x_i, x_{i+1}) \qquad \epsilon_1 = \min(x_1, x_2) \qquad \epsilon_n = \min(x_{n-1}, x_n) \tag{5.4}$$

et sa $j^{ème}$ répétition :

$$\epsilon_i^j = \epsilon_i\left(\epsilon_i^{j-1}\right) = \min\left(\epsilon_{i-1}^{j-1}, \epsilon_i^{j-1}, \epsilon_{i+1}^{j-1}\right) \quad \text{et} \quad \epsilon_i^0 = x_i \tag{5.5}$$

Comme au moins une valeur de X est supposée nulle (cf. éq. (5.1) p. 98), l'érosion répétée des valeurs de la série mène à son érosion totale où toutes ses valeurs sont nulles. Le score d'érosion est la somme normalisée de ces érosions successives jusqu'à érosion totale de X. Pour chaque x_i de X, le score d'érosion *non normalisé* est défini comme :

$$es_i = \sum_{j=0}^{z_i} \epsilon_i^j \tag{5.6}$$

où z_i est le nombre d'érosions nécessaires pour atteindre l'érosion totale de x_i :

$$z_i = \arg\min_{j \in \mathbb{N}} \left\{\epsilon_i^j = 0\right\} \tag{5.7}$$

Le score d'érosion normalisé est ensuite défini comme :

$$es_i^* = \frac{es_i}{\max\limits_{i=1,\ldots n} es_i} \tag{5.8}$$

La figure 5.4 illustre le score d'érosion et sa capacité à posséder des valeurs élevées au milieu des groupes de valeurs hautes ainsi qu'à lisser les données bruitées. Cet effet est un des bénéfices classiquement attendus des outils de morphologie mathématique. Nous proposons ci-dessous d'employer ces deux propriétés pour effectuer le regroupement des données.

Mattioli & Schmitt (1992) utilisent également des érosions successives dans le cadre de la granulométrie par érosion. Appliquée aux séries temporelles, celle-ci s'écrit :

$$\psi^j = \sum_{i=1}^{n} \epsilon_i^j$$

Le lien avec le score d'érosion défini dans l'éq. (5.8) apparaît ici clairement : la granulo-

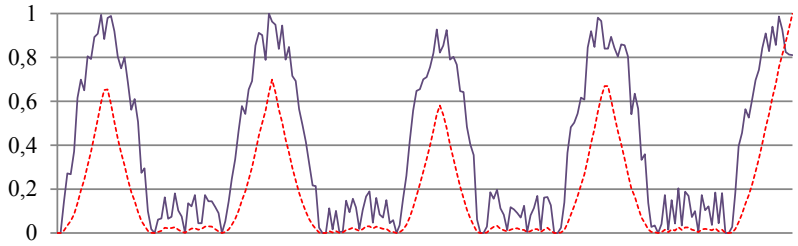


FIGURE 5.4 – Les données en trait plein et le score d'érosion en pointillés

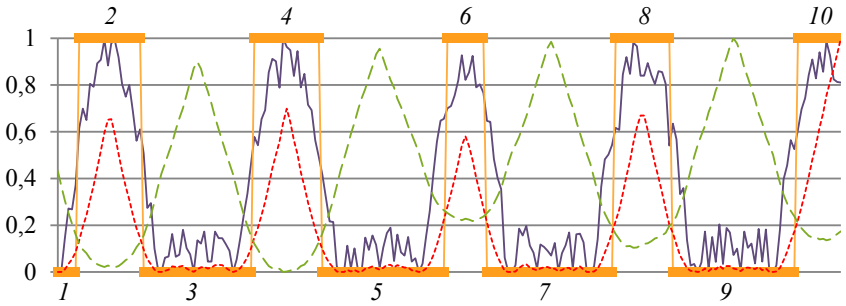


FIGURE 5.5 – En trait plein, les données X en entrée, en pointillés courts rouges, le score d'érosion es de X, en pointillés longs verts, le score d'érosion $\overline{es}$ de $\overline{X}$, au-dessus et en-dessous, les indices des groupes hauts et bas respectivement.

métrie par érosion agrège, pour un niveau d'érosion fixé, les valeurs obtenues pour chaque donnée, tandis que le score d'érosion agrège, pour une donnée fixée, les valeurs obtenues à chaque niveau d'érosion.

Clustering par score d'érosion Le score d'érosion en lui-même ne suffit pas à définir γ puisqu'il permet d'évaluer dans quelle mesure un point appartient à un groupe haut mais pas à un groupe bas. Pour ce faire, le score d'érosion $\overline{es}$ est calculé pour $\overline{X}$, le complémentaire de X, l'idée étant que les groupes hauts de $\overline{X}$ correspondent aux groupes bas de X. Comme les x_i sont supposés dans [0,1], les $\overline{x}_i$ de $\overline{X}$ définis comme $1 - x_i$ appartiennent également à [0,1]. À l'instar d'es_i basé sur z_i, $\overline{es}_i$ utilise $\overline{z}_i$ dont l'existence est garantie par la contrainte $\exists j$ tel que $x_j = 1$ (cf. éq. (5.1) p. 98).

La fonction de regroupement basée sur le score d'érosion est notée γ_{es} (es pour *erosion score*) et définie comme :

$$\gamma_{es}(x_i) = \begin{cases} H & \text{si } es_i^* > \overline{es}_i^* \\ L & \text{sinon} \end{cases} \tag{5.9}$$

La figure 5.5 illustre le résultat de cette méthode : les groupes sont correctement identifiés sans qu'aucun paramètre ne soit spécifié.

Par la suite, nous utilisons les notations suivantes : τ désigne le type d'un groupe dans $\{H, L\}$ et $G_k^\tau \in G$ le $k^{ème}$ groupe de type τ. g^τ représente le nombre de groupes de type τ et n^τ le nombre total de points contenus dans ces groupes. Comme tous les points appartiennent soit à un groupe haut soit à un groupe bas, $n = n^H + n^L$. Comme tous les

groupes sont soit hauts soit bas, $g = g^H + g^L$.

5.2.3 Variantes de regroupement

Nous détaillons ci-dessous deux variantes de regroupement γ_{BL} et γ_W que nous avons proposées, la première comme méthode de référence et la seconde basée sur la ligne de partage des eaux, une autre approche issue de la morphologie mathématique.

Méthode de référence γ_{BL} La première variante proposée γ_{BL} (BL pour *baseline*) est une méthode de référence à laquelle les autres variantes sont comparées. C'est une méthode de seuillage simple du même type que celles mentionnées au début de cette section p. 98. Elle est basée sur un seuil $t_v \in [0; 1]$ défini par l'utilisateur et considère les points dont la valeur est au-dessus de ce seuil comme haut et les autres comme bas :

$$\gamma_{BL}(x_i) = \begin{cases} H & \text{si } x_i > t_v \\ L & \text{sinon} \end{cases} \tag{5.10}$$

Nous utilisons de plus un taux t_m défini par l'utilisateur pour fusionner les groupes proches et limiter les effets du bruit. Plus spécifiquement, un groupe de taille inférieure à nt_m est fusionné avec les deux groupes de type opposé qui l'encadrent.

Méthode γ_W basée sur la ligne de partage des eaux La seconde variante, comme la méthode de référence, se base sur un seuil global pour séparer les données, mais le seuil ici n'est pas fourni par l'utilisateur mais déterminé à partir des données. L'objectif de cette méthode est de déterminer un seuil qui sépare les données hors des zones de valeurs où elles sont bruités. Dans le cas contraire en effet, la méthode de regroupement renvoie de petits groupes hauts et bas consécutifs car dans la zone de bruits les valeurs sont supérieures ou inférieures au seuil de manière non significative. Sur la figure 5.5 par exemple, les zones bruitées sont dans les valeurs [0 ; 0,2] et [0,8 ; 1], aussi la définition d'un seuil égal à 0,9 par exemple n'est pas pertinente car de nombreux groupes hauts seront identifiés au lieu des cinq correctement détectés sur la figure. A l'inverse, un seuil autour de 0,5 paraît adapté en ce cas car il ne coupe pas les zones bruitées de la courbe.

Principe de la ligne de partage des eaux Afin de déterminer automatiquement ce seuil, nous définissons une méthode inspirée de la Ligne de Partage des Eaux (LPE) utilisée en morphologie mathématique. Vincent & Soille (1991) proposent pour la décrire la métaphore de l'immersion des reliefs de la courbe, illustrée sur la figure 5.6. Sur cette figure, le niveau d'eau est symbolisé par la ligne mauve clair qui monte d'une image à la suivante. Lorsque des extrema locaux sont rencontrés, leur intersection avec le niveau d'eau est indiqué par un losange coloré et leur ordonnée par une ligne orange.

Ces extrema sont liés aux zones bruitées de la courbe. Le seuil cherché ne doit donc pas les contenir et même en être le plus éloigné possible. Le meilleur seuil t_W est donc

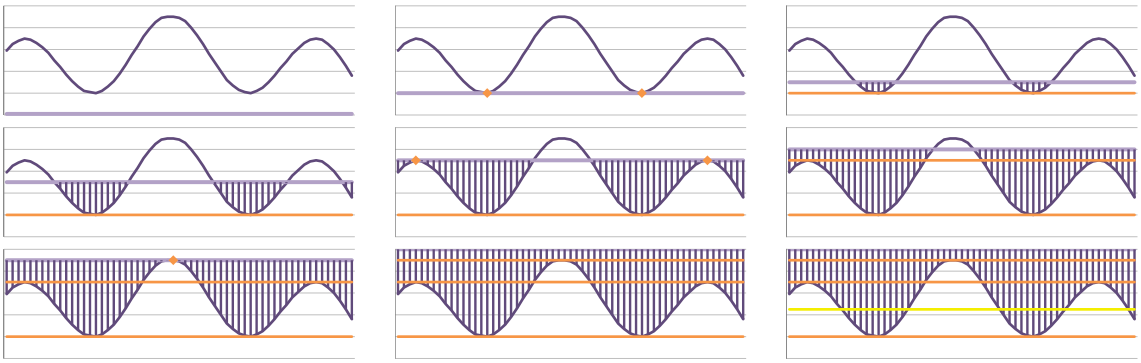


FIGURE 5.6 – Illustration de la méthode de Ligne de Partage des Eaux : de gauche à droite, de haut en bas, immersion progressive du relief défini par la courbe.

défini comme le milieu du plus grand intervalle entre deux niveaux successifs contenant un ou plusieurs extrema. Il est indiqué en jaune sur la dernière image de la figure 5.6.

Implémentation Deux étapes de pré-traitement sont effectuées afin de rendre la méthode plus efficace. La première permet d'atténuer le bruit de la la série par application d'une moyenne mobile de largeur w, qui est un paramètre de la méthode. Ce premier lissage réduit le nombre d'extrema locaux et donc de niveaux qui seront retenus par la suite.

La seconde consiste à ne garder qu'une valeur lorsque plusieurs sont égales consécutivement. Ainsi, les seuls extrema locaux à identifier sont tels que les deux points qui leur sont adjacents sont soit strictement supérieurs (dans le cas d'un minimum local), soit strictement inférieurs (dans le cas d'un maximum local).

L'ensemble de données obtenu après ces deux étapes de préparation est noté W. L'ensemble L des niveaux où un extrema est détecté est alors défini comme :

$$L = \{w_i \in W / (w_i > w_{i+1} \wedge w_i > w_{i-1}) \vee (w_i < w_{i+1} \wedge w_i < w_{i-1})\}$$

Le seuil t_W est alors calculé ainsi :

$$t_W = \tfrac{1}{2}(l_m + l_{m+1}) \text{ où } m = \underset{i=1\ldots|L|-1}{\arg\max}\ l_{i+1} - l_i \tag{5.11}$$

où les l_i sont les éléments de L classés par ordre croissant. La fonction de regroupement basée sur la ligne de partage des eaux (*watershed* en anglais) est donc :

$$\gamma_W(x_i) = \begin{cases} H & \text{si } x_i > t_W \\ L & \text{sinon} \end{cases} \tag{5.12}$$

Les expériences décrites plus tard dans la section 7.2 p. 139 montrent que les seuils globaux utilisés avec γ_{BL} et γ_W sont moins adaptés que les seuils locaux utilisés avec γ_{ES}.

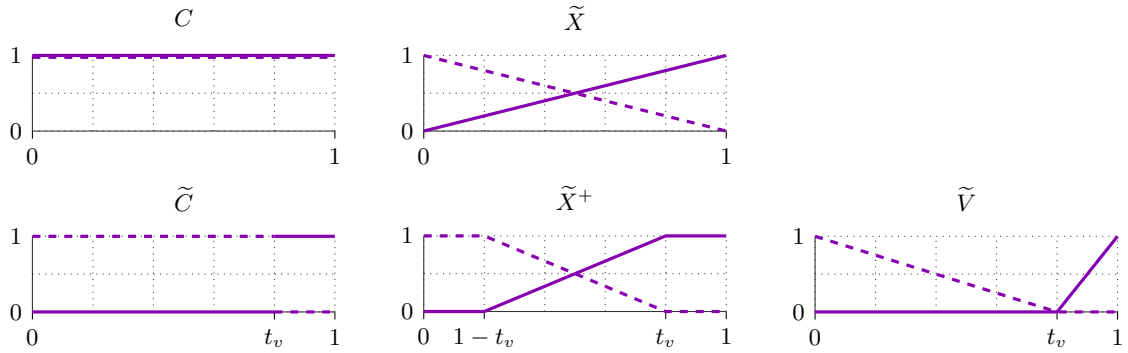


FIGURE 5.7 – Cardinalités pondérées pour la taille des groupes. En abscisse, la valeur d'un point, en ordonnée, le poids associé. En trait plein, la contribution d'un point H, en pointillés celle d'un point L.

5.3 Période et périodicité

La deuxième étape de la méthode DPE concerne le calcul de la période et de la périodicité à partir des groupes déterminés à l'étape précédente. Ce calcul est réalisée en trois étapes présentées dans les sous-sections suivantes : dans un premier temps, la taille des groupes est évaluée, puis leur régularité est calculée, enfin le degré de périodicité et la période candidate sont déterminés.

5.3.1 Taille des groupes

La taille des groupes H et L est simplement la cardinalité du groupe considéré. s_j^τ la taille du $j^{ème}$ groupe de type τ est donc calculée comme $s_j^\tau = \left|G_j^\tau\right|$.

Variantes La taille du groupe est ici calculée comme une cardinalité classique puisque chaque point compte pour 1 dans son calcul. Nous avons proposé d'examiner d'autres schémas de pondération qui n'attribuent pas le même poids aux points du groupe et qui calculent la taille du groupe comme leur somme pondérée.

Les différents schémas de pondération utilisés pour chaque point sont représentés sur la figure 5.7. Les poids des points des groupes H sont en traits pleins et ceux des groupes L en pointillés.

Les schémas C et $\widetilde{X}$ n'utilisent pas de paramètre. C est la cardinalité classique, i.e. tous les points du groupe comptent pour 1 dans sa taille. $\widetilde{X}$ est la fonction identité et considère qu'un point compte d'autant plus que sa valeur est élevée (resp. basse) pour un groupe H (resp. L).

Les schémas $\widetilde{C}$, $\widetilde{X}^+$ et $\widetilde{V}$ sont définis avec le seuil t_v et donc applicables pour la méthode γ_{BL}. Les tests réalisés avec ces trois schémas ont pour but de déterminer si l'utilisation d'un décompte pondéré pour la taille des groupes permet ou non de compenser la rigidité du seuil utilisé par la méthode.

$\widetilde{X}^+$ est du même type que $\widetilde{X}$, mais pondère légèrement plus les points des groupes

hauts et bas : ceux qui sont supérieurs (resp. inférieurs) à t_v (resp. $1 - t_v$) contribuent tous pour 1 à la taille des groupes H (resp. L). $\widetilde{V}$ est plus stricte car les points inférieurs (resp. supérieurs) à t_v ne comptent pas pour les groupes H (resp. L). De plus, elle n'est pas symétrique, ce qui peut biaiser la comparaison des tailles de groupes H et L telle que décrite dans la section 5.3.2.

Enfin, $\widetilde{C}$ peut sembler similaire à C puisque, pour la méthode γ_{BL}, les groupes H (resp. L) sont composés de points dont les valeurs sont supérieures (resp. inférieures) à t_v. En fait, la différence entre ces modalités vient de la technique de fusion appliquée avec γ_{BL} et détaillée dans la section 5.2.3 p. 103. En effet, lorsqu'un groupe contenant peu de points est fusionné avec les deux groupes adjacents de type opposé, les valeurs qu'il contient sont inchangées. Avec le schéma de pondération C tous les points des trois groupes contribuent à 1, mais avec $\widetilde{C}$ seuls ceux de même type sont pris en compte puisque ceux de type opposé, donc « de l'autre côté » de $x = t_v$, ont un poids nul.

Les tests menés sur l'ensemble des méthodes et des schémas sont détaillés dans la section 7.2 p. 139. Ils montrent que la cardinalité crisp C est la mieux adaptée pour la méthode DPE. Pour les modalités liées à γ_{BL}, la sensibilité au bruit liée à l'application dans un premier temps du seuil crisp n'est pas contrebalancée par le calcul pondéré de la taille qui lui est postérieur. La cardinalité classique est donc utilisée dans la suite de la thèse.

5.3.2 Régularité des groupes

L'étude de la régularité de l'occurrence d'un événement est couramment utilisée pour calculer la période d'une série temporelle : parmi les approches détaillées au chapitre 4, Durnerin (1999, p.116) calcule la régularité de l'espacement des pics de la séquence d'autocorrélation et Otunba et al. (2014) celle de motifs similaires dans une série symbolique.

Pour DPE, la régularité étudiée n'est pas celle de l'occurrence d'un événement mais celle de la taille des groupes hauts et bas. Nous proposons d'étudier cette régularité ρ au travers de la variabilité des tailles des groupes H et L. Cette variabilité est calculée par le coefficient de variation CV qui rapporte une mesure de dispersion d à la taille moyenne μ des groupes. En notant τ le type de groupe dans $\{H, L\}$, nous définissons :

$$\begin{aligned} \mu^\tau &= \frac{1}{g^\tau}\sum_{j=1}^{g^\tau} s_j^\tau \qquad & d^\tau &= \frac{1}{g^\tau}\sum_{j=1}^{g^\tau}\left|s_j^\tau - \mu^\tau\right| \\ CV^\tau &= \frac{d^\tau}{\mu^\tau} & \rho^\tau &= 1 - \min\left(CV^\tau, 1\right) \end{aligned} \tag{5.13}$$

Un certain nombre de choix ont été réalisés pour le calcul de ces variables : une moyenne pour la mesure de tendance centrale, une déviation absolue moyenne (DAM) pour la mesure de dispersion, un coefficient de variation CV pour celle de variabilité et son complément à 1 avec seuillage pour celle de régularité.

Chacun de ces choix ainsi que certaines de leurs variantes sont détaillés dans les para-

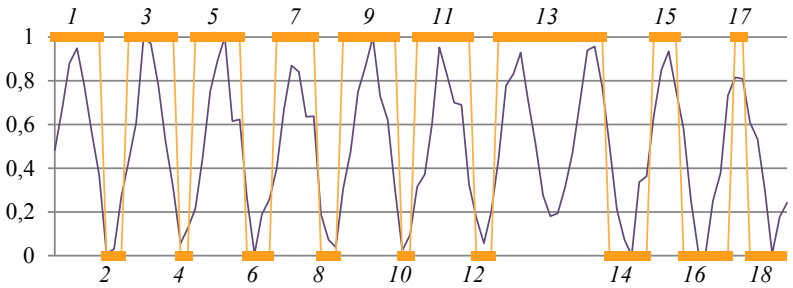


FIGURE 5.8 – Erreurs de classification ayant un impact sur la régularité

TABLEAU 5.1 – Combinaisons de moyenne et de médiane pour l'évaluation de la variabilité

Dispersion \ Taille	Moyenne	Médiane
Moyenne	Combinaison $\mu\mu$ $\mu = 1/n \sum s_i$ $d\mu = 1/n \sum \lvert\mu - s_i\rvert$	Combinaison $m\mu$ $m = \text{med}(s_i)$ $d\mu = 1/n \sum \lvert m - s_i\rvert$
Médiane	Combinaison μm $\mu = 1/n \sum s_i$ $dm = \text{med}(\lvert\mu - s_i\rvert)$	Combinaison mm $m = \text{med}(s_i)$ $dm = \text{med}(\lvert m - s_i\rvert)$

graphes suivants.

Tendance centrale Concernant la mesure de tendance centrale de la taille des groupes, la médiane, plus robuste que la moyenne, permet de ne pas prendre en compte les tailles extrêmes pouvant apparaître suite à une erreur de regroupement durant la première étape. Sur la figure 5.8 par exemple, les groupes 13 et 17 sont mal identifiés : le premier est trop grand et englobe deux groupes tandis que le second est trop petit et ne prend pas en compte le groupe dans toute sa largeur.

Pour autant, la médiane peut être trop robuste dans certains cas, comme détaillé dans la discussion plus approfondie donnée dans l'étude expérimentale de la section 7.2 p. 139. Ainsi, la moyenne est utilisée dans le reste de nos travaux, mais une poursuite des investigations à ce sujet constitue une perspective de nos travaux.

Dispersion Nous avons comparé expérimentalement (voir section 7.2.5 p. 149) l'écart-type à la DAM pour la mesure de dispersion des tailles de groupes. Les résultats montrent que l'écart-type est trop sensible au bruit et que l'utilisation de la DAM pour le calcul est préférable dans le contexte de DPE. Gorard (2005) présente également une étude détaillée illustrant les avantages de la DAM sur l'écart-type pour mesurer la dispersion.

De plus, la DAM mesure la moyenne des écarts en valeur absolue à la moyenne, mais d'autres variantes sont envisageables avec la médiane pour la mesure de tendance centrale et/ou de dispersion de la taille des groupes. Les différentes combinaisons testées sont résumées dans le tableau 5.1.

Les résultats des expériences menées avec ces différentes variantes, détaillés dans la sec-

tion 7.2.6 p. 150, montrent que l'utilité de la médiane par rapport à la moyenne n'est pas systématiquement vérifiée. La moyenne et la DAM sont donc retenues par la suite.

Variabilité L'utilisation du coefficient de variation qui rapporte une mesure de dispersion à une mesure de tendance centrale permet de définir une mesure générique de la variabilité adaptée à la taille des groupes : CV est élevé avec une dispersion de 1 pour des groupes de taille moyenne 5 mais faible pour une même dispersion et une taille moyenne de groupe égale à 100. Nous n'avons pas testé d'autres variantes pour ce coefficient qui donne de bons résultats en l'état.

Régularité A l'aide de CV, nous définissons ρ^τ pour mesurer la régularité de la taille des groupes. Nous contraignons ρ^τ dans [0,1] par seuillage à l'aide d'un min dans l'éq. (5.13) p. 106.

Nous avons étudié d'un point de vue théorique la borne supérieure de CV afin de normaliser ρ via une multiplication par un coefficient, i.e. $\rho^\tau = CV^\tau/\eta$. Les résultats, détaillés en annexe D p. 221, indiquent qu'avec des valeurs de x_i dans $[0, 1]$, $\max(CV^\tau) = 2$, et donc que $\eta = 2$ permet de normaliser ρ^τ. Or les cas où le coefficient de variation est supérieur à 1 correspondent à des tailles de groupes dont la dispersion est supérieure à la moyenne, i.e. des groupes de tailles très inégales qui ne peuvent représenter une série périodique. De plus, le facteur $\eta = 2$ « écrase » les valeurs de CV plus faibles qui sont a priori liées à des données périodiques, rendant leur analyse moins précise. La normalisation du coefficient de variation est donc réalisée par le minimum qui « coupe » les valeurs trop importantes sans écraser les valeurs plus faibles, également les plus intéressantes.

5.3.3 Degré de périodicité et période candidate

Une fois calculées la taille moyenne et la régularité des groupes, le degré de périodicité π et la période candidate p_c sont évalués par :

$$\pi = \frac{\rho^H + \rho^L}{2} \qquad\qquad p_c = \mu^H + \mu^L \tag{5.14}$$

Le degré de périodicité π est simplement la moyenne des régularités des groupes, ce qui correspond à l'hypothèse initiale liant la périodicité à l'alternance de groupes hauts et bas de tailles régulières.

La période candidate p_c est la somme des tailles moyennes des groupes H et des groupes L : si le signal est périodique et découpé en zones de valeurs hautes et en zones de valeurs basses, alors une zone haute et une zone basse définissent une période. C'est sous l'hypothèse de l'alternance de groupes H et L de tailles régulières, et donc lorsque π est suffisamment élevé, que la période candidate a un sens.

Concernant le calcul de la périodicité π, nous avons également testé l'opérateur min pour agréger les régularités ρ^H et ρ^L, au lieu de la moyenne utilisée dans l'éq. (5.14). Comme le min représente la conjonction logique, il aurait pu être plus pertinent car

conforme à l'interprétation "*les données sont périodiques si les tailles de groupes H ET celles des groupes L sont régulières*". Les expériences détaillées dans la section 7.2 p. 139 montrent qu'en pratique ce dernier est trop strict et renvoie des degrés de périodicité π trop faibles par rapport à ceux obtenus en utilisant la moyenne comme opérateur d'agrégation.

5.4 Rendu linguistique

La dernière étape, de rendu linguistique, permet de générer une phrase après l'obtention du degré de périodicité π et de la période candidate p_c. Comme p_c, la phrase n'a de sens que si le degré de périodicité π est suffisamment élevé.

5.4.1 Principe

L'étape de rendu linguistique s'inspire de la manière dont nous supposons que les humains expriment le temps. Trois caractéristiques sont prises en compte et détaillées dans les paragraphes suivants : le choix d'une unité pertinente, l'approximation de la période et l'adverbe de caractérisation.

Le choix de l'unité permet d'éviter les nombres trop petits ou trop grands pour exprimer la période candidate. Par exemple, si la période renvoyée est 168 et que les unités du jeu de données sont des heures, l'unité la plus appropriée ne sera pas l'heure mais la semaine car 168 heures = 1 semaine.

De la même manière, l'approximation de la période vise à éviter les mesure trop précises, rarement utilisées, sauf dans des contextes particuliers comme la compétition sportive ou la mesure scientifique par exemple. Dans le cas général qui est celui dans lequel se place notre étude, il semble par exemple plus approprié de parler de 45 minutes (voire de trois quarts d'heure) plutôt que de 44,87 minutes.

Cette approximation est ensuite caractérisée linguistiquement. En fonction de l'erreur commise, l'un des adverbes "exactement", "environ" ou "grossièrement" est utilisé.

5.4.2 Choix de l'unité

Afin de déterminer l'unité la plus adaptée pour représenter la période candidate, un ensemble d'unités ainsi que leurs rapports successifs sont définis par l'utilisateur, par exemple : secondes, minutes, heures, jours, semaines, avec les rapports : 86400 s = 1440 min = 24 h = 1 jour = 0,143 semaine. Le choix des unités utilisées permet à l'utilisateur de définir le domaine sur lequel les périodes sont exprimées.

Un intervalle de valeurs acceptables est ensuite défini, qui spécifie l'intervalle de valeurs pouvant être utilisé pour le choix d'une unité. Supposons que l'intervalle $[1, 60]$ soit choisi, alors l'unité sélectionnée sera telle que la période puisse s'exprimer dans cet intervalle. Si par exemple la période candidate est 3708 s, alors les conversions suivantes sont réalisées : 3708 s = 67,8 min = 1,03 heure = 0,04 jour = 0,006 semaine. L'unité "heure" est retenue car c'est la seule permettant d'exprimer la période dans $[1, 60]$. Lorsque plusieurs unités sont possibles, la plus petite représentation est choisie.

Il faut noter que l'intervalle retenu doit permettre d'exprimer tous les cas de figure. Avec l'exemple précédent, si l'intervalle est $[10, 60]$, aucune unité n'est possible car aucune des conversions ne renvoie de valeur dans cet intervalle. Pour éviter ce cas de figure, l'intervalle peut être déterminé de manière automatique en prenant 1 comme valeur minimale et le plus grand rapport de conversion entre deux unités successives comme valeur maximale. Avec les unités secondes, minutes, heures, jours, semaines, le plus grand rapport de conversion est 60 et l'intervalle $[1, 60]$ permet d'exprimer toutes les périodes entre 1 seconde et 60 semaines. Si une période hors de cet intervalle doit être exprimée, les unités et les rapports de conversion correspondants doivent être ajoutés.

5.4.3 Période approchée

Pour exprimer la période candidate de manière intuitive pour un utilisateur humain, les multiples de 5 et les valeurs entières sont privilégiées. Par exemple, il paraît plus naturel de dire 1 heure plutôt que 59 minutes ou 42 secondes plutôt que 41,79 secondes.

Afin de réaliser cette approximation, le pourcentage ϵ représentant l'erreur maximale tolérée pour arrondir la période candidate est défini par l'utilisateur. Plusieurs arrondis p_{ling} de la période candidate p_c sont testés jusqu'à ce que l'erreur relative commise soit inférieure à ϵ.

En notant $AM5\left(p_c\right)$ la fonction renvoyant l'arrondi de p_c au multiple de 5 le plus proche, $A\left(p_c, d\right)$ la fonction renvoyant l'arrondi de p_c à d décimales et $er\left(x, y\right) = \left|x - y\right| / y$ l'erreur relative de x par rapport à y, l'arrondi retenu p est défini par :

$$p = \begin{cases} AM5\left(p_c\right) & \text{si } er\left(AM5\left(p_c\right), p_c\right) < \epsilon \\ A\left(p_c, d\right) & \text{avec } d = \underset{d \geq 0}{\arg\min}\left(er\left(A\left(p_{c,} d\right), p_c\right) < \epsilon\right) \text{ sinon} \end{cases}$$

p est donc égale soit au multiple de 5 le plus proche, soit à l'arrondi avec le moins de décimales dont l'erreur relative par rapport à p_c est inférieure à ϵ.

Avec l'exemple précédent où la période 3708 s est exprimée comme 1,03 h avec l'unité retenue et en utilisant $\epsilon = 5\%$, alors $AM5\left(1, 03\right) = 1$ et $er\left(1, 03,\ 1\right) = 0, 03 < \epsilon$: l'erreur commise avec l'arrondi au multiple de 5 le plus proche est acceptable donc $p = 1$.

5.4.4 Sélection de l'adverbe

La dernière étape du rendu linguistique est la sélection de l'adverbe caractérisant l'approximation réalisée à l'étape précédente lors de l'arrondi de la période candidate. L'adverbe est l'une des modalités *Exactement*, *Environ* et *Grossièrement* de la variable linguistique *Précision* illustrée sur la figure 5.9.

L'adverbe retenu est celui pour lequel l'appartenance de l'erreur d'approximation $er(p_c, p)$ est la plus grande parmi les modalités de *Précision*. En notant M cette modalité et $m \in$

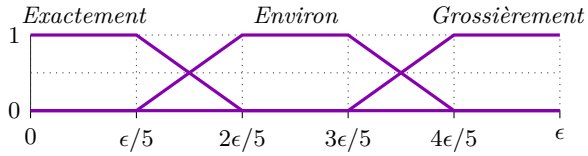


FIGURE 5.9 – Variable linguistique « Précision »

Précision les fonctions d'appartenance des modalités de la variable linguistique, on définit :

$$M = \underset{m \in Précision}{\arg\max}\ m\left(er\left(p_c, p\right)\right)$$

Dans l'exemple précédent, $p = 1$ et $p_c = 1,03$ donc $er\left(p_c, p\right) = 0,03$ et la modalité retenue est *Environ*.

Cet adverbe ne fait référence qu'à l'erreur d'approximation réalisée dans le rendu de la période calculée. Cependant, le degré de périodicité porte également une partie de l'information sur l'exactitude de la périodicité. Afin de restituer cette information à l'utilisateur, il est proposé d'une part de ne retourner la formulation linguistique que lorsque le degré de périodicité est suffisamment élevé, et d'autre part de l'intégrer entre parenthèses à la fin de la phrase, comme cela est l'usage pour les protoformes classiques. Ainsi, avec un degré de périodicité égal à 0,86 par exemple, la phrase renvoyée est « Environ toutes les heures, les valeurs sont élevées ($\pi = 0,86$) ».

5.5 Bilan

Ce chapitre a présenté la méthode DPE qui renvoie à partir d'une série temporelle un degré de périodicité π, une période candidate p_c et une phrase de la forme « M toutes les p unités, les valeurs sont élevées ». Les trois étapes nécessaires au calcul de ces résultats sont présentées : d'abord le regroupement des données en groupes de valeurs hautes et groupes de valeurs basses selon une approche basée sur la morphologie mathématique, puis le calcul de différentes statistiques sur la régularité de la taille des groupes permettant l'évaluation de π et p_c et enfin le rendu linguistique.

La méthode DPE répond donc à nos contraintes initiales : elle est interprétable car fondée sur un principe simple, ne nécessite ni paramètre ni hypothèse sur les données, calcule leur périodicité et retourne des résultats sous forme linguistique.

Nous montrons dans le prochain chapitre comment elle peut être implémentée en pratique de manière efficace.

Chapitre 6

Mise en œuvre de DPE

> Ce n'est pas grand-chose d'avoir des idées, le tout est de les appliquer, c'est-à-dire de penser par elles les dernières différences.
>
> —Alain, *Propos sur l'éducation*

Alors que le chapitre précédent propose une description détaillée des principes fonctionnels de DPE, celui-ci en décrit la mise en œuvre *efficace*.

La complexité de DPE dépendant principalement de la performance du calcul du score d'érosion, nous présentons dans une première section des approches incrémentales et par niveaux qui en permettent un calcul exact et rapide. Ces travaux ont fait l'objet de la publication (Moyse & Lesot, 2014).

Dans la deuxième section de ce chapitre nous présentons des algorithmes permettant d'implémenter ces approches efficaces et calculons leur complexité. Enfin, la dernière section introduit une présentation algorithmique de DPE permettant son exécution sur des flux de données.

6.1 Différentes approches pour le calcul du score d'érosion

Le score d'érosion présenté dans la section 5.2.2 p. 99 nécessite le calcul d'érosions successives jusqu'à érosion totale du jeu de données. Si les expériences détaillées au chapitre 7 montrent que le score d'érosion est efficace pour déterminer les groupes de valeurs hautes et ceux de valeurs basses, il est cependant coûteux en termes de calcul dans sa version naïve. Ainsi, nous proposons dans cette section trois autres approches destinées à l'optimiser.

Dans la section 6.1.1, nous rappelons les autres approches existantes en morphologie mathématique ainsi que leurs optimisations basées sur des érosions successives.

La section 6.1.1 décrit l'implémentation naïve du calcul du score d'érosion. Nous en proposons dans la section 6.1.3 un calcul *par niveaux*, basé sur la simplification des opérations répétitives de l'approche naïve par l'identification de valeurs particulières et de leurs répétitions.

La complexité du calcul naïf provenant également de la manière dont les données à traiter sont intégrées, nous proposons dans la section 6.1.4 p. 119 un traitement *incrémental* permettant de l'accélérer considérablement.

Enfin, nous détaillons dans la section 6.1.5 p. 122 une approche par niveaux *et* incrémentale, combinant les avantages des deux précédentes.

6.1.1 Optimisations de calculs en morphologie mathématique

L'application répétée d'érosion, de dilatation, d'ouverture ou de fermeture sont courantes en MM et leur implémentation naïve est souvent coûteuse (Vincent, 1992).

Différentes propositions ont été faites pour les améliorer. La première porte sur l'élément structurant dont la taille pénalise linéairement la complexité des méthodes naïves. Pecht (1985) propose différentes approches pour le calcul des opérations classiques de MM dont la complexité ne dépend pas de la taille de l'élément structurant. Pour DPE néanmoins, cette optimisation n'est pas pertinente car l'élément structurant est simple et ne varie pas (cf. éq. (5.4) p. 101).

D'autres optimisations reposent sur l'usage de représentations spécifiques, comme une file de pixels (Vincent & Dougherty, 1994), ou d'implémentations récursives efficaces (Chen & Haralick, 1995) pour le calcul des érosions successives d'une image 2D. Ces deux algorithmes, bien que très performants, ne sont pas optimisés pour les signaux 1D que nous traitons et ne sont pas non plus incrémentaux.

Certaines propriétés des opérateurs morphologiques sont également mises à profit pour accélérer le calcul des ouvertures et des fermetures (Van Droogenbroeck & Buckley, 2005; Dokládal & Dokládalová, 2011; Morard et al., 2012). Ces optimisations ne sont pas non plus appropriées dans notre cas puisqu'elles ne se concentrent pas sur des érosions mais sur des ouvertures et/ou n'en optimisent qu'une seule tandis que les méthodes que nous proposons ci-dessous les optimisent toutes jusqu'à érosion totale.

6.1.2 Méthode naïve

L'implémentation naïve du score d'érosion défini dans l'éq. (5.8) p. 101 est réalisée par le calcul des érosions successives de toutes les données jusqu'à leur érosion totale : tant qu'au moins une valeur n'est pas complètement érodée, une nouvelle érosion de l'ensemble du jeu de données est exécutée. L'algorithme naïf qui exploite cette représentation de base est détaillé dans la section 6.2.2 p. 124.

6.1.3 Méthode par niveaux

L'observation des érosions répétées d'un jeu de données permet de constater que les valeurs de deux érosions successives d'un point sont fréquemment égales. La figure 6.1 illustre l'érosion d'un jeu de données avec la valeur x_7 mise en valeur : sa valeur change trois fois à la $2^{ème}$, $3^{ème}$ et $5^{ème}$ érosion mais reste constante à la $1^{ère}$ et à la $4^{ème}$.

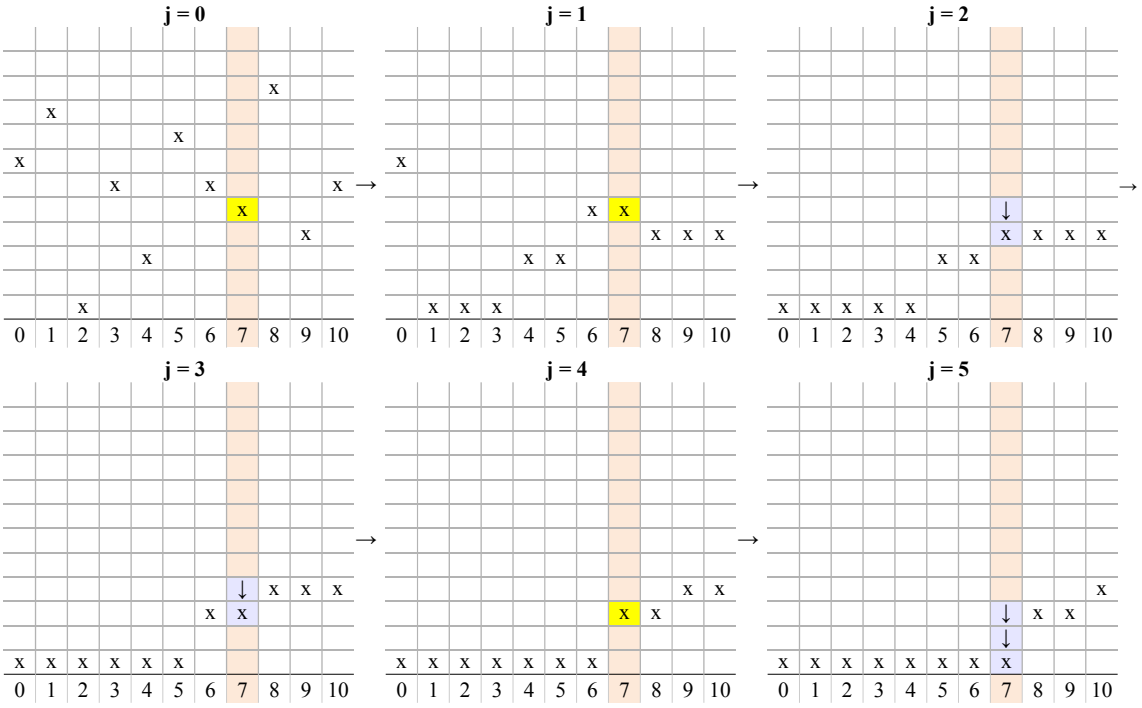


FIGURE 6.1 – Érosions successives d'un jeu de données. x_7 est mis en valeur, en jaune lorsque sa valeur ne change pas et en mauve lorsqu'elle change d'une érosion à la suivante.

Les érosions qui entraînent un changement de valeur sont indiquées en mauve sur la figure et appelées *érosions clés*. Si la $j^{ème}$ érosion est clé, la valeur prise est celle du point inférieur dans un voisinage de j points autour du point considéré. Cette valeur est appelée *valeur clé* et la valeur j est la *distance* entre le point considéré et la valeur clé.

Le principe de l'approche par niveaux est donc de détecter les valeurs clés autour du point considéré, de déterminer leur distance et de calculer le score d'érosion à l'aide de ces valeurs uniquement.

Une fois les notations définies dans le paragraphe ci-dessous, nous montrons dans celui d'après que cette intuition est correcte et introduisons le théorème de calcul du score d'érosion par niveaux. L'exemple de la figure 6.1 est aussi illustré sur la figure 6.2 p. 117 avec les variables du théorème.

Notations La définition des érosions successives ϵ_i^j (cf. éq. (5.5) p. 101) entraîne par récurrence :

$$\epsilon_i^j = \min\left(\epsilon_{i-1}^{j-1}, \epsilon_i^{j-1}, \epsilon_{i+1}^{j-1}\right) = \min\left(x_{i-j}, ..., x_i, ..., x_{i+j}\right) \tag{6.1}$$

ce qui signifie que j érosions avec un élément structurant unitaire reviennent à une érosion avec un élément structurant de taille j ou de manière équivalente que la $j^{ème}$ érosion de x_i est égale à la plus petite valeur autour de x_i dans un rayon de j points à droite et à gauche.

Afin de s'affranchir des particularités liées aux bords du jeu de données, nous posons :

$$\forall i \in \{-\infty, ..., 0\} \cup \{n+1, ..., +\infty\}\,,\; x_i = +\infty$$

L'opération utilisée étant un min, les valeurs égales à $+\infty$ n'ont pas d'influence.

De plus, à partir de l'éq. (6.1) et en vertu de la décroissance du min, on peut établir que les érosions successives sont décroissantes :

$$\forall j > k,\; \epsilon_i^j \leq \epsilon_i^k \tag{6.2}$$

Les différentes variables détaillées ci-dessous sont illustrées sur la figure 6.2. De manière formelle, nous nous intéressons aux érosions successives qui ne sont pas égales donc aux j tels que $\epsilon_i^j \neq \epsilon_i^{j-1}$. Comme les érosions sont décroissantes, ce sont les j tels que $\epsilon_i^j < \epsilon_i^{j-1}$. Notons D_i l'ensemble ordonné par valeurs croissantes défini comme :

$$D_i = \left\{ j \in \{1, \ldots, n\} \text{ tels que } \epsilon_i^j < \epsilon_i^{j-1} \right\} \tag{6.3}$$

ω_i son cardinal et les d_{il} ses éléments, également appelés distances, pour $l = 1...\omega_i$.

Nous ajoutons la valeur 0 à la position 0 de D_i, si bien que $d_{i0} = 0$ et $D_i = (d_{i0}, ..., d_{i\omega_i})$ est un sous-ensemble ordonné de $\{0, ..., n\}$ ne contenant que les étapes des érosions clés de x_i. Par définition, la $z_i^{ème}$ érosion de x_i est la première qui est égale à 0 et donc la dernière érosion clé, d'où $d_{i\omega_i} = z_i$. Avec l'exemple de la figure 6.2, les érosions successives de x_7 sont :

$$\underbrace{x_7 = \epsilon_7^0}_{\text{VC 0}} = \epsilon_7^1 > \underbrace{x_9 = \epsilon_7^2}_{\text{VC 1}} > \underbrace{x_4 = \epsilon_7^3}_{\text{VC 2}} = \epsilon_7^4 > \underbrace{x_2 = \epsilon_7^5}_{\text{VC 3}} = 0$$

et ses valeurs clés (notées VC) sont x_7, x_9, x_4 et x_2 correspondants aux érosions 0, 2, 3 et 5, d'où $D_7 = (0, 2, 3, 5)$.

Ainsi, les érosions j appartenant à D_i sont telles que $\epsilon_i^j < \epsilon_i^{j-1}$ et celles qui n'y appartiennent pas telles que $\epsilon_i^j = \epsilon_i^{j-1}$. Comme par définition $d_{i0} = 0$, $\epsilon_i^0 = x_i$ et D_i est trié par ordre croissant i.e. $d_{il} < d_{il+1}$, la relation suivante est vérifiée :

$$\begin{aligned} x_i = \epsilon_i^{d_{i0}} = \epsilon_i^{d_{i0}+1} = \cdots = \epsilon_i^{d_{i1}-1} > \epsilon_i^{d_{i1}} = \epsilon_i^{d_{i1}+1} = \cdots = \epsilon_i^{d_{i2}-1} > \epsilon_i^{d_{i2}} \ldots \\ \ldots \epsilon_i^{d_{il}} = \epsilon_i^{d_{il}+1} = \cdots = \epsilon_i^{d_{i,l+1}-1} > \epsilon_i^{d_{i,l+1}} \ldots \epsilon_i^{d_{i,\omega_i}-1} > \epsilon_i^{d_{i\omega_i}} = 0 \end{aligned} \tag{6.4}$$

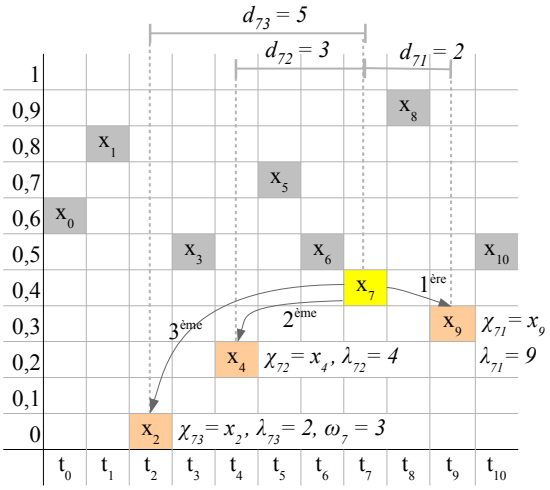
Afin de simplifier l'écriture, nous introduisons :

$$\chi_{il} = \epsilon_i^{d_{il}} \text{ et } \lambda_{il} \text{ tel que } \chi_{il} = x_{\lambda_{il}} \tag{6.5}$$

Avec l'exemple précédent, $\chi_7 = \{x_7, x_9, x_4, x_2\}$ et $\lambda_7 = \{7, 9, 4, 2\}$.

On peut alors interpréter d_{il} comme le nombre de points entre x_i et χ_{il}. Avec les indices λ_{il}, on obtient :

$$d_{il} = |\lambda_{il} - i| \tag{6.6}$$



$$D_7 = (\ 0\ \ 2\ \ 3\ \ 5\ )$$
$$j\ \ =\ \ 0\ \ 1\ \ 2\ \ 3\ \ 4\ \ 5$$
$$\epsilon_7^j = (\ x_7\ x_7\ x_9\ x_4\ x_4\ x_2\ )$$
$$\chi_7 = (\ x_7\ x_9\ x_4\ x_2\ )$$
$$\lambda_7 = (\ 7\ \ 9\ \ 4\ \ 2\ )$$

Correspondances entre les valeurs de j, D, ϵ_i^j, χ et λ pour x_7 dans l'exemple ci-contre.

FIGURE 6.2 – Illustration des valeurs clés et des vecteurs D, χ et λ associés à x_7

Les valeurs ci-dessus peuvent donc être interprétées ainsi : d_{il} est la distance en nombre de points entre x_i et sa $l^{ème}$ valeur clé, χ_{il} est cette valeur et λ_{il} son index.

Calcul du score d'érosion par niveaux Le théorème suivant donne une expression du score d'érosion par niveaux, à l'aide des variables introduites ci-dessus.

Théorème 1. *Calcul du score d'érosion par niveaux*

$$es_i = \sum_{l=0}^{\omega_i - 1} (d_{i,l+1} - d_{il})\, \chi_{il} = \sum_{l=0}^{\omega_i - 1} (|\lambda_{i,l+1} - i| - |\lambda_{il} - i|) x_{\lambda_{il}} \tag{6.7}$$

L'expression la plus à droite montre que le score d'érosion peut n'être calculé qu'en utilisant les indices des valeurs clés λ_{il}.

Démonstration. La démonstration découle directement des définitions de χ, λ, d et de la définition du score d'érosion donnée par l'éq. (5.8) p. 101 :

$$\begin{aligned} es_i &= \underbrace{x_i^0 + \cdots + x_i^{d_{i1}-1}}_{(d_{i1}-d_{i0})\chi_{i0}} + \underbrace{x_i^{d_{i1}} + \cdots + x_i^{d_{i2}-1}}_{(d_{i2}-d_{i1})\chi_{i1}} + \ldots + \underbrace{x_i^{d_{i,\omega_i-1}} + \cdots + x_i^{d_{i,\omega_i}-1}}_{(d_{i\omega_i}-d_{i,\omega_i-1})\chi_{i,\omega_i-1}} + \underbrace{x_i^{d_{i\omega_i}}}_{0} \\ &= \sum_{l=0}^{\omega_i-1} (d_{i,l+1} - d_{il})\, \chi_{il} \\ &= \sum_{l=0}^{\omega_i-1} (|\lambda_{i,l+1} - i| - |\lambda_{il} - i|) x_{\lambda_{il}} \text{ par l'éq. (6.6)} \end{aligned}$$

□

Représentations de la matrice λ D'autre part, la représentation par niveaux donne une vision intéressante du score d'érosion. En effet, l'ensemble des λ_{il} peut se représenter sous forme matricielle, où chaque ligne i représente la *chaîne des indices des ω_i valeurs clés* de x_i.

$$\lambda = \begin{pmatrix} 0 & 2 & . & . \\ 1 & 2 & . & . \\ 2 & . & . & . \\ 3 & 2 & . & . \\ 4 & 2 & . & . \\ 5 & 4 & 2 & . \\ 6 & 7 & 4 & 2 \\ 7 & 9 & 4 & 2 \\ 8 & 9 & 4 & 2 \\ 9 & 4 & 2 & . \\ 10 & 9 & 4 & 2 \end{pmatrix} \qquad \omega = \begin{pmatrix} 2 \\ 2 \\ 1 \\ 2 \\ 2 \\ 3 \\ 4 \\ 4 \\ 4 \\ 3 \\ 4 \end{pmatrix}$$

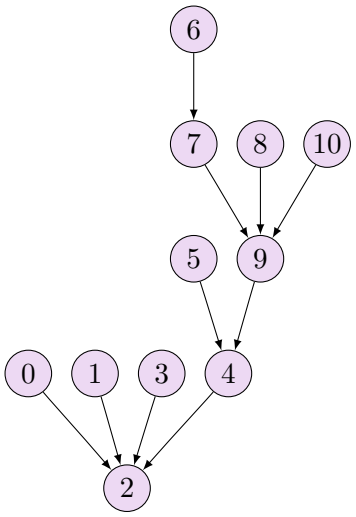


FIGURE 6.3 – Matrice λ des données illustrées sur la figure 6.2 (à gauche) représentée sous forme d'arbre (à droite)

A titre d'illustration, la matrice λ illustrée sur la figure 6.3 correspond aux λ_{ij} et ω du jeu de données illustré sur la figure 6.2, où les indices situés dans des colonnes supérieures à ω_i sont représentés par des points.

Comme une partie des données de la matrice λ n'est pas porteuse d'information, nous en avons étudié une représentation arborescente, plus parcimonieuse a priori. Un exemple d'arbre généré à partir de la matrice λ ci-dessus est illustré dans la partie droite de la figure 6.3.

L'intérêt à première vue de ce type de représentation est que seul le successeur d'un point doit être stocké pour mener à bien le calcul du score d'érosion, au lieu des chaînes de valeurs de clés de chacun des x_i. Malheureusement, la connaissance du seul successeur d'un point n'est pas suffisante pour retrouver la chaîne des valeurs clés. En effet, l'arbre indique que la chaîne partant de x_6 est x_7, x_9, x_4 et x_2 or x_9 n'est pas présente dans la chaîne de x_6. Cela est dû au fait que l'arbre ne tient pas compte de la position du point dans la chaîne : si x_9 est effectivement la seconde valeur clé de la chaîne de x_7, elle ne l'est pas pour x_6 car x_4 est plus proche de x_6 et $x_4 < x_9$.

Nous avons donc proposé de définir les deux fonctions auxiliaires $l(x_i)$ et $r(x_i)$ qui désignent respectivement l'indice du point inférieur le plus proche à gauche et celui du point inférieur le plus proche à droite :

$$l(x_i) = \arg\max_{j<i} x_j < x \qquad r(x_i) = \arg\min_{j>i} x_j < x_i \tag{6.8}$$

Comme vérifié expérimentalement, les scores d'érosion peuvent également être calculés à l'aide de l et r. De plus, ces fonctions permettent un stockage réduit puisqu'elles ne consomment que $2n$ en mémoire contre n^2 dans le pire des cas pour λ, comme discuté dans la section 7.3 p. 151.

Δ	0	1	2	3	4	5	6	7	8	9	10
es0	2	0	0	0	0	0	0	0	0	0	0
es1	0	1	0	0	0	0	0	0	0	0	0
es2	0	0	0	0	0	0	0	0	0	0	0
es3	0	0	0	1	0	0	0	0	0	0	0
es4	0	0	0	0	2	0	0	0	0	0	0
es5	0	0	0	0	2	1	0	0	0	0	0
es6	0	0	0	0	2	0	1	1	0	0	0
es7	0	0	0	0	2	0	0	2	0	1	0
es8	0	0	0	0	2	0	0	0	1	3	0
es9	0	0	0	0	2	0	0	0	0	5	0
es10	0	0	0	0	2	0	0	0	0	5	1

FIGURE 6.4 – Matrice Δ pour les données de la figure 6.2 p. 117

Matrice Δ Le théorème 1 p. 117 permet également d'introduire une notation matricielle pour le calcul par niveaux. Notons $\delta_{il} = d_{i,l+1} - d_{il}$ et Δ la matrice $n \times n$ composée des éléments Δ_{il} définis par :

$$\forall i = 1, ..., n, \forall j = 1, ..., n, \Delta_{ij} = \begin{cases} \delta_{il} & \text{si } j \in \lambda_i \\ 0 & \text{sinon} \end{cases}$$

où $\lambda_i = \{\lambda_{il}, l = 0...\omega_i\}$. Le score d'érosion est donné par :

$$ES = \Delta X^T$$

où X^T est la transposée de X et $X = (x_1, ..., x_n)$ et $ES = (es_1, ..., es_n)$ sont les vecteurs $1 \times n$ des données et des scores d'érosion.

La matrice Δ correspondant aux données de la figure 6.2 p. 117 est illustrée sur la figure 6.4. Sur cette figure, on voit par exemple que $es_0 = 2x_0$ et l'on retrouve $es_7 = 2x_4 + 2x_7 + x_9$.

Cette matrice a de plus certaines propriétés intéressantes. Par exemple, pour tout $i = 1, ..., n$, $\sum_j \Delta_{ij} = z_i$, ce qui signifie que la somme des éléments de la ligne i donne la distance de x_i au 0 le plus proche. D'autre part, si $x_i = 0$, alors $\forall j$, $\Delta_{ij} = \Delta_{ji} = 0$. Au contraire, si $x_i > 0$, alors $\Delta_{ii} > 0$. Une étude plus approfondie de cette matrice est proposée dans les perspectives.

6.1.4 Méthode incrémentale

Nous proposons dans ce paragraphe un autre type d'optimisation, basé sur une intégration incrémentale des données. En ce cas, le jeu de données considéré est vide au temps t_0 puis reçoit la première donnée au temps t_1, puis la seconde en t_2 et ainsi de suite tout en mettant à jour à chaque nouvelle donnée reçue les score d'érosions déjà calculés.

Comme formalisé plus bas, l'approche incrémentale permet de n'avoir à rechercher les plus proches valeurs inférieures qu'à gauche du dernier point intégré, ce qui réduit considérablement la complexité de l'algorithme. La méthode de Dokládal & Dokládalová (2011) présentée dans la section 6.1.1 p. 114 tire également parti de ce constat.

L'intérêt de l'approche incrémentale est d'une part sa rapidité, illustrée par les expériences décrites dans la section 7.3 p. 151, et d'autre part sa capacité à utiliser DPE sur des données reçues en flux, comme détaillé dans la section 6.3 p. 127.

Les notations utilisées sont présentées dans un premier temps, suivies des théorèmes de mise à jour des érosions et des scores d'érosion existants.

Notations Dans le cadre d'un flux de données, x_{n+1} représente la nouvelle donnée reçue au temps $t+1$. On note :

$$x_i(t) = \begin{cases} 0 & \text{si } i = 0 \\ x_i & \text{si } i \in \{1, ..., n\} \\ +\infty & \text{sinon} \end{cases} \tag{6.9}$$

La valeur $x_0 = 0$ est ajoutée au début de la série afin de s'assurer qu'à tout moment il existe au moins une valeur nulle dans les données, conformément à la contrainte exprimée dans l'éq. (5.1) p. 98. De plus, lorsque le contexte ne présente pas d'ambiguïté, x_i est utilisé pour représenter $x_i(t)$.

Au temps t, la $j^{ème}$ érosion de x_i est notée $\epsilon_i^j(t)$ et son score d'érosion $es_i(t)$.

Mise à jour incrémentale des érosions L'objectif de la méthode incrémentale est le calcul du nouveau score d'érosion $es_i(t+1)$ à partir des scores d'érosion existants $es_i(t)$ et de la nouvelle valeur x_{n+1}. Pour ce faire, deux théorèmes sont établis ci-dessous. Le premier concerne la mise à jour des érosions ϵ_i^j selon qu'ils sont influencés ou non par x_{n+1} et le second donne la valeur des scores d'érosion sur la base des érosions mises à jour.

Théorème 2. *Mise à jour des érosions successives à l'arrivée de x_{n+1}*
En notant $q = l(x_{n+1})$ et $m = (n+1+q)/2$, on a :

$$\epsilon_i^j(t+1) = \begin{cases} \epsilon_i^j(t) & \text{si } i \leq m & (6.10) \\ \epsilon_i^j(t) & \text{si } i > m \text{ et } j < n+1-i & (6.11) \\ x_{n+1} & \text{si } i > m \text{ et } n+1-i \leq j < i-q & (6.12) \\ \epsilon_q^{j-(i-q)} & \text{si } i > m \text{ et } j \geq i-q & (6.13) \end{cases}$$

Principe de preuve Ce paragraphe donne une idée intuitive de la démonstration de ce théorème, dont la preuve détaillée est présentée en annexe F p. 227. La figure 6.5 illustre les variables utilisées sur un jeu de données différent de l'exemple précédent.

Le théorème indique que si $x_{n+1} \geq x_n$ alors la première valeur inférieure à gauche de x_{n+1} est x_n, soit $q = n$, donc $m = n+1/2$ si bien que $\epsilon_i^j(t+1) = \epsilon_i^j(t)$ pour tout $i = 1...n$ par l'éq. (6.10), donc les érosions existantes sont inchangées.

Dans le cas contraire, la première valeur inférieure ou égale à x_{n+1} sur la gauche est recherchée et notée x_q. Puisque les valeurs sont dans [0,1] et que $x_0 = 0$, cette valeur

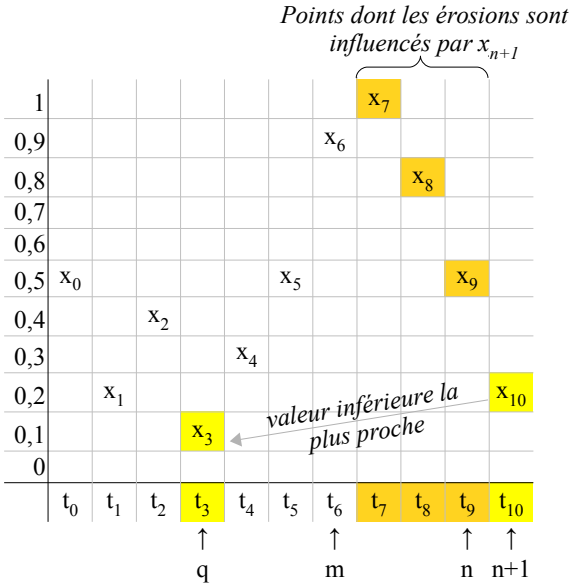


FIGURE 6.5 – Impact du nouveau point sur les érosions précédentes

existe toujours. Le théorème dit que les seules valeurs dont l'érosion est impactée par l'arrivée de x_{n+1} sont celles qui sont après l'indice m défini comme le milieu de l'intervalle entre x_q et x_{n+1} (cf. éq. (6.10)). Les premières érosions de ces valeurs sont inchangées (cf. éq. (6.11)), les érosions suivantes sont égales à x_{n+1} (cf. éq. (6.12)), et les dernières érosions jusqu'à érosion totale sont égales à celles de x_q (cf. éq. (6.13)).

Mise à jour incrémentale du score d'érosion Le théorème suivant détaille la mise à jour incrémentale du score d'érosion :

Théorème 3. *Calcul incrémental de* $es_i(t+1)$
En notant $q = l(x_{n+1})$ *et* $m = (n+1+q)/2$, *on a :*

$$es_i(t+1) = \begin{cases} es_i(t) & \text{si } i \leq m \\ \sum_{j=0}^{n-i} \epsilon_i^j(t) + 2(i-m)\, x_{n+1} + es_q(t) & \text{sinon} \end{cases} \tag{6.14}$$

Démonstration. Comme d'après le théorème 2 pour tout $i \leq m$, $\epsilon_i^j(t+1) = \epsilon_i^j(t)$, on a :

$$\forall i \leq m,\ es_i(t+1) = \sum_{j=0}^{z_i} \epsilon_i^j(t+1) = \sum_{j=0}^{z_i} \epsilon_i^j(t) = es_i(t) \tag{6.15}$$

Pour $i > m$, on décompose la somme qui définit es_i selon la valeur de j :

$$j \leq n-i : \sum_{j=0}^{n-i} \epsilon_i^j (t+1) = \sum_{j=0}^{n-i} \epsilon_i^j (t)$$

$$n-i < j < i-q : \sum_{j=n+1-i}^{i-q-1} \epsilon_i^j (t+1) = \sum_{j=n+1-i}^{i-q-1} x_{n+1} = 2(i-m) x_{n+1}$$

$$j \geq i-q : \sum_{j=i-q}^{z_i} \epsilon_i^j (t+1) = \sum_{j=i-q}^{z_i} \epsilon_q^{j-(i-q)} (t) = \sum_{\gamma=0}^{z_q} \epsilon_q^{\gamma} (t) = es_q (t)$$

D'où :

$$es_i (t+1) = \sum_{j=0}^{n-i} \epsilon_i^j (t+1) + 2(i-m) x_{n+1} + es_q (t)$$

□

6.1.5 Méthode incrémentale par niveaux

Nous proposons de combiner les deux approches précédentes pour obtenir un mode de calcul rapide *et* incrémental. La méthode incrémentale par niveaux fonctionne de manière similaire à la méthode incrémentale simple : sa première étape consiste en la mise à jour de la structure des λ issue de la méthode par niveaux et sa seconde étape est dédiée à la mise à jour des scores d'érosion.

Mise à jour incrémentale de λ_{il} La mise à jour incrémentale de λ_{il} est le pendant pour les érosions clés *uniquement* de la mise à jour incrémentale des ϵ_i^j pour toutes les érosions.

Théorème 4. *Mise à jour incrémentale de λ_{il}*

En notant $q = l(x_{n+1})$, $m = (n+1+q)/2$, et k_i tel que $d_{i,k_i-1}(t) < n+1-i \leq d_{ik_i}(t)$ avec $k_{n+1} = 0$, on a :

$$\forall i, \forall l, \ \lambda_{il} (t+1) = \begin{cases} \lambda_{il} (t) & \text{si } i \leq m \quad (6.16) \\ \lambda_{il} (t) & \text{si } i > m \text{ et } l < k_i \quad (6.17) \\ n+1 & \text{si } i > m \text{ et } l = k_i \quad (6.18) \\ \lambda_{q,l-k_i-1} (t) & \text{si } i > m \text{ et } l > k_i \quad (6.19) \end{cases}$$

$$\forall i, \ \omega_i (t+1) = \begin{cases} \omega_i (t) & \text{si } i \leq m \quad (6.20) \\ k_i + \omega_q (t) & \text{si } i > m \quad (6.21) \end{cases}$$

Principe de preuve Les équations démontrées pour le calcul en incrémental s'appliquent pour l'ensemble des données d'indice i et l'ensemble des étapes d'érosion j. Pour la mise à jour en incrémental par niveaux, seules les érosions clés sont considérées. Ce

théorème est donc démontré de la même manière que le théorème du calcul en incrémental, en n'utilisant plus les étapes d'érosion j mais les indices l dans l'ensemble croissant D_i des érosions clés.

Concernant la valeur k_i introduite dans ce théorème, elle représente l'érosion clé l correspondant à la première étape d'érosion j faisant intervenir x_{n+1}, dont on a montré avec le théorème incrémental qu'elle était comprise entre $n+1-i$ et $i-q$ (cf. éq. (6.11) p. 120).

La démonstration complète de ce théorème est donnée en annexe F p. 227.

Mise à jour incrémentale par niveaux du score d'érosion Le théorème suivant montre comment calculer le score d'érosion dans le cadre incrémental par niveau.

Théorème 5. *Mise à jour incrémentale par niveaux de es_i*

En notant $q = l(x_{n+1})$, $m = (n+1+q)/2$, k_i tel que $d_{i,k_i-1}(t) < n+1-i \leq d_{ik_i}(t)$ avec $k_{n+1} = 0$ et p_i défini pour $i > m$ tel que $\lambda_{ip_i}(t) = q$, on a :

$$\forall i,\ es_i(t+1) = \begin{cases} es_i(t) & \text{si } i \leq m \\ \chi_{i,k_i-1}(t)(n+1-i-d_{ik_i}(t)) + 2x_{n+1}(i-m) & \text{si } m < i < n+1 \\ \quad -\sum_{j=k_i}^{p_i-1} \chi_{ij}(t)(d_{i,j+1}(t) - d_{ij}(t)) + es_i(t) & \\ 2x_{n+1}(n+1-m) + es_q(t) & \text{si } i = n+1 \end{cases}$$

Les notations d et χ sont utilisées pour des raisons de simplicité d'écriture, mais l'expression de $es_i(t+1)$ peut être réalisée avec λ uniquement en utilisant $d_{il} = |\lambda_{il} - i|$ et $\chi_{il} = x_{\lambda_{il}}$.

Principe de preuve Dans ce théorème, une nouvelle variable p_i est utilisée, qui indique le numéro d'érosion clé valant x_q. Cette dernière permet d'éviter le calcul du terme $\sum_{j=0}^{n-i} \epsilon_i^j(t)$ dans l'éq. (6.14) p. 121.

D'autre part, la mise à jour du du score d'érosion repose ici sur le fait que les valeurs clés avant k_i et après p_i sont les mêmes en t et en $t+1$. Donc les seules valeurs clés impactant la mise à jour de es_i sont situées entre k_i et p_i.

La démonstration complète du théorème est donnée en annexe F p. 227.

6.2 Implémentations de DPE

Nous présentons dans cette section les implémentations des méthodes de calcul du score d'érosion présentées dans la section précédente. Leurs performances sont comparées au travers de l'étude expérimentale présentée dans la section 7.3 p. 151.

Le cadre général de leur fonctionnement est décrit dans la section 6.2.1 puis chacune des implémentations, naïve, par niveaux, incrémentale et incrémentale par niveaux, est décrite avec sa complexité dans les sous-sections 6.2.2 p. 124 à 6.2.5 p. 127.

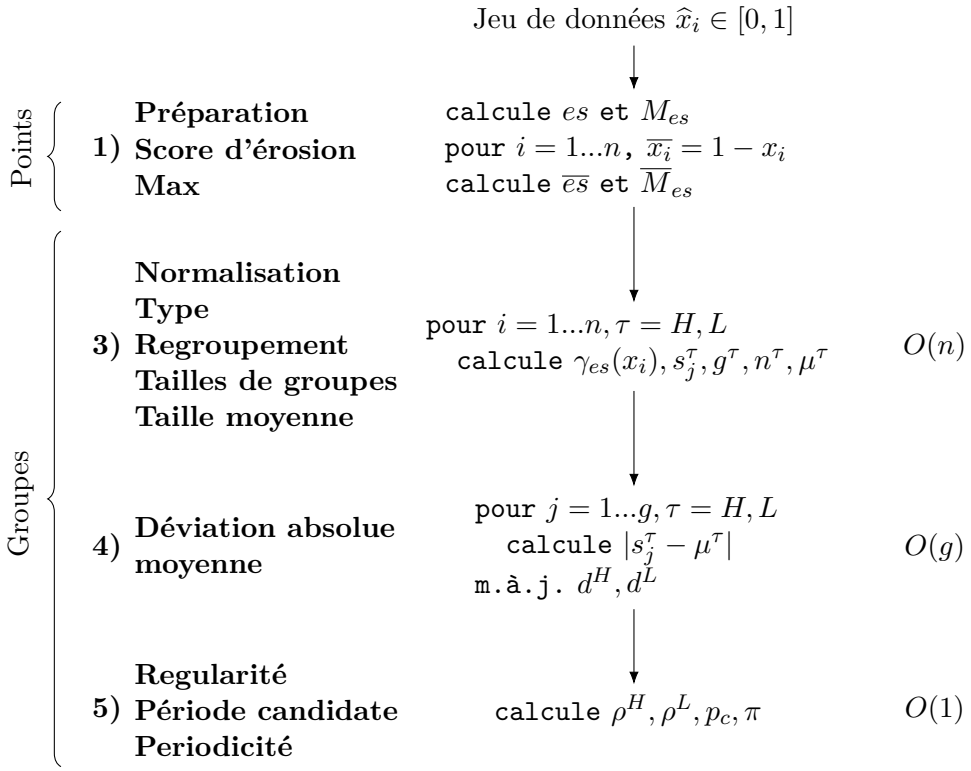


FIGURE 6.6 – DPE en traitement par lot avec un calcul de score d'érosion

6.2.1 Cadre général des implémentations de DPE

Les implémentations de DPE présentées ici supposent des données dans $[0, 1]$, conformément à la contrainte de l'éq. (5.1) p. 98. Comme illustré sur la figure 6.6, chaque méthode calcule dans un premier temps le score d'érosion es avec M_{es} la plus grande valeur de es, puis $\overline{X}$, $\overline{es}$ et $\overline{M}_{es}$. Enfin, les résultats obtenus divisés par M_{es} et $\overline{M}_{es}$ sont utilisés pour déterminer les groupes H et L et enfin la périodicité et la période de la série.

Les paragraphes suivants détaillent le calcul du score d'érosion ainsi que sa complexité en fonction de l'approche retenue. Le calcul des valeurs M_{es} et $\overline{M}_{es}$ est réalisé de manière identique d'une méthode à l'autre par comparaison du max courant avec la nouvelle valeur d'érosion calculée et n'est pas détaillé dans les algorithmes pour des raisons de lisibilité.

6.2.2 Algorithme naïf

L'algorithme 6.1 décrit le calcul naïf du score d'érosion. Il effectue autant de passes sur le jeu de données qu'il est nécessaire pour son érosion complète et ajoute à chaque itération la valeur actuelle des points au score d'érosion.

Puisque l'algorithme s'arrête lorsque toutes les valeurs sont totalement érodées, le nombre de passes sur l'ensemble des données dépend de celle qui sera la dernière à être érodée, soit celle dont la distance z_i au zéro le plus proche est la plus élevée. La complexité de cette approche est donc $O(n \times \max(z_i))$.

Algorithme 6.1 Algorithme naïf pour le score d'érosion

```
Tant que notEroded
    notEroded ← false
    Pour i = 1...n
        x_i ← min(x_{i-1}, x_i, x_{i+1})      # érosion du point
        es_i ← es_i + x_i                      # m.à.j du score d'érosion
        si x_i > 0 notEroded ← true            # boucle tant qu'une valeur est > 0
```

Algorithme 6.2 Algorithme par niveaux pour le score d'érosion

```
Pour i = 1...n
    j ← 1
    es_i ← x_i
    Tant que x_{i-j} > 0 et x_{i+j} > 0      # recherche du zéro le plus proche
                                             # z_i itérations
        es_i ← es_i + min(x_{i-j}, ..., x_{i+j})  # ajout de la valeur la plus faible
                                                  # dans un voisinage de j
        j ← j + 1
```

Le pire des cas correspond à un jeu de données dans lequel une seule valeur 0 est située en x_1 ou en x_n. En ce cas, le nombre d'érosions nécessaires à l'érosion totale du point à l'autre extrémité a pour valeur $z_i = n-1$, puisqu'il est séparé de $n-1$ points de la valeur 0 et le nombre d'itérations à réaliser est $n\,(n-1)$ soit $O\,(n^2)$.

6.2.3 Algorithme par niveaux

L'algorithme 6.2 décrit le calcul *par niveaux* du score d'érosion. Pour chaque point du jeu de données, le zéro le plus proche est recherché en z_i itérations. La mise à jour du score d'érosion utilise le résultat de l'éq. (6.1) p. 115 et ajoute à chaque itération la valeur la plus faible sur un voisinage de j points autour de x_i.

L'algorithme ne maintient pas la structure λ pour des raisons de simplicité, bien qu'elle soit présente dans les équations du théorème 1 p. 117 pour le calcul par niveaux. En effet, cette structure est nécessaire à la formalisation du théorème mais peut être omise dans l'algorithme. Elle est en revanche calculée dans le calcul incrémental par niveaux détaillé dans la section 6.2.5 p. 127.

La boucle interne est donc réalisée z_i fois pour chacun des n points, d'où une complexité de $O\,(\sum z_i)$. Ainsi, la complexité de cette méthode est inférieure à la complexité $O\,(n \times \max z_i)$ de l'implémentation naïve. Dans le pire des cas toutefois, la valeur nulle est située à l'une des extrémité du jeu de données ce qui implique que chaque point est situé à une distance $i-1$ de cette valeur nulle, autrement dit que $z_i = i-1$ et donc $\sum z_i = (n-1)(n-2)/2$ dans les deux cas, soit une complexité quadratique comme dans le cas de la méthode naïve. Ce cas est néanmoins rare, ce qui permet à la méthode par niveaux d'être plus rapide que celle naïve, comme le confirment l'étude de performance de la section 7.3 p. 151.

Algorithme 6.3 Algorithme incrémental pour le score d'érosion

```
Pour i = 1...n
    Si x_i > x_{i-1}                                  # en ce cas, q = i − 1
        es_i = es_{i-1} + x_i                         # calcul. rapide d'es_i
    Sinon
        j ← 2
        Tant que x_{i-j} > x_i                        # recherche de q
            j ← j + 1                                 # O(log(n))
        m ← (n + 1 + q)/2
        Pour i' = m + 1...n                           # O(log(n)/2) itérations
            Pour j = 0...n − i'                       # calcule Σ_{j=0}^{n-i'} ε_{i'}^j
                es_{i'} ← es_{i'} + min(x_{i'-j}, ..., x_{i'+j})   # cf. éq. (6.14) p. 121
            es_{i'} ← es_{i'} + 2(i' − m)x_i + es_q
```

6.2.4 Algorithme incrémental

L'algorithme 6.3 décrit le calcul *incrémental* du score d'érosion. Contrairement aux cas précédents, les seules données accessibles dans cette approche sont celles d'indice inférieur à i. Une description plus complexe des approches par flux est donnée dans la section 6.3.

Pour l'implémentation de score d'érosion incrémental, afin de garantir la contrainte de l'éq. (5.1) p. 98 liée à la présence d'une valeur 0 et d'une valeur 1 dans le jeu de données, nous ajoutons ces deux valeurs au début du jeu de données. Ce choix est également discuté dans la section suivante.

L'algorithme fait un traitement particulier du cas $x_i > x_{i-1}$ car il permet d'améliorer notablement et simplement les performances et qu'il apparaît une fois sur deux en moyenne si les données sont i.i.d. par exemple. Sinon, q le premier point inférieur à gauche de x_i est recherché puis m est calculé et les scores d'érosion entre $m+1$ et n sont mis à jour conformément au théorème 3 p. 121.

En termes de complexité, le calcul incrémental du score d'érosion dépend de la complexité c_q liée au calcul de q et de celle c_{es} associée à la mise à jour des scores d'érosion.

La complexité de c_q est directement égale à la distance $dist = n - q$. c_{es} en dépend également du fait de la mise à jour des scores d'érosion $es_{i'}$ pour $i' = m+1...n$ dans le terme $\sum_{j=0}^{n-i'} \epsilon_{i'}^j$. En pratique, la mise à jour de es_i nécessite une itération, celle de es_{i-1} deux itérations et ainsi de suite jusqu'à $n - m$. Comme $m = (n+1+q)/2$, on a :

$$\begin{aligned} c_{es} &= \sum_{i=1}^{n-m} i = \frac{1}{2}\left(n - \frac{n+1+q}{2}\right)\left(n - \frac{n+1+q}{2} + 1\right) \\ &= \frac{1}{8}\left(dist - 1\right)\left(dist + 1\right) = \frac{1}{8}\left(dist^2 - 1\right) \end{aligned}$$

La complexité de la méthode est donc $c = c_q + c_{es} = dist + (dist^2 - 1)/8 = O(dist^2)$. Elle dépend donc quadratiquement de la distance à la plus proche valeur inférieure.

En faisant l'hypothèse forte que les données sont i.i.d., nous montrons dans l'annexe G p. 233 qu'indépendamment de leur distribution, $dist = \log(n)$ en moyenne, auquel cas la

complexité de la méthode est $O(\log^2(n))$.

Il est possible d'optimiser la recherche de q en conservant la dernière valeur calculée avant réception de x_i. Néanmoins, la complexité de la recherche de q n'intervenant que linéairement dans celle de la méthode, cette approche n'est pas détaillée ici.

6.2.5 Algorithme incrémental par niveaux

L'algorithme *incrémental par niveaux* du score d'érosion implique un certain nombre de manipulations d'indices et n'est pas représenté ici pour des raisons de simplicité. Son principe est donné ci-après.

Comme pour l'algorithme incrémental présenté ci-dessus, nous supposons pour la méthode incrémentale par niveaux que les valeurs sont comprises dans [0,1] et qu'une valeur 0 et une valeur 1 sont présentes au début des données, donc qu'il existe toujours une valeur inférieure à gauche de la nouvelle valeur reçue.

Comme indiqué dans le théorème 5 p. 123, la réception d'une nouvelle valeur implique dans cet algorithme la mise à jour de la matrice λ dans un premier temps suivie de la mise à jour des scores d'érosion en conséquence.

Ici également, le cas $x_i > x_{i-1}$ est traité de manière spécifique pour simplifier les calculs. En ce cas, le vecteur λ_i des indices des valeurs clés de x_i est égal au vecteur λ_{i-1} préfixé de i qui est toujours la première valeur clé de x_i pour tout i.

Lorsque $x_i \leq x_{i-1}$, la valeur q est cherchée à partir de la liste λ_{i-1} des indices clés de x_{i-1} puisque la première valeur inférieure à gauche de x_i est aussi une des valeurs clé de x_{i-1}, cf. la démonstration de la section F.2 p. 229 en annexe F p. 227.

m est ensuite calculé à l'aide de q. Pour $i' = m+1...i-1$, $k_{i'}$ est déterminé comme l'indice dans la chaîne des indices clés $\lambda_{i'}$ où l'indice i de la valeur reçue doit être insérée.

Après la mise à jour des $\lambda_{i'}$, celle des $es_{i'}$ est réalisée à l'aide des valeurs calculées ci-dessus avec en plus l'indice $p_{i'}$ qui désigne la position de q dans la suite des $\lambda_{i'}$. Ainsi, le score d'érosion n'est recalculé que pour les valeurs clés ayant changé, i.e. dont les indices sont compris entre $k_{i'}$ et $p_{i'}$.

La complexité de la méthode incrémentale par niveaux n'est pas détaillée ici, mais les expériences présentées dans la section 7.3 p. 151 montrent qu'elle est bien plus rapide que les autres méthodes, incrémentale notamment. Cette efficacité s'explique par l'utilisation de valeurs clés qui permet de calculer q à partir des valeurs clés des points précédents, mais également de mettre à jour les scores d'érosions des points impactés par l'arrivée du nouveau point à partir de leurs valeurs clés et donc sans avoir à recalculer toutes les érosions précédant celle où le nouveau point est impliqué.

6.3 DPE en flux

Les différentes approches détaillées ci-dessus permettent le calcul efficace du score d'érosion qui est l'étape la plus complexe de DPE. Nous présentons dans cette section une

approche pour le fonctionnement complet de DPE en flux, i.e. incluant également le calcul de $\overline{es}$, des groupes, de leurs tailles et des valeurs de période et de périodicité.

Lorsque les données sont accessibles par *lot* (*batch*) [1], l'algorithme a accès à toutes les données depuis son lancement jusqu'à sa fin. L'information d'une base de données statique à laquelle rien n'est ajouté, modifié, ou retiré, est par exemple accessible par lot.

Lorsque les données sont accessibles par *flux* (*stream*), l'algorithme y accède l'une après l'autre, de la plus ancienne à la plus récente. Dans ce cadre, les données peuvent être accessibles en *flux incrémental* (*landmark*) ou en *flux fenêtré* (*sliding window*). Dans le premier cas, toutes les données reçues sont conservées tandis que dans le second seules les w dernières le sont, auquel cas w représente la taille de la fenêtre.

Ces différents types d'accès sont de complexité variable, le plus simple étant celui par lot, puis celui par flux incrémental et enfin celui par flux fenêtré. Ces derniers sont de plus des cas particuliers de ceux de complexité supérieure : le flux fenêtré est un flux incrémental avec une taille de fenêtre infinie et le flux incrémental peut traiter des données par lot en y accédant l'une après l'autre, comme dans le cas des méthodes incrémentales présentées dans la section précédente.

Le premier paragraphe de cette sous-section décrit un bref état de l'art des méthodes d'analyse des flux tandis que le second donne un point de vue algorithmique général sur la méthode DPE en flux.

6.3.1 Méthodes d'analyses des flux de données

Si les algorithmes d'analyse de données (fouille de données et apprentissage artificiel) développés pour des accès par lot sont aujourd'hui nombreux, les approches destinées à traiter des flux (*stream mining*) sont plus récentes et connaissent aujourd'hui un développement important (Shaker & Hüllermeier, 2015).

Les approches par flux sont basées sur les principes suivants : les données ne sont accessibles qu'une fois ou seules les w dernières le sont et w est très inférieur au nombre de données du flux, les données arrivent rapidement et doivent être traitées immédiatement et le système de traitement doit ne consommer que peu de mémoire (Shaker & Hüllermeier, 2015). Par rapport à l'analyse *big data* en lot qui s'intéresse à des volumes de données toujours plus importants, l'analyse *fast data* en flux permet celle d'information arrivant toujours plus vite (Lemaire, 2014).

Les raisons du développement récent de ces méthodes sont multiples. D'abord, la production de données augmente de manière plus rapide que la capacité à les traiter et l'approche par flux permet de les analyser en temps réel sans avoir à les stocker (Lemaire et al., 2015). De plus, le traitement de données anciennes n'est pas toujours pertinent et la mise à jour de modèles en fonction des plus récentes l'est souvent plus pour l'utilisateur (Lemaire, 2014). L'analyse des seules données récentes permet également de s'affranchir de leur stationnarité, contrairement aux études couvrant un intervalle temporel important (Shaker

1. La terminologie des types d'accès est celle donnée par Lemaire (2014)

& Hüllermeier, 2015). Enfin, certaines données ne sont accessibles qu'en flux, comme celles issues de capteurs temps-réel (Krempl et al., 2014).

Leskovec et al. (2014) présentent un ensemble de techniques permettant d'effectuer des tâches simples lorsque toutes les données sont accessibles sans contrainte sur leur temps de traitement mais potentiellement complexes sur des flux. Ces approches sont approximatives et basées pour la plupart sur des techniques de hachage : filtre de Bloom pour tester l'appartenance d'un élément reçu à une liste, décompte de données uniques ou estimation de moments statistiques. D'autres calculs de statistiques sont proposés, pour le min et le max (Lemire, 2006) ou la médiane (Manku et al., 1998) par exemple.

Les techniques d'apprentissage supervisé sur des flux sont également étudiées sous l'angle du *concept drift*, i.e. de l'évolution du modèle dans le temps (Gama et al., 2014).

Enfin, dans le but d'accélérer le traitement des données reçues, Gaber et al. (2005) présentent des approches permettant de n'en conserver que certaines par échantillonnage (*sampling*) ou suppression selon certains critères (*load shedding*) ou encore en les stockant sous forme synthétique à l'aide de statistiques (*synopsis data structure*) mentionnées plus haut ou par projection sur des espaces de dimensions inférieures (*sketching*). La dernière approche, retenue pour DPE, consiste à n'étudier que les w dernières données reçues.

La recherche sur la fouille de flux est en plein développement et contient encore un grand nombre de questions ouvertes, listées entre autres par Krempl et al. (2014).

6.3.2 Algorithme général

Nous présentons sur la figure 6.7 un schéma indiquant les étapes principales d'un algorithme pour DPE en flux.

Vis-à-vis de l'approche DPE présentée au chapitre 5, le score d'érosion es_i du point est calculé en même temps que $\overline{es}_i$, le score d'érosion de $\overline{x}_i$. Les deux branches des étape de niveau *Points* sont donc exécutées en parallèle.

Il convient de noter que la première étape de ces branches est la normalisation du point reçu. En notant $W \subset X$ l'ensemble des points disponibles dans la fenêtre à un moment donné, il n'est plus possible de valider la contrainte de l'éq. (5.1) p. 98, en particulier qu'à tout moment $\exists x_i, x_j$ tels que $x_i = 0$ et $x_j = 1$. Cette contrainte est simple à obtenir par prétraitement lorsque l'ensemble des données à étudier est disponible au début de l'algorithme, mais ne peut plus être exigée en flux, sauf à les contraindre de manière très importante. Ainsi, nous proposons d'étendre le calcul du score d'érosion et de la fonction γ_{es} pour des valeurs dans $\mathbb{R}$. Les formules à jour pour le calcul de la normalisation des points, du score d'érosion et de γ_{es} sont détaillées dans l'annexe E p. 223. L'impact de ce changement sur les algorithmes est discuté dans le paragraphe suivant.

Les étapes de niveau *Points* suivant celles de normalisation sont identiques à celles de DPE telle que présentées au chapitre 5.

Une fois déterminé le type du nouveau point reçu lors de la dernière étape de niveau *Points*, les étapes de niveau *Groupes* démarrent. Ces dernières sont également présentées en deux branches, exclusives cette fois, dépendant du type de point déterminé : si le point

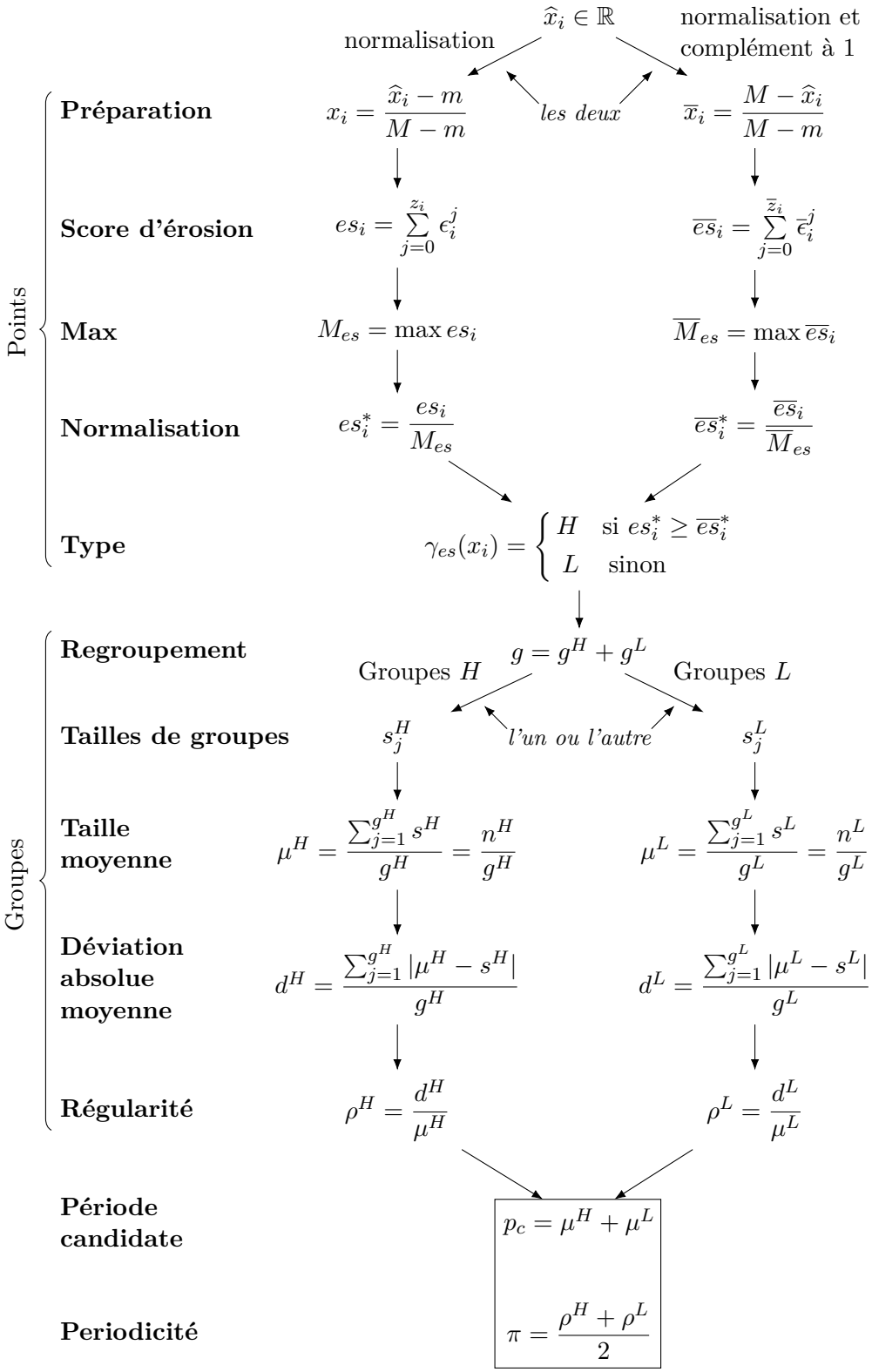


FIGURE 6.7 – Étapes pour DPE en flux

reçu est de type H (resp. L), la branche de gauche (resp. droite) est exécutée. Les étapes représentées pour le calcul des groupes sont identiques à celles du chapitre précédent.

Spécificités de DPE en flux Le relâchement de la contrainte de l'éq. (5.1) p. 98 rend l'implémentation de DPE plus complexe mais également plus générale. En effet, comme précisé dans la section 5.2.2 p. 101, la convergence du score d'érosion es est assurée par la présence d'au moins une valeur nulle dans les données sur lesquelles il est calculé. De manière symétrique, la convergence d'$\overline{es}$ l'est par la présence d'une valeur 1.

Dans le cadre d'un flux, seule la présence d'une valeur minimale m et d'une valeur maximale M est garantie dans les données de la fenêtre W. Ainsi, comme détaillé dans l'annexe E p. 223, es et $\overline{es}$ dépendent des valeurs égales à m et M respectivement. Leur mise à jour dépend du type de flux considéré, incrémental ou fenêtré.

Dans le premier, les données reçues sont conservées et les seuls changement dans W sont liés à l'arrivée de nouvelles données. En ce cas, les valeurs de m et M sont simples à maintenir et sont monotones, i.e. m ne peut que décroître et M que croître. Nous détaillons ci-après l'impact de l'arrivée d'une nouvelle donnée sur les valeurs déjà calculées d'es_i, les mêmes principes étant applicables à $\overline{es}_i$.

Lorsqu'une nouvelle donnée est reçue, elle est ou non inférieure à m. Si elle ne l'est pas, la méthode décrite dans la section 6.1.4 p. 119 pour la mise à jour incrémentale des scores d'érosion est directement applicable. Si elle l'est en revanche, l'ensemble des scores d'érosion de la fenêtre doivent être recalculés car la valeur m change de position. D'une manière générale, lorsque m décroît, les es_i sont plus importants relativement aux $\overline{es}_i$ et les groupes hauts sont plus nombreux et/ou plus larges. Ce phénomène s'explique intuitivement par l'effet de la normalisation des données dans $[m, M]$. Supposons par exemple des oscillations comprises entre 9,9 et 10, interprétées comme des alternances de groupes haut et bas, suivies d'une valeur soudainement plus basse, 0 par exemple : une fois renormalisées dans [0,10], les oscillations sembleront planes et les données seront vues comme composées d'un groupe haut contenant les données précédentes et d'un groupe bas débutant avec la valeur 0.

Dans le cas d'un flux fenêtré, le problème est plus complexe car les changements de W sont dus à l'arrivée de nouvelles données mais également au retrait des plus anciennes. Dans ce cadre, le maintien des valeurs m et M n'est plus trivial et leur évolution n'est plus monotone, i.e. la valeur de m peut croître et celle de M décroître d'un instant au suivant.

Un certain nombre de solutions sont toutefois envisageables dans ce cadre. D'une part, des méthodes efficaces de mise à jour de m et M existent, comme celle de Lemire (2006) mentionnée plus haut. D'autre part, la technique de mise à jour incrémentale du score peut être appliquée de manière symétrique au cas du retrait des données en ne mettant à jour que les scores d'érosion des points situés sur la moitié gauche de l'intervalle entre le point retiré et le premier point inférieur dans W.

De plus, l'analyse d'un flux fenêtré a également des impacts concernant les groupes

identifiés et donc la période et la périodicité. En effet, puisque les groupes haut et bas peuvent être amenés à changer avec l'arrivée de données et le départ des plus anciennes, il est possible qu'un point appartenant à un groupe H à un instant appartienne à un groupe L à l'instant d'après. En reprenant l'exemple ci-dessus, lorsque seules les oscillations dans [9,9 ; 10] sont présentes, le flux peut paraître périodique, mais lorsque la valeur 0 arrive, alors seuls deux groupes sont définis et la série n'est plus périodique.

Plusieurs approches peuvent être retenues pour traiter ce cas, notamment la contextualisation des résultats, i.e. l'association d'une période et d'une périodicité à un sous-ensemble de points de la fenêtre. Nous proposons une méthode de ce type au chapitre 8.

Plus généralement, la formalisation des approches par flux constitue une perspective de la méthode DPE.

Comparaison avec les implémentations présentées

À différents égards, les algorithmes incrémentaux présentés dans les sous-sections 6.2.4 et 6.2.5 pp. 126-127 sont plus spécifiques que la méthode générale présentée ici.

D'abord, ils fonctionnent en flux incrémental et non en flux fenêtré, ce qui pose le problème de l'occupation mémoire au bout d'un certain moment. De plus, les calculs de es et $\overline{es}$ sont réalisés en deux passes séquentielles au lieu du calcul simultané proposé ici. D'autre part, ils n'intègrent pas le calcul incrémental des groupes et la question de leur mise à jour. Enfin, même dans le cadre du flux incrémental, ils éludent la question du maintien de m et M en ajoutant en début de flux une valeur 0 et une valeur 1. Cette approche est correcte si les données suivantes contiennent des 0 et des 1 auquel cas ces valeurs deviendront les nouvelles valeurs de référence pour les données ultérieures. Si au contraire le 0 et le 1 ajoutés en début de flux sont les seuls, alors ils entraînent un biais important dans l'analyse des valeurs suivantes.

Ainsi, les implémentations présentées dans la section précédente apportent un éclairage indispensable pour la compréhension du score d'érosion mais méritent d'être généralisées au cadre présenté dans cette section.

6.4 Bilan

Nous avons présenté dans ce chapitre les différentes approches retenues pour mettre en œuvre la méthode DPE introduite au chapitre 5. Du fait de l'importance du score d'érosion dans la méthode et de sa nouveauté, nous nous sommes attaché dans un premier temps à introduire des approches efficaces pour son calcul : par niveaux, incrémental, et incrémental par niveaux.

Par la suite, nous avons présenté dans la deuxième section différentes implémentations de ces approches ainsi que leurs complexités.

Dans la dernière section, nous avons proposé un modèle général en flux pour DPE, étendant les méthodes de calcul efficace du score d'érosion à l'ensemble des étapes de la méthode.

Chapitre 7

Expériences

> Plongeur sous-marin débutant, cherche équipement réduit pour expérimentation en lavabo.
>
> —Pierre Dac, *L'Os à moelle*

La méthode DPE et ses variantes, présentées au chapitre 5, ainsi que les algorithmes et leurs implémentations, détaillés au chapitre 6, ont fait l'objet d'un nombre important d'expériences présentées dans ce chapitre.

Ces dernières sont réparties en deux études expérimentales : la première vise à valider la pertinence de DPE et la seconde ses performances. Ces deux études utilisent de nombreuses données synthétiques de formes, de tailles et de bruit variés, créées par un *générateur* présenté dans la section 7.1. Elles se réfèrent également à un ensemble de *critères* et à un *protocole* expérimental permettant de les vérifier.

Dans l'étude de la *pertinence* de DPE, détaillée dans la section 7.2 p. 139, les critères retenus sont liés au comportement des méthodes de regroupement et de calcul de la période et de la périodicité en fonction des différents types de bruit utilisés pour générer le jeu de données. Plus précisément, le degré de périodicité doit décroître avec le bruit dans les données, cette décroissance doit être régulière, pour des niveaux de bruits équivalents la méthode doit renvoyer des résultats équivalents, la période estimée doit être la plus proche possible de la période « réelle » et l'appartenance des points aux groupes hauts ou bas renvoyée par la méthode de regroupement doit être la plus précise possible.

Cette première étude présente les résultats obtenus selon ces critères avec les différentes variantes de DPE sur des jeux de données créés dans le cadre de *scénarios* qui détaillent les paramètres de génération utilisés. Parmi ces paramètres, l'un contrôle le bruit par paliers successifs entre 0 et 1 afin de permettre la création de jeux de données du moins bruité, donc strictement périodique, au plus bruité, apériodique.

Dans l'étude de la *performance* de DPE présentée dans la section 7.3 p. 151, les différentes approches liées à l'étape de regroupement par score d'érosion sont comparées. Une attention particulière est portée à cette étape du fait de sa complexité importante,

Tableau 7.1 – Étapes et paramètres de génération des données artificielles

	Étape	**Paramètres** ($\tau \in \{H, L\}$)
1.	Génération des étiquettes H et L	p^τ : taille des groupes ν_s^τ : bruit sur la taille des groupes ν_o^τ : bruit d'oubli sur les groupes
2.	Génération des valeurs	forme : Rectangle, Sinus, Vague, Triangle ν_v : bruit sur les valeurs ν_t : paramètre de tendance
3.	Normalisation des données	-

montrée dans dans la section 6.2 p. 123 du chapitre précédent.

Les critères retenus pour les performance sont le temps et l'occupation mémoire consommés par les différentes variantes de DPE. Le scénario utilisé génère des données croissantes en taille jusqu'à un million de points, avec des paramètres de formes et de bruits variés afin de comparer les méthodes dans les cas les plus divers. Nous présentons également une version plus précise de la complexité des méthodes en fonction des paramètres de génération, permettant de retrouver analytiquement les résultats obtenus expérimentalement.

Enfin, la section 7.4 p. 157 présente une application de la méthode DPE sur un jeu de données réelles permettant de vérifier la pertinence des trois résultats qu'elle renvoie, à savoir le degré de périodicité, la période et le rendu linguistique.

Les expériences menées sur la pertinence de DPE ont été publiées dans (Moyse et al., 2013a) et (Moyse et al., 2013b) et celles sur sa performance dans (Moyse & Lesot, 2014).

7.1 Générateur de données artificielles

Un grand nombre d'expériences ont été exécutées sur des données artificielles dans le but de tester un ou plusieurs points spécifiques des méthodes comparées. L'intérêt des données artificielles réside dans la connaissance des paramètres utilisés pour leur génération et donc de la réponse attendue pour la méthode.

Le générateur de données que nous présentons permet de tester l'ensemble des cas de période et de périodicité présentés sur la figure 4.2 p. 71 à l'exception des séries symboliques et de période non constante, ces dernières faisant l'objet du chapitre 8.

Il fonctionne en trois étapes : d'abord des groupes H et L de points hauts et bas respectivement sont créés selon les paramètres de taille donnés en entrée, puis des valeurs sont associées à chaque point des groupes selon les paramètres de forme et de bruit et enfin normalisées dans $[0, 1]$.

Ces étapes sont résumées dans le tableau 7.1 et détaillées dans les trois sous-sections suivantes. De plus, la figure 7.1 donne des exemples de jeux de données générés par la méthode en fonction de ses paramètres.

La dernière sous-section détaille le calcul des valeurs de références associées aux para-

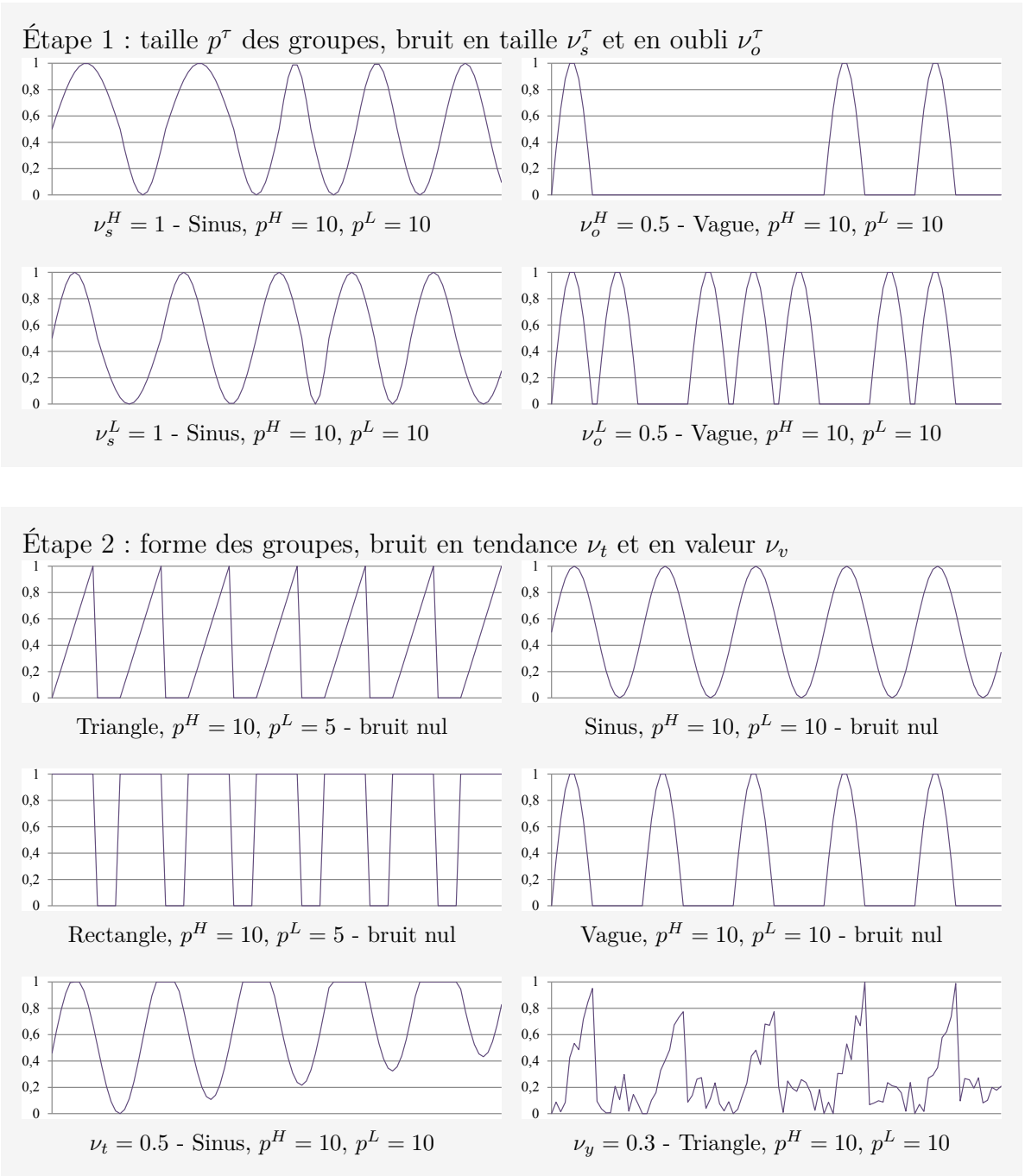


FIGURE 7.1 – Jeux de données générés. Les figures illustrent l'influence du premier paramètre indiqué dans la légende.

mètres de génération, utilisées par la suite pour l'évaluation des critères de qualité.

7.1.1 Étape 1 : Génération des étiquettes H et L

La première étape de génération consiste en la création de groupes d'étiquettes H ou L de taille p^H et p^L spécifiées en entrée. Ces derniers sont ajoutés alternativement les uns à la suite des autres en commençant par les groupes H jusqu'à ce que le nombre d'étiquettes

générées soit égal au nombre n de données souhaitées.

Avec $n = 10$, $p^H = 3$ et $p^L = 2$, les étiquettes crées sont $HHHLLHHHLL$ et les tailles de groupes sont $s_1^H = 3$, $s_2^H = 3$, $s_1^L = 2$ et $s_2^L = 2$ puisque le premier et le troisième groupes sont de type H et de longueur 3 et que le second et le quatrième sont de type L et de longueur 2.

Le paramètre de bruit en taille ν_s^τ est utilisé pour agrandir ou rétrécir aléatoirement la taille s_j^τ des groupes de type τ, avec $\tau \in \{H, L\}$ selon le type de groupe. Formellement :

$$s_j^\tau = \lceil 1 + \operatorname{sgn}(0.5 - \epsilon_1) \times \nu_s^\tau \times \epsilon_2 \rceil \, p^\tau \tag{7.1}$$

où ϵ_1 et ϵ_2 sont des variables aléatoires uniformes dans $[0; 1]$ et $\operatorname{sgn}(x)$ renvoie le signe de x. ϵ_1 donne le sens de la modification de la taille du groupe, agrandissement ou diminution selon qu'il est supérieur ou non à 0,5, ϵ_2 donne l'intensité de cette modification et ν_s^τ pondère le tout. Ainsi, la taille des groupes varie entre $(1 - \nu_s^\tau)\, p^\tau$ et $(1 + \nu_s^\tau)\, p^\tau$.

Le bruit en « oubli » ν_o^τ affecte l'alternance des groupes H et L en retirant aléatoirement un groupe. Plus précisément, les points d'un groupe de type τ ne sont générés que si $\epsilon_1 \geq \nu_o^\tau$, sinon le groupe est ignoré et les points sont créés selon les paramètres du type opposé (H si $\tau = L$ et L sinon). Ainsi, plusieurs groupes de même type peuvent être générés à la suite avec une probabilité d'autant plus grande que ν_o^τ l'est aussi.

7.1.2 Étape 2 : Génération des valeurs

La génération des valeurs est réalisée par le calcul successif de deux séries intermédiaires, $\dot{X}$ et $\ddot{X}$. La première est obtenue par application du paramètre de forme à la séquence d'étiquettes calculée à l'étape précédente et la seconde par ajout d'un bruit en valeurs donné par le paramètre ν_v et d'une tendance donnée par ν_t.

Le calcul de $\dot{X}$ repose sur l'assignation de valeurs élevées aux étiquettes H et faibles aux étiquettes L selon un paramètre de forme parmi Rectangle, Sinus, Vague et Triangle.

En notant a l'indice du premier point du $j^{ème}$ groupe et τ son type, ses valeurs $\dot{x}_i$ sont calculées pour $i = a...a + s_j^\tau$ par, pour la forme Rectangle :

$$\dot{x}_i = \begin{cases} 1 & \text{si } \tau = H \\ 0 & \text{sinon} \end{cases} \tag{7.2}$$

pour la forme Sinus :

$$\dot{x}_i = \frac{1}{2} + \frac{\lambda}{2} \sin\left(\pi \frac{i - a}{s_j^\tau}\right) \text{ avec } \lambda = \begin{cases} 1 & \text{si } \tau = H \\ -1 & \text{sinon} \end{cases} \tag{7.3}$$

pour la forme Vague :

$$\dot{x}_i = \begin{cases} \sin\left(\pi \dfrac{i - a}{s_j^\tau}\right) & \text{si } \tau = H \\ 0 & \text{sinon} \end{cases} \tag{7.4}$$

et pour la forme Triangle :

$$\dot{x}_i = \begin{cases} \dfrac{i-a}{s_j^\tau} & \text{si } \tau = H \\ 0 & \text{sinon} \end{cases} \tag{7.5}$$

La série $\ddot{X}$ est ensuite obtenue par application de la tendance ν_t et du bruit en valeurs ν_v sur les éléments de $\dot{X}$:

$$\ddot{x}_i = \frac{\nu_t}{n} + \begin{cases} \dot{x}_i - \nu_v\epsilon & \text{si } \dot{x}_i \text{ est dans un groupe } H \\ \dot{x}_i + \nu_v\epsilon & \text{sinon} \end{cases} \tag{7.6}$$

ou ϵ est une variable uniforme $\mathcal{U}(0;1)$. Le bruit ν_v augmente donc aléatoirement les valeurs des groupes L et diminue celles des groupes H, et la tendance ν_t augmente linéairement les valeurs des points indifféremment de leur groupe.

7.1.3 Étape 3 : Normalisation

Enfin, le jeu de données final X est obtenu par normalisation de $\ddot{X}$:

$$x_i = \frac{\ddot{x}_i - m}{M - m}$$

où $m = \min \ddot{x}_i$ et $M = \max \ddot{x}_i$, ce qui permet de vérifier que toutes les données sont comprises dans $[0, 1]$ et qu'une au moins est égale à 0 et une autre à 1, conformément à la contrainte spécifiée dans l'éq. (5.1) p. 98.

7.1.4 Calcul des valeurs de référence

L'usage de données générées à partir d'un ensemble de paramètres permet la définition d'une vérité terrain établissant les résultats attendus. Comme discuté plus bas, sa définition n'est pas triviale du fait de l'usage de paramètres de bruit pour la génération des données.

Les valeurs de référence sont la période de référence p_{ref} ainsi que les étiquettes de référence e_{ref}, détaillées ci-dessous. La périodicité de référence n'est pas définissable car non déductible des paramètres de génération. La section 7.2 p. 139 détaille néanmoins un certain nombre de critères pour définir sa qualité.

Période de référence

Lorsqu'aucun bruit ν_s sur la taille des groupes n'est utilisé, la période de référence p_{ref} est définie comme $p^H + p^L$, i.e. la somme de la taille des groupes H et des groupes L qui définissent le motif périodique de base et donc la période de la série.

En revanche, lorsque le bruit ν_s^τ est utilisé, ces tailles ne sont plus celles effectivement présentes dans la série créée. Nous proposons donc de définir la période de référence p_{ref}

comme la somme des tailles moyennes des groupes H et L générés, soit :

$$p_{ref} = \frac{1}{2}\left(\frac{1}{n^H}\sum_{j=1}^{n^H} s_j^H + \frac{1}{n^L}\sum_{j=1}^{n^L} s_j^L\right) \tag{7.7}$$

Étiquettes de référence

Les étiquettes de référence $e_{ref}(x_i)$ sont celles créées à la première étape de génération. Selon les types de bruits utilisés, les étiquettes de référence peuvent être différentes de ce qui serait intuitivement attendu d'un groupe haut ou bas. Par exemple, si une étiquette H est transformée en 1 avec la forme « Rectangle » mais qu'un bruit en valeur important est utilisé, sa valeur peut être proche de 0 (cf. éq. (7.6)). En ce cas, l'étiquette de référence fait foi et le point est tout de même considéré comme de type H.

En pratique, comme les bruits utilisés sont uniformes, les cas extrêmes sont rares en moyenne et les séries renvoyées par le générateur coïncident majoritairement avec les valeurs intuitivement associées à un groupe H ou un groupe L. De plus, ces cas sont également intéressants pour tester la robustesse de la DPE dont l'objectif est le calcul de la période et de la périodicité et non l'identification de groupes H et L. Autrement dit, même si certains sont étiquetés de manière contre-intuitive, DPE doit être en mesure de retrouver le résultat de période sans être trop affectée par ces valeurs aberrantes.

7.1.5 Protocole expérimental

Afin d'effectuer les études expérimentales en termes de pertinence et de performance, nous définissons un ensemble de scénarios permettant de générer des données avec un paramètre de bruit dont la valeur augmente progressivement.

À chaque pas d'évolution du paramètre variable, plusieurs jeux de données (20 ou 40 selon les scénarios) sont créés avec les mêmes paramètres afin d'évaluer la robustesse de la méthode testée sur des séries différentes mais de bruit comparable.

Les paramètres variables que nous utilisons sont le bruit ν_s sur la taille des groupes et celui ν_v sur les valeurs. Ils varient de 0 à 1 par pas de 0,05, soit 21 valeurs différentes. La même valeur de paramètre ν_s est utilisée pour ν_s^H et ν_s^L (cf. tableau 7.1 p. 134). Les autres paramètres de générations ν_o et ν_t ne sont pas utilisés dans les scénarios, mais peuvent l'être pour créer des jeux de données spécifiques, comme le paramètre d'oubli ν_o utilisé pour l'exemple de la figure 7.6 p. 148.

Le tableau 7.2 présente les scénarios S1 à S4 définis pour l'étude expérimentale de la pertinence et SP pour celle de la performance. Sont spécifiés pour chaque scénario, le paramètre variable ν_s ou ν_v, et les paramètres fixes : tailles p^H et p^L pour les groupes H et L respectivement, formes, et nombre de séries générées pour chaque jeu de paramètres.

Enfin, chaque scénario est associé aux aspects étudiés, liés aux étapes de la méthode DPE ou à sa performance, aux critères, définis dans la section 7.2.1 p. 140 pour l'étude de pertinence et dans la section 7.3.1 p. 151 pour celle de performance et aux résultats graphiques obtenus.

TABLEAU 7.2 – Protocole pour les études expérimentales de pertinence (S1 à S4) et de performance (SP). *Regroup.* désigne les méthodes de regroupement, s le calcul de la taille des groupes, μ celui de leur tendance centrale, d de leur déviation, π de leur périodicité et *Perf.* la performance de la méthode.

Scénarios	S1	S2	S3	S4	SP
$\nu_s = 0...1$, $\nu_v = 0$		✓	✓	✓	✓
$\nu_s = 0...1$, $\nu_v = 0,3$		✓		✓	✓
$\nu_v = 0...1$, $\nu_s = 0$	✓	✓	✓	✓	✓
$\nu_v = 0...1$, $\nu_s = 0,5$		✓		✓	✓
$\left(p^H, p^L\right)$	$(35, 15)$	$(25, 25)$, $(10, 40)$	$(6, 6)$	$(25, 25)$, $(20, 5)$	$(50, 50)$, $(90, 10)$
Forme(s)	Rectangle	Rectangle, Sinus	Rectangle	Rectangle, Triangle	Rectangle, Triangle, Vague, Sinus
Répétitions	40	20	40	20	20
Aspect(s) étudié(s)	Regroup., s	Regroup.	μ, d, π	μ, d	Perf.
Critères	Tous sauf C4	Tous	Tous sauf C4	Tous sauf C4	Temps, mémoire
Résultats	figure 7.2 p. 141	figure 7.3 p. 142	figure 7.4 p. 143	figure 7.5 p. 144	figure 7.8 p. 153

7.2 Étude expérimentale de la pertinence de la méthode DPE et de ses variantes

La première série d'expériences présentée a pour objectif de valider la pertinence de la méthode DPE, donc sa capacité à déterminer la période et la périodicité d'une série temporelle affectée de différents types de bruits.

L'étude du rendu linguistique n'est pas effectué ici car difficile à réaliser simplement avec les paramètres de génération. Comme mentionné dans les perspectives de cette thèse, cette étude pourrait être réalisée par la biais de questionnaires utilisateurs, dans le même ordre d'idée que ceux présentés par Laurent et al. (2004) ou Newstead et al. (1987) évoqués dans la section 2.1.2 p. 28.

Les expériences présentées dans cette section permettent d'identifier les variantes de DPE les plus efficaces parmi celles présentées dans les sections 5.2 p. 98 et 5.3 p. 105.

Afin de pouvoir les comparer précisément, des critères de qualité concernant les résultats attendus sont définis dans la section 7.2.1. Les résultats obtenus sont quant à eux présentés dans la section 7.2.2 p. 141 puis rapprochés des critères de qualité selon chacun des aspects étudiés de DPE : méthodes de regroupement dans la section 7.2.3 p. 145, évaluation de la taille des groupes dans la section 7.2.4 p. 148, de leur tendance centrale dans la section 7.2.5 p. 149, de leur dispersion dans la section 7.2.6 p. 150 et de la périodicité

de la série dans la section 7.2.7 p. 150.

7.2.1 Critères de qualité

Nous définissons cinq critères de qualité C1 à C5 afin de spécifier le comportement souhaité de la méthode : les critères C1 et C2 mesurent la qualité du degré de périodicité, C3 la précision du calcul de la période, C4 la qualité du regroupement et C5 la robustesse de la méthode. Plus précisément, ces critères sont définis ainsi :

C1 : le degré de périodicité π doit être égal à 1 pour une série strictement périodique, i.e. vérifiant l'éq. (4.2) p. 70, comme la série `(a)` de la figure 4.2 p. 71.
Le générateur utilisé avec des valeurs de bruit nulles renvoie une telle série. Ce critère permet de pénaliser une méthode qui renverrait un degré de périodicité plus petit que 1 dans ce cas.

C2 : π doit décroître de manière régulière lorsque le bruit utilisé pour générer la série augmente et qu'elle devient moins périodique.
Comme illustré plus bas par les résultats des expériences, la décroissance de π n'est pas constante en fonction du type de bruit utilisé. En revanche, le critère permet de vérifier que la périodicité ne croît pas avec le bruit et d'autre part que la décroissance est régulière, afin de pénaliser une méthode qui renverrait deux périodicités très différentes pour des niveaux de bruit voisins.

C3 : la différence entre la période de référence p_{ref} définie dans la section 7.1.4 p. 137 et celle renvoyée par la méthode doit être égale à 0.
Dans le cas où la série étudiée est périodique ou pseudo-périodique, ce critère quantitatif permet de mesurer la capacité de la méthode à évaluer correctement la période. La valeur utilisée est le pourcentage d'erreur entre la période de référence p_{ref} et la période p_c renvoyée par DPE :

$$\Delta p = |p_c - p_{ref}|/p_{ref} \tag{7.8}$$

C4 : les étiquettes H et L de référence définies dans la section 7.1.4 p. 138 doivent coïncider avec les étiquettes H et L renvoyées par la méthode de regroupement γ, i.e. $\forall i = 1...n$, $e_{ref}(x_i) = \gamma(x_i)$.
Ce critère correspond au taux de bonne classification utilisé pour évaluer les méthodes en apprentissage supervisé. Dans le cadre des tests, nous le définissons comme :

$$Acc = \frac{1}{n} \sum_{x \in X} \mathbb{1} \left(\gamma \left(x \right) = e_{ref} \left(x \right) \right) \tag{7.9}$$

C5 : l'écart-type de π, Δp et Acc des résultats obtenus sur plusieurs expériences avec un même niveau de bruit doit être égal à 0.
Ce critère valorise les méthodes robustes. En pratique, la variabilité des résultats renvoyés par une méthode, même la meilleure, n'est jamais 0 car des données générées aléatoirement avec un niveau de bruit important peuvent être très périodiques

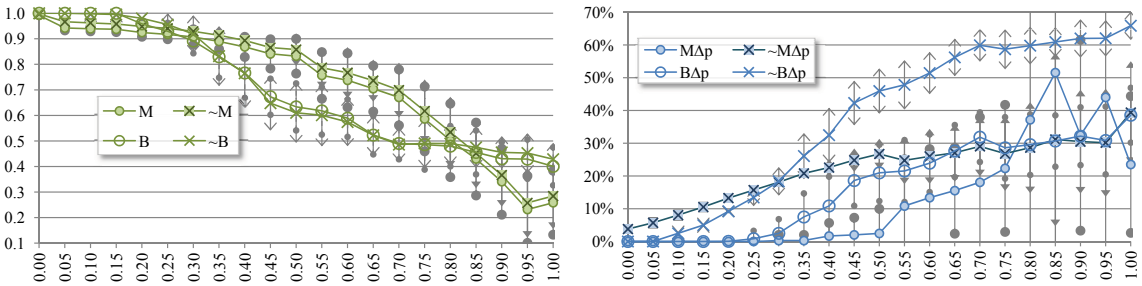


FIGURE 7.2 – Résultats du scénario S1 : comparaison pour un bruit en valeurs ν_v croissant des méthodes de regroupement γ_{es} et γ_{BL} et des cardinalités C et $\widetilde{X}$ pour le calcul de la taille des groupes.

par hasard. Cependant, le nombre de répétitions des expériences avec les mêmes paramètres permet d'assurer que ces jeux de données particuliers restent marginaux et ne biaisent pas statistiquement les résultats obtenus.

7.2.2 Résultats

Les résultats obtenus avec les scénarios S1 à S4 définis dans le tableau 7.2 p. 139 sont présentés ci-dessous. La discussion des différents aspects qu'ils permettent de mettre en valeur est donnée dans les sous-sections suivantes.

S1 Le scénario S1 étudie les méthodes de regroupement γ_{es} et γ_{BL} décrites dans la section 5.2 p. 98 et les cardinalités C et $\widetilde{X}$ détaillées dans la section 5.3.1 p. 105 pour un bruit en valeurs ν_v croissant. Les paramètres $t_v = 0,8$ et $t_m = 2\%$ sont retenus pour γ_{BL}.

Les résultats du scénario sont donnés sur la figure 7.2. La moyenne et l'écart-type de la périodicité π sont illustrés sur le graphique de gauche et ceux de l'erreur de calcul de la période Δp sur celui de droite. Les courbes dont le nom commence par une tilde utilisent la cardinalité $\widetilde{X}$, sinon la cardinalité C. Celles dont le nom contient M utilisent la méthode γ_{es} et celles dont le nom contient B la méthode γ_{BL}.

S2 Le scénario S2 compare les méthodes γ_{es}, γ_{BL} et γ_W pour des bruits croissants en valeurs et en taille, deux formes et deux couples de tailles (p^H, p^L).

Les paramètres de la méthode γ_{BL} sont $t_v = 0,5$ et $t_m = 0$, donc la fusion entre groupes adjacents n'est pas utilisée. Pour γ_W, la taille de la fenêtre de lissage est $w = 10$.

Ce scénario entraîne la création de 3 200 jeux de données du fait des 4 combinaisons de paramètres évoluant sur 10 paliers, des 2 couples de tailles p^H et p^L, des deux formes Rectangle et Sinus et des 20 répétitions. Ainsi, tous les graphiques renvoyés ne sont pas représentés ici et seuls ceux obtenus pour les combinaisons Rectangle, Sinus, $\nu_v = 0...1$ et $\nu_s = 0...1$, significatifs de l'ensemble, sont représentés sur la figure 7.3.

Le bruit ν_s en taille est noté « Noise Grp Size » et ν_v en valeurs « Noise Y ». Les graphiques sont groupés par trois, représentant la moyenne et l'écart-type de π pour celui du haut, de Δp pour celui du milieu et de *Acc* pour celui du bas. La colonne de gauche

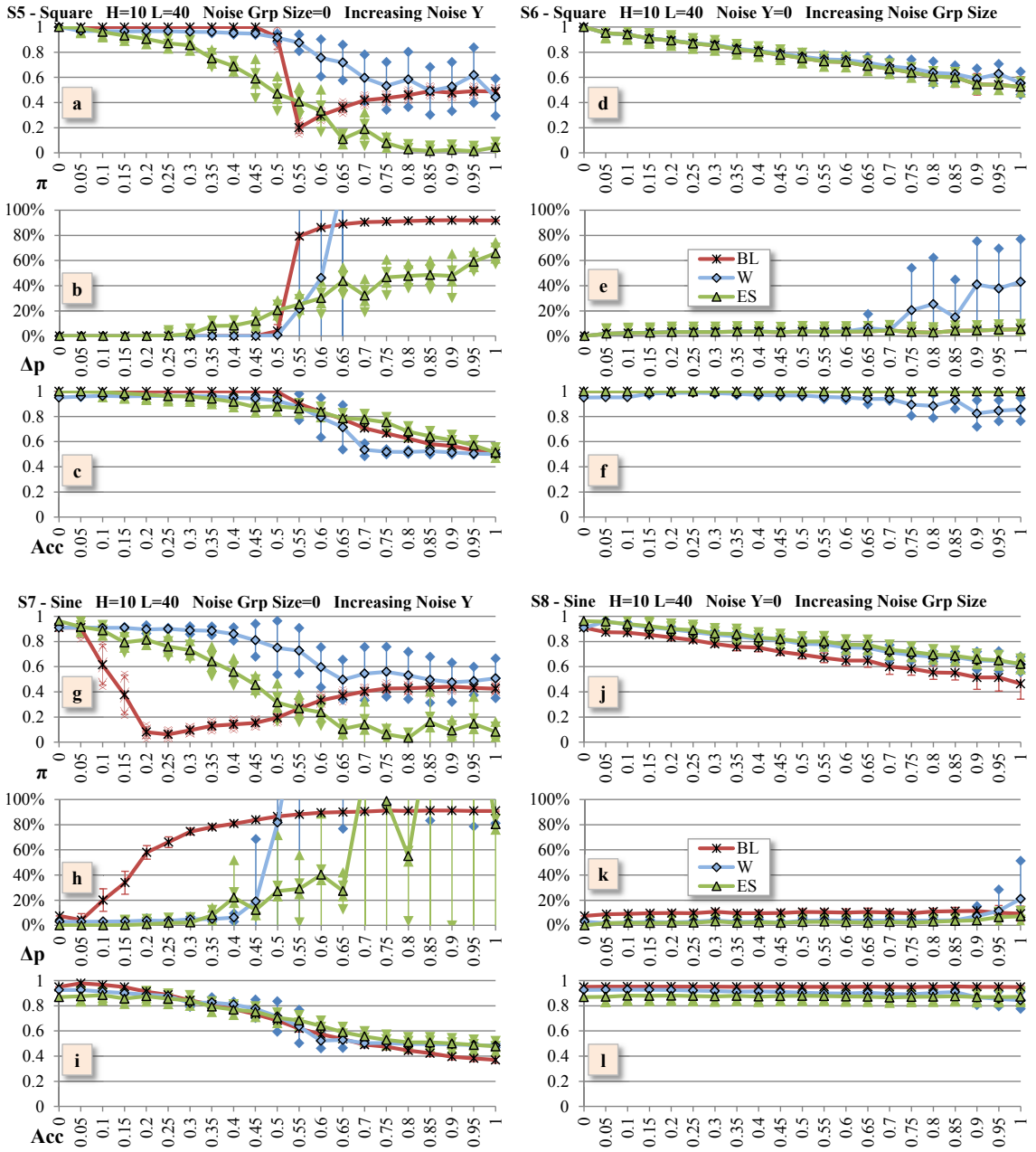
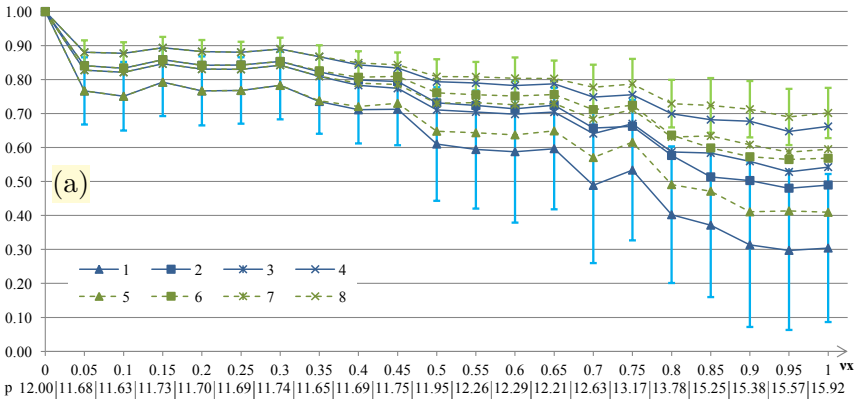
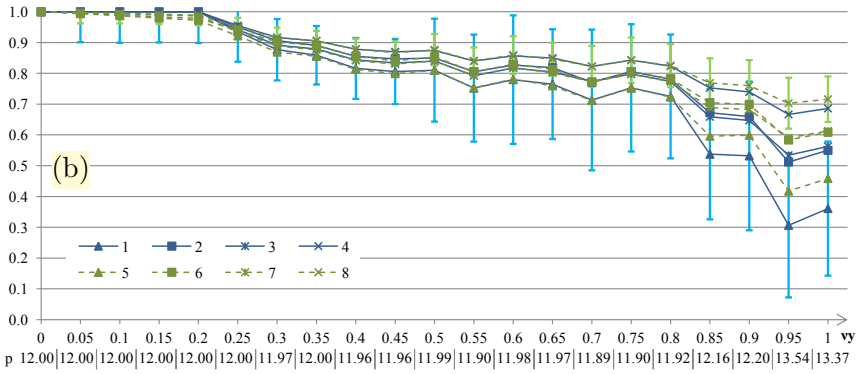


FIGURE 7.3 – Résultats du scénario S2 : comparaison des méthodes de regroupement γ_{es}, γ_W et γ_{BL} pour deux formes et des bruits variables en valeurs et en taille.



Moyenne et écart-type de π - bruit en taille ν_s croissant



Moyenne et écart-type de π - bruit en valeurs ν_v croissant

1 : C,σ,Min 2 : C,d,Min 3 : C,σ,Moy 4 : C,d,Moy
5 : $\widetilde{X}$,σ,Min 6 : $\widetilde{X}$,d,Min 7 : $\widetilde{X}$,σ,Moy 8 : $\widetilde{X}$,d,Moy

FIGURE 7.4 – Résultats du scénario S3 : comparaison pour des bruits croissants en valeurs en en taille des cardinalités C et $\widetilde{X}$ pour la taille des groupes, des mesures d et σ pour leur dispersion et des fonctions d'agrégation min et moyenne pour le calcul de la périodicité.

contient les expériences utilisant un bruit croissant en valeurs et celle de droite un bruit croissant en taille. Les six graphiques du haut de `(a)` à `(f)` sont obtenus avec la forme Rectangle et les six du bas de `(g)` à `(l)` avec la forme Sinus.

Sur chaque graphique, trois courbes indiquent les résultats renvoyés par chaque méthode de regroupement : BL en rouge pour la méthode de regroupement γ_{BL}, W en bleu pour γ_W et ES en vert pour celle basée sur le score d'érosion γ_{ES}.

S3 Le scénario S3 utilise exclusivement la méthode de regroupement γ_{BL} avec les paramètres $t_v = 0,8$ et $t_m = 3\%$ et permet la comparaison des cardinalités C et $\widetilde{X}$ pour le calcul de la taille des groupes (cf. section 5.3.1 p. 105), de la déviation absolue moyenne d et de l'écart-type σ pour leur dispersion (cf. section 5.3.2 p. 106) et des fonctions d'agrégation min et moyenne pour le calcul de la périodicité π (cf. section 5.3.3 p. 108).

Les résultats de ce scénario sont donnés sur la figure 7.4. Les deux graphiques contiennent 8 courbes correspondants aux triplets décrits sous la figure issus des combinaisons des deux

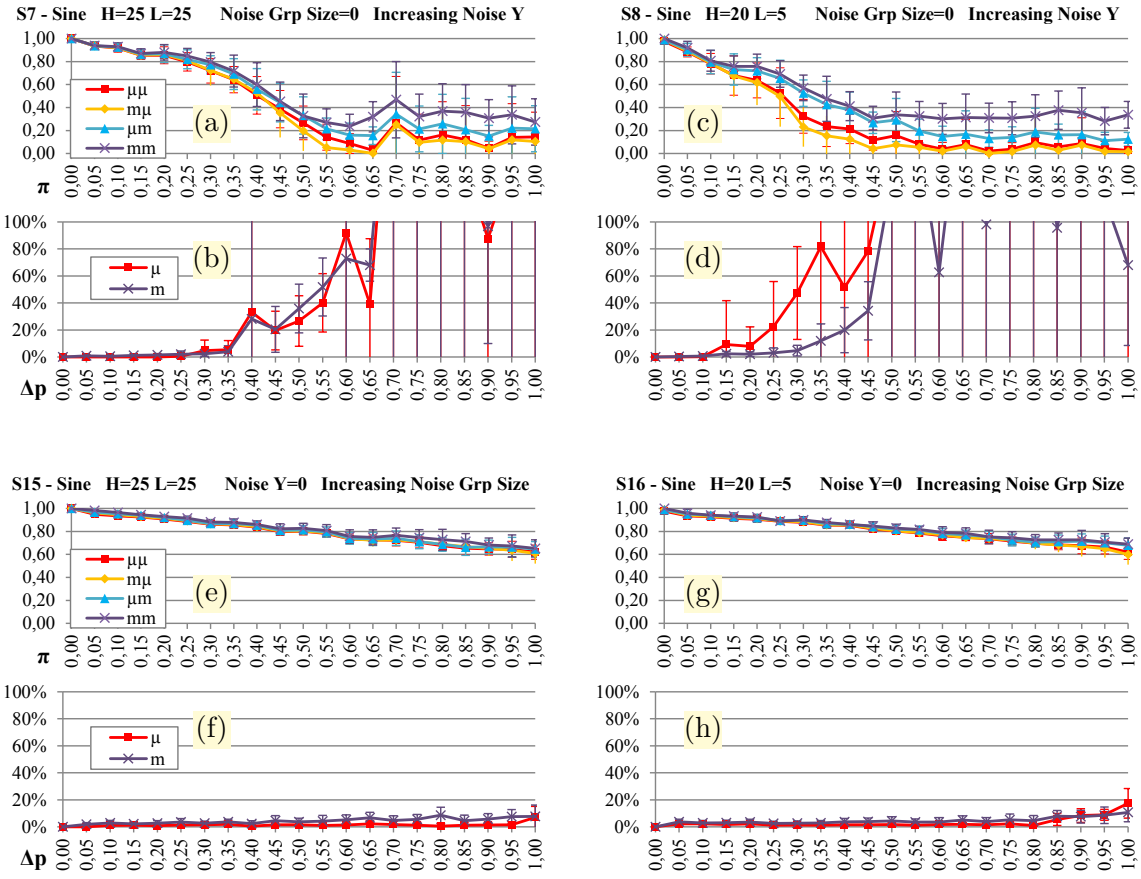


FIGURE 7.5 – Résultats du scénario S4 : comparaison des mesures de tendance centrale et de dispersion pour les tailles de groupes avec des bruits croissants en valeurs et en taille, deux types de formes et deux couples de tailles (p^H, p^L). Les notations $\mu\mu$, $m\mu$, μm et mm correspondent aux combinaisons moyenne / médiane décrites dans le tableau 5.1 p. 107.

cardinalités, des deux mesures de déviation et deux fonctions d'agrégation.

D'une manière générale sur cette figure, les lignes pointillées représentent les tailles de groupe mesurées avec $\widetilde{X}$, les lignes pleines celles mesurées avec C, les triangles représentent les combinaisons (σ, Min), les rectangles (d, Min), les étoiles (σ, Moy) et les croix (d, Moy).

Enfin, pour chaque valeur de bruit testée, les périodes candidates obtenues sont indiquées sous les graphiques.

S4 Le scénario S4 compare les mesures de tendance centrale et de dispersion pour les tailles de groupes basées sur des moyennes et des médianes décrites dans le tableau 5.1 p. 107, avec des bruits croissants en valeurs et en taille, deux formes et deux couples de tailles (p^H, p^L). Seule la méthode de regroupement γ_{es} est utilisée.

Comme pour le scénario S2, tous les graphiques obtenus ne sont pas représentés du fait du nombre important de jeux de données créés. Seules les combinaisons pour les deux couples de tailles de groupes (25,25) et (20,5), la forme Sinus et les bruits croissants en taille et en valeurs, significatives de l'ensemble, sont représentés sur la figure 7.5.

Les graphiques sont groupés par deux, celui du haut donnant la moyenne et l'écart-type de la périodicité et celui du bas ceux de l'erreur de calcul de la période Δp. Sur les graphiques du haut liés à la périodicité, quatre courbes sont présentes correspondant aux quatre combinaisons moyenne / médiane détaillées dans le tableau 5.1 p. 107. Sur ceux du bas liés à Δp, seules deux courbes sont affichées, liées à l'utilisation de la moyenne ou de la médiane comme mesure de tendance centrale de la taille de groupes. Seules ces deux variantes sont présentées car le calcul de la période n'est basé que sur la mesure de tendance centrale et n'utilise pas celle de dispersion (cf. éq. (5.14) p. 108).

Sur la colonne de gauche, les jeux de données sont générés avec des groupes H et L de taille 25 tandis que sur celle de droite les groupes H sont de taille 20 et les groupes L de taille 5.

7.2.3 Méthodes de regroupement

L'efficacité des différentes méthodes de regroupement est testée au travers des scénarios S1 comparant γ_{BL} avec γ_{es} et S2 intégrant de plus γ_W.

Le paragraphe suivant présente les résultats obtenus avec l'ensemble des méthodes tandis que les trois suivants détaillent les particularités de chacune d'entre elles.

Résultats communs à toutes les méthodes de regroupement

De manière générale, les résultats du scénario S2 illustrés sur la figure 7.3 p. 142 montrent que le type de bruit (taille ou valeurs) est le paramètre qui a le plus d'influence sur les résultats obtenus, suivi par la forme des séries (Rectangle ou Sinus). La taille des groupes ou l'utilisation simultanée des deux types de bruit n'induisent pas de comportement qualitativement différent.

Nous détaillons désormais les résultats au regard des cinq critères définis dans la section 7.2.1 p. 140. Tout d'abord, le critère C1 est vérifié par toutes les méthodes, i.e. la périodicité est toujours égale à 1 lorsque le bruit est nul. Ce résultat est explicable par la méthode de génération des données qui créé des séries constituées de motifs exactement répétés simplement détectables par des méthodes à seuil global ou local lorsqu'aucun bruit ne les perturbe. Lorsque les groupes de même taille sont correctement identifiés, la dispersion de leur taille est nulle et la périodicité vaut donc 1.

De manière peut-être plus surprenante, le critère C4 de bonne identification des étiquettes est également vérifié pour toutes les méthodes. Cela est dû au fait que *Acc* est calculé à partir du nombre de données bien classées *sans tenir compte de leur ordre*. Supposons par exemple une série strictement périodique composée de 5 groupes H et 5 groupes L de 20 points chacun pour laquelle la méthode de regroupement détecte par erreur un point L au milieu de chaque groupe H. En ce cas, seules 5 erreurs d'étiquetage sont réalisées sur 200 points, et $Acc = 97,5\%$. Au niveau des groupes en revanche, la méthode en identifie 10 de types H, de taille 9 ou 10, et 10 de type L dont 5 de taille 20 et 5 de taille 1. La régularité ρ^H des groupes L, et par conséquent la périodicité renvoyée,

est donc faible, alors que la série initiale est strictement périodique. Ainsi, même si *Acc* est élevé le calcul de la périodicité peut-être faussé. En ce sens, *Acc* n'est pas pertinent pour juger de la qualité d'une méthode de regroupement.

De plus, le bruit ν_s sur les tailles de groupes est peu discriminant pour les méthodes de regroupement. Comme illustré sur la figure 7.1 p. 135, ses effets sont limités et les groupes restent simplement identifiables par l'ensemble des méthodes de regroupement. Même pour des valeurs élevés de ν_s, les cinq critères sont globalement vérifiés. Seule la méthode γ_W réalise des erreurs d'évaluation de la période pour $\nu_s > 0,7$ avec des rectangles (cf. figure 7.3 p. 142 `(e)`) pour des raisons expliquées plus bas dans le paragraphe dédié à cette méthode.

Concernant les effet du bruit en valeurs ν_v, le critère C2 est vérifié pour les méthodes γ_W et γ_{es} : π décroît lorsque le bruit augmente. Il est également vérifié pour γ_{BL}, dans tous les cas pour le scénario S1 et uniquement pour les valeurs de bruits inférieures à 0,5 dans le cas de Rectangles et 0,2 dans le cas de Sinus pour le scénario S2. Ces résultats sont justifiés dans le paragraphe suivant.

Le critère C3 est approximativement respecté quand $\nu_v < 0,4$: l'erreur d'estimation de la période Δp reste très faible dans ce cas. A partir de ce seuil néanmoins, les méthodes renvoient des estimations erronées.

Enfin, le critère C5 sur l'écart-type des valeurs renvoyées pour un même niveau de bruit est étudié dans les paragraphes suivants car les résultats des différentes méthodes sont variables pour ce critère.

Méthode de référence γ_{BL}

La méthode γ_{BL} est la plus sensible au bruit conformément à nos attentes. Il est intéressant de voir cependant que cette sensibilité est nettement diminuée dans le scénario S1 grâce l'utilisation du paramètre de fusion des groupes adjacents (cf. section 5.2.3 p. 103) : les résultats de la méthode γ_{BL} (légende B) sont alors comparables à ceux de la méthode γ_{es}. Comme illustré sur la figure 7.2 p. 141 `(a)`, le comportement de γ_{BL} (légende B) est comparable à la méthode γ_{es} (légende M). Le paramètre de fusion a toutefois été fixé afin de bien fonctionner sur les jeux de données générés, mais les expériences de S2 sans ce paramètre montrent que γ_{BL} est inefficace. D'autre part, comme discuté plus bas avec l'utilisation de la moyenne mobile pour la méthode γ_W, les paramètres utilisés sont sensibles et diminuent considérablement la robustesse de ces approches.

La suite de l'analyse porte sur le scénario S2, plus complet pour la comparaison des méthodes de regroupement. Dans ce scénario, la méthode γ_{BL} est la moins robuste. En effet, avec la forme rectangle (figure 7.3 p. 142 `(a)`, `(b)`), π décroît rapidement et Δp augmente subitement dès que ν_v atteint 0,5. Ces changements brusques sont liés au fait que le paramètre de seuil t_v est fixé à 0,5 : tant que le bruit en valeurs ν_v est inférieur à 0,5, les valeurs générées pour les groupes H sont dans]0,5 ;1] et celles des groupes L dans [0 ;0,5[, donc γ_{BL} les détecte exactement si bien que le degré de périodicité vaut 1 et Δp est nul. Cette absence d'écart est intéressante, mais la constance du degré de périodicité est

problématique car elle signifie que la méthode ne fait aucune différence entre un jeu de données sans bruit et un jeu bruité.

De plus, dès lors que ν_v est supérieur ou égal à 0,5, les courbes changent brusquement car γ_{BL} identifie des groupes H et L de petites tailles au milieu de groupes de type opposé, augmentant ainsi la dispersion des tailles de groupes et réduisant nettement la valeur de π. Il est intéressant de constater que pour des valeurs de $\nu_v > 0,5$, π augmente jusqu'à atteindre 0,5 lorsque $\nu_v = 1$. Pour ce niveau de bruit en effet et pour la forme Rectangle, la valeur d'un élément est purement aléatoire et suit une loi $\mathcal{U}(0;1)$ d'espérance $\mu = 1/2$. Concernant son écart moyen, il est calculé par :

$$\mathbb{E}\left[|X-\mu|\right] = \int_0^1 |x-\mu|\,dx = \int_0^{0,5} -(x-0,5)\,dx + \int_{0,5}^1 (x-0,5)\,dx = \frac{1}{4}$$

Donc $\rho^H = \rho^L = 1 - \frac{1/4}{1/2} = 1/2$ d'où $\pi = 1/2$, conformément à ce qu'indique le graphique.

Comparaison de γ_{es} et γ_W

D'une manière générale, γ_{es} est plus souple que γ_W car elle varie de manière moins abrupte du fait de l'utilisation d'un seuil local déterminé à partir des données. De plus, toujours dans le cadre du scénario S2, elle est généralement plus précise dans l'évaluation de la période (11 cas sur 16) et dans le taux de bonne classification (12 cas sur 16). Enfin, elle est plus robuste car ses valeurs d'écart-types sont plus faibles.

Concernant l'estimation de la période avec l'augmentation du bruit en taille, il est intéressant de noter que lorsque $\nu_s > 0,7$, γ_W est la seule à faire des erreurs assez importantes (figure 7.3 p. 142 `(e)`). Ce phénomène est lié à l'utilisation d'une moyenne mobile en prétraitement (cf. section 5.2.3 p. 103) qui peut réduire de manière drastique les groupes très fins qui apparaissent lorsque le bruit en taille est important et/ou lorsque la taille de la fenêtre est mal choisie. Ainsi, lorsqu'un groupe H ou un groupe L n'est pas identifié durant le regroupement, un groupe H ou un groupe L particulièrement grand est identifié, entraînant un biais considérable dans l'évaluation de la taille des groupes et donc dans celle de la période. La figure 7.6 illustre ce phénomène : pour γ_W, le seuil est mal placé suite à la diminution de pics par application de la moyenne mobile et seuls deux groupes H sont détectés.

Le point faible de γ_W réside donc dans l'utilisation d'un seuil global qui biaise toute l'évaluation lorsqu'il est mal déterminé. γ_{es} peut commettre des erreurs de regroupement, mais son seuil local permet de ne pas les propager. Ainsi, les évaluations de la période sont meilleures en moyenne avec γ_{es} qu'avec γ_W pour les données très bruitées en valeur (figure 7.3 p. 142 `(b)`, `(h)`). La figure 7.7 illustre un tel jeu de données : γ_{es} identifie mal certains groupes mais γ_W les manque tous.

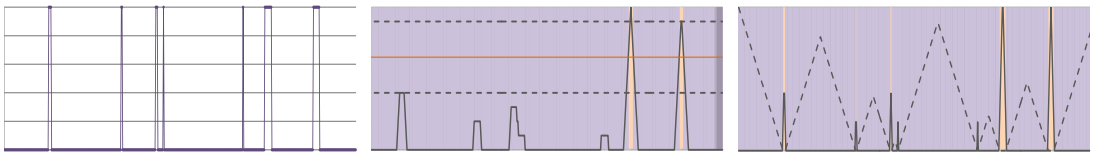


FIGURE 7.6 – Comportement de γ_W et γ_{es} avec de petits groupes. Sur fond clair (resp. sombre), les groupes identifiés comme H (resp. L). Pour γ_W, le trait plein est obtenu par passage d'une moyenne mobile sur les données et le trait rouge représente le seuil retenu. Pour γ_{es}, le trait plein est le score d'érosion de X et le trait en pointillés celui de $\overline{X}$.

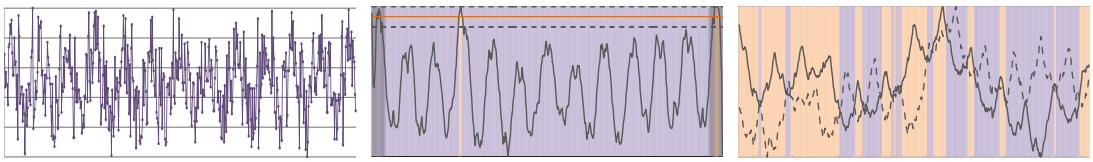


FIGURE 7.7 – Comportement de γ_W et γ_{es} avec des données sinusoïdales très bruitées en valeurs, légende similaire à celle de la figure 7.6

Méthode γ_{es}

La figure 7.7 montre également un des biais de γ_{es} dans le cas de données très bruitées, particulièrement en forme de Sinus. Les scores d'érosion de X et $\overline{X}$ qui servent de base au regroupement ne se croisent pas au moment où ils l'auraient dû, comme dans la partie droite du graphe où des groupes H sont ignorés. Cependant, le score de X croît et celui de $\overline{X}$ décroît aux endroits où un groupe H aurait dû être détecté, et inversement aux endroits où un groupe L aurait dû être détecté.

Ceci est dû au fait que le score d'érosion est très sensible au contraste des données initiales, i.e. au fait ce que les valeurs hautes atteignent 1 et que les valeurs basses redescendent en 0 à chaque cycle. Même si ces dernières sont normalisées avant calcul du score, il est possible qu'une seule valeur atteigne 0 et que les valeurs basses suivantes ne redescendent pas assez bas, en 0,1 par exemple. En ce cas, il faut de multiples érosions pour qu'elles atteignent finalement 0 (voir la section 6.1.2 p. 114). Comme à chaque nouvelle érosion la valeur de l'érodé est ajoutée aux valeurs précédentes, le score d'érosion en ce point est artificiellement élevé. Lorsque le score d'érosion est normalisé en fin de calcul, seule une valeur est égale à 0 et les autres sont élevées en rapport. Ce phénomène pour les valeurs basses se retrouve aussi pour les valeurs hautes lors du calcul du score d'érosion de $\overline{X}$. Des solutions à ce problème sont discutées en perspective de cette thèse.

7.2.4 Évaluation de la taille des groupes

La sous-section précédente contient une discussion de la première étape de DPE, le regroupement. Dans cette sous-section et les suivantes, nous détaillons les résultats obtenus avec les différentes variantes proposées pour la seconde étape de DPE, i.e. le calcul de la taille des groupes, de leur tendance centrale, de leur dispersion et enfin de la période et

de la périodicité de la série.

Concernant le calcul de la taille des groupes, plusieurs variantes sont proposées dans la section 5.3.1 p. 105. Ces dernières utilisent différents schéma de pondération pour mesurer la taille d'un groupe en fonction des valeurs qui le composent.

Parmi celles-ci, les schémas C et $\widetilde{X}$ sont comparées dans les scénarios S1 et S3. Le premier est égal au nombre d'éléments du groupe indifféremment de leurs valeurs et le second mesure la taille d'un groupe comme la somme des valeurs de ses éléments. Cette approche est similaire à σ-count dans le contexte de sous-ensembles flous (cf. section 1.2.5 p. 13). Les autres schémas présentés dans la section 5.3.1 p. 105 ne sont pas étudiés de manière aussi détaillée car peu différents de $\widetilde{X}$, comme nous ont confirmé les tests que nous avons réalisés avec ces derniers.

Le schéma C renvoie les meilleurs résultats pour DPE. Même si les deux renvoient des résultats équivalents pour le calcul de la périodicité π comme l'illustrent les résultats de S1 sur la figure 7.2 p. 141 `(a)` et de S3 sur les deux graphes de la figure 7.4 p. 143, le calcul de la période p_c est plus précis avec C comme le montre la figure 7.2 p. 141 `(b)`. En effet, si le schéma $\widetilde{X}$ similaire à σ-count capture bien la notion d'appartenance à un sef, elle est moins adaptée à la mesure de la distance entre le début et la fin du groupe, qui est à l'inverse bien appréhendée par C puisque tous les éléments ont le même poids. La période étant précisément définie comme la distance entre deux motifs identiques, cette dernière est donc mieux évaluée avec C.

De plus, la possibilité pour une donnée d'appartenir à deux types de groupes qu'offre $\widetilde{X}$ n'est pas utile car le regroupement en groupes H et L est effectué avant le calcul de leur cardinalité et ces dernières n'appartiennent finalement qu'à un seul type de groupe.

7.2.5 Tendance centrale de la taille des groupes

La mesure de tendance centrale μ de la taille des groupes a une influence sur la période et la périodicité calculées, comme détaillé dans la section 5.3.2 p. 106. Le scénario S4 propose de comparer deux variantes de cette mesure, l'une basée sur la moyenne et l'autre sur la médiane.

Le graphe `(d)` de la figure 7.5 p. 144 qui illustre les résultats de ce scénario pour le calcul de la période montre une meilleure performance de la médiane sur la moyenne pour le calcul de la taille des groupes. Ce cas correspond à des groupes H larges (20 points en moyenne) et des groupes L étroits (5 points en moyenne). Dans ce cas de figure en effet, deux groupes H très proches pour lesquels les valeurs du groupe L intermédiaire ne sont pas suffisamment faibles peuvent être interprétés comme un seul groupe H très grand par la méthode γ_{es}, comme illustré sur la figure 5.8 p. 107. Ce groupe affecte par la suite le calcul de la taille moyenne, mais pas celui de la taille médiane dont la robustesse est une caractéristique classique, cf. l'étude approfondie sur les statistiques robustes de Rousseeuw & Croux (1993).

Dans la méthode DPE cependant, différentes raisons nous ont conduit à conserver la moyenne pour le calcul des tailles de groupes. Tout d'abord, elle n'est moins efficace que la

médiane que dans certains cas particuliers, et pas de manière excessive. D'autre part, elle est nettement plus simple à calculer, surtout dans un contexte incrémental fenêtré. Enfin, les meilleurs résultats renvoyés par la médiane sont liés à un défaut du score d'érosion dans certains cas. Nous avons donc privilégié l'amélioration du calcul de ce score plutôt que l'utilisation d'une technique destinée à en masquer les faiblesses. L'étude de la médiane pour le calcul de la valeur centrale des tailles peut toutefois constituer une perspective à cette thèse.

7.2.6 Dispersion de la taille des groupes

Comme rappelé dans la sous-section précédente, la mesure de dispersion sert avec celle de tendance centrale à calculer le régularité de la taille des groupes et donc la périodicité de la série. La dispersion n'a toutefois pas d'influence sur le calcul de la période.

Dans le scénario S3, la déviation absolue moyenne définie par l'éq. (5.13) p. 106 et l'écart-type non biaisé $\sigma = \sqrt{1/(n-1)\sum(x_i - \mu)^2}$ sont comparés. Les résultats illustrés sur la figure 7.4 p. 143 montrent que l'écart-type σ est plus sensible au bruit que la déviation moyenne absolue d, ce qui est un résultat connu d'un point de vue théorique. A l'usage, il nous est apparu plus pertinent d'utiliser d plutôt σ dans le calcul de la régularité des tailles de groupes du fait de sa stabilité. Au regard du critère C2, le caractère « régulier » de la décroissance de π avec le bruit est donc mieux vérifié avec d.

Le scénario S4 propose également la comparaison des différentes mesures de dispersion définies dans le tableau 5.1 p. 107. Les résultats illustrés sur la figure 7.5 p. 144 montrent que les calculs de périodicité sont équivalents et ne dépendent donc pas de la mesure de dispersion retenue. Ainsi, la déviation absolue moyenne est sélectionnée pour DPE.

7.2.7 Périodicité

Le calcul de la périodicité repose sur l'agrégation des régularités des tailles des groupes H et L, comme indiqué dans l'éq. (5.14) p. 108. Deux fonctions d'agrégation, le min et la moyenne, sont comparées dans le scénario S3.

Le min représentant la conjonction logique, son usage aurait pu être pertinent dans l'interprétation "*les données sont périodiques si les tailles de groupes H* et *celle des groupes L sont régulières*". En pratique, ce dernier est trop strict et renvoie des degrés de périodicité π trop faible par rapport à la moyenne, comme illustré par les courbes 1, 2, 5 et 6 de la figure 7.4 p. 143. L'agrégation par la moyenne, plus « optimiste », est donc retenue.

En conclusion de cette étude expérimentale, les résultats détaillés ci-dessus montrent que l'approche de DPE est pertinente puisque les cinq critères décrits dans la section 7.2.1 p. 140 sont globalement vérifiés pour l'ensemble des variantes avec des niveaux de bruit raisonnables. Néanmoins, certaines combinaisons de variantes sont meilleures que les autres. Nous retenons donc la meilleure d'entre elles pour spécifier DPE, qui est la combinaison de γ_{es} pour la méthode de regroupement avec la cardinalité crisp C pour le calcul des tailles, la moyenne pour la mesure de leur tendance centrale, la déviation absolue moyenne

pour leur dispersion et la moyenne pour le calcul de la périodicité.

7.3 Étude expérimentale de la performance des méthodes de calcul du score d'érosion

L'étude expérimentale de la performance se concentre sur la comparaison en termes de temps d'exécution et d'occupation mémoire des méthodes de calcul du score d'érosion dans ses implémentations naïves, par niveaux, incrémentale et incrémentale par niveaux présentées dans le chapitre précédent section 6.1 p. 113.

Les critères de qualité dans ce cadre sont présentés dans la section 7.3.1, puis le protocole expérimental dans la section 7.3.2 et enfin les résultats obtenus dans la section 7.3.3.

7.3.1 Critères de qualité

Les performance sont évaluées sur le temps d'exécution et l'utilisation mémoire des différentes approches pour le calcul du score d'érosion. L'occupation mémoire est déterminée par la taille des structures utilisées pour chacun des algorithmes plutôt que par leur occupation mémoire réelle, cette dernière étant potentiellement moins précise car dépendante du système d'exploitation utilisé et plus complexe à obtenir.

7.3.2 Protocole

Comme la complexité des méthodes de calcul du score d'érosion est directement fonction de la taille n des données, la comparaison de leur performance est réalisée sur des jeux de données de taille croissante. Leur génération est décrite par le scénario SP, défini dans le tableau 7.2 p. 139, qui présente l'avantage d'en produire un nombre important selon des paramètres variés, garantissant ainsi que les résultats de performance mesurés ne sont pas propres à telle ou telle de leurs caractéristiques mais bien à la méthode étudiée.

Le temps d'exécution est mesuré depuis l'appel à la méthode jusqu'au retour du résultat. L'occupation mémoire est dépendante du type de méthode utilisée, selon qu'elle utilise ou non une structure pour stocker les indices clés λ décrits dans la section 6.1.3 p. 114 dédiée au score d'érosion par niveaux.

L'implémentation est réalisée en VB.NET et les expériences ont lieu sur une machine virtuelle Windows configurée avec 4 CPUs et 4 Go de mémoire vive sur une machine physique équipée d'un CPU Intel i7 et de 16 Go de mémoire vive.

7.3.3 Résultats

La figure 7.8 p. 153 illustre les temps de calculs moyens avec leur écart-type pour les quatre méthodes étudiées avec plusieurs tailles de jeux de données. Les tailles ont été ajustées en fonction de l'efficacité des méthodes testées, afin d'éviter qu'une méthode ne soit trop longue à terminer du fait d'un jeu de données trop grand.

La figure montre sur le graphique du haut que les méthodes fonctionnent en un temps raisonnable pour de petits jeux de données de moins de 10 000 points, sur celui du milieu que seules les deux méthodes incrémentales sont rapides pour des jeux de données moyens de moins de 100 000 points et enfin sur celui du bas que seule la méthode incrémentale par niveaux est acceptable pour des jeux de données d'un million de points.

La figure illustre clairement la supériorité des méthodes incrémentales sur les non incrémentales. La méthode naïve est significativement plus lente et moins robuste comme le montrent les écarts-types importants enregistrés. Dans une situation réelle, les méthodes incrémentales sont encore plus rapides en comparaison puisqu'elles permettent d'intégrer un nouveau point en un temps négligeable tandis que les autres nécessitent un recalcul sur l'ensemble du nouveau jeu de données.

La comparaison des méthodes incrémentales montre la supériorité de celle par niveaux sur la méthode incrémentale simple. Là encore, le résultat est vérifié sur les temps moyens de calcul comme sur les écarts-types, indiquant la moins grande stabilité en temps d'exécution de la méthode incrémentale simple.

Enfin, le graphique du bas illustre l'efficacité de la méthode incrémentale par niveaux sur de grands jeux de données, entre 100 000 et 1 000 000 de points. Il en ressort que cette méthode peut traiter 1 million de nouveaux points en 1,5 seconde.

7.3.4 Discussion

La discussion des résultats porte d'abord sur les temps de calcul puis sur l'utilisation mémoire des différentes méthodes.

Temps de calcul et complexité

Nous établissons dans ce paragraphe la représentation formelle de la complexité des méthodes en fonction des jeux de données utilisés afin de la comparer aux résultats obtenus expérimentalement. La figure 7.9 montre les courbes représentant les expressions analytiques déterminées, similaires à celles illustrées sur la figure 7.8 issues des expériences.

La complexité des méthodes non incrémentales est présentée dans un premier temps, suivie de celle de la méthode incrémentale simple.

Méthodes non incrémentales La complexité des méthodes non incrémentales sont $O(n \times \max z_i)$ pour la méthode naïve (cf. section 6.2.2 p. 124) et $O(\sum z_i)$ pour celle par niveaux (cf. section 6.2.3 p. 125) et dépendent toutes deux de la distance z_i du zéro le plus proche du point x_i.

Afin d'en étudier la complexité plus finement, nous détaillons la valeur de ces expressions en intégrant les paramètres utilisés pour la génération des données. Nous définissons également $p = p^H + p^L$ la taille du motif répété, $g = n/p$ le nombre de ses répétitions, égal au nombre de groupes H ou L, et supposons que même lorsque du bruit sur la taille des groupes ν_s est utilisé, la taille moyenne des groupes H est p^H et celles groupes L est p^L.

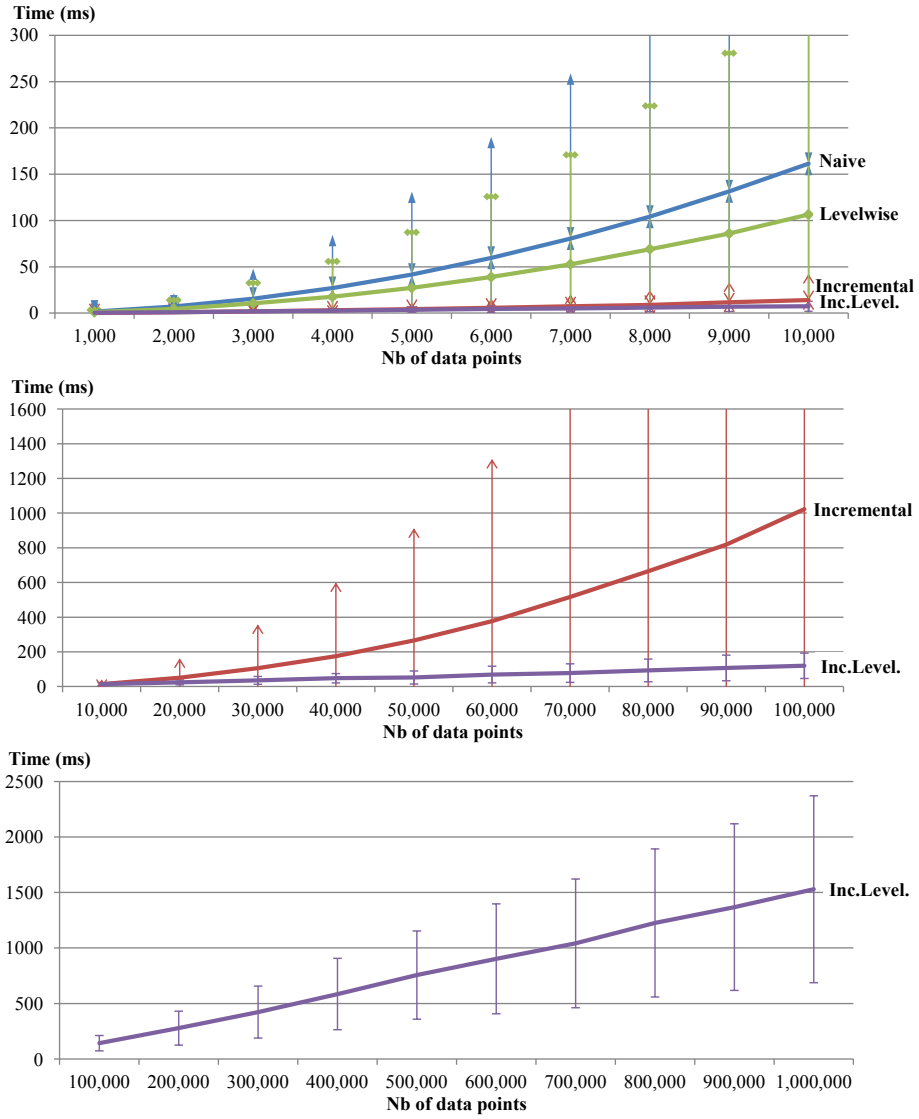


FIGURE 7.8 – Temps de calcul, en haut, des quatre méthodes sur de petits jeux de données, au milieu, des méthodes incrémentales sur des jeux de données moyens et en bas de la méthode incrémentale par niveaux sur de grands jeux de données

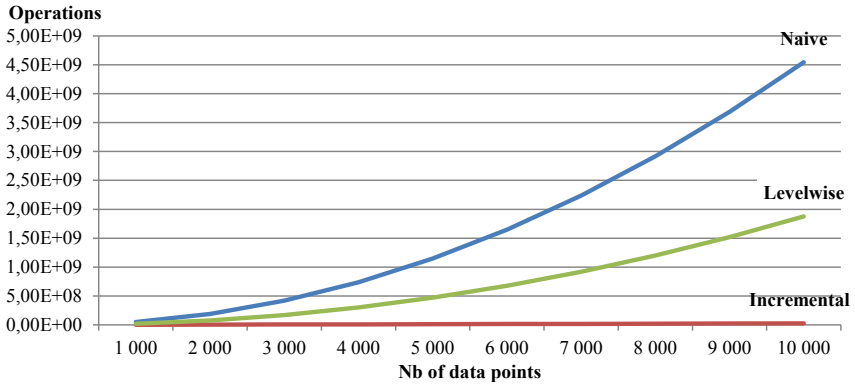


FIGURE 7.9 – Complexité analytique du calcul du score d'érosion avec les méthodes naïve, par niveaux et incrémentale.

Tableau 7.3 – Complexité des méthodes non incrémentales en fonction des paramètres de génération des données

Forme	ν_v	$n \times \max z_i$	ζ	$\sum z_i$	Nb
Triangle, Vague	≥ 0	$np^H/2$	p^H	$gp^H(p^H+2)/4$	80
Rectangle,Sinus	> 0	$3n^2/4$	-	$(5n^2+8n)/16$	60
Rectangle	$= 0$	np^H	p^H	$gp^H(p^H+2)/4$	10
Sinus	$= 0$	$n(p^H+p^L/2)$	$p/2$	$gp(p+2)/16$	10

Les expressions de $n \times \max z_i$ et $\sum z_i$ en fonction des paramètres de génération sont synthétisées dans le tableau 7.3. Ces résultats sont détaillés ci-dessous après une exposition des principes de raisonnement permettant de les établir.

Principes Nous considérons dans un premier temps des données générées sans bruit composées de valeurs supérieures à 0 pour les groupes H et égales à 0 pour les groupes L. Nous définissons ζ la plus petite distance entre deux groupes de zéros successifs. Comme illustré sur la figure 7.1 p. 135, $\zeta = p^H$ pour les formes Rectangle, Triangle et Vague et $\zeta = p$ pour la forme Sinus qui ne dispose que d'un zéro au milieu des groupes L.

Ainsi, $\max z_i = \zeta/2$ pour un groupe H quelconque dans le jeu de données, i.e. le point le plus éloigné d'un zéro est au milieu du groupe H. Cependant, comme la génération débute toujours par un groupe H, $z_i = \zeta$ pour le premier point de ce groupe car il ne dispose pas de zéro à un indice inférieur. Ainsi, lorsque le premier groupe H ne contient pas de valeur nulle, $\max z_i = \zeta$.

L'établissement de $\sum z_i$ suit une logique légèrement différente. En ce cas en effet, nous ignorons la particularité liée au premier groupe car les z_i de tous les groupes sont considérés. Au sein d'un groupe H, $z_i = 1$ pour le premier point, $z_i = 2$ pour le second et ainsi de suite jusqu'au point d'indice $\zeta/2$. Après ce point, les z_i décroissent jusqu'à valoir 1 pour la dernière valeur du groupe. Ainsi, la somme des z_i pour un groupe H est égal à $2\sum_{i=1}^{\zeta/2} i = \zeta(\zeta+2)/4$. Cette somme est multipliée par le nombre de groupes H, noté g, d'où $\sum z_i = g\zeta(\zeta+2)/4$.

Dans les paragraphes suivants, ces expressions sont adaptées selon les formes utilisées.

Triangle et Vague Les jeux de données générés avec ces deux formes ont la particularité de posséder des 0 espacés de p^H points indépendamment du bruit en valeurs ν_v. Cette propriété découle directement des équations de la section 7.1.2 p. 136 pour $\nu_v = 0$ mais se vérifie également avec $\nu_v > 0$ du fait d'un effet de bord de la méthode de génération. En effet, la première valeur des groupes H pour ces formes est égale à 0 avant l'application du bruit, mais comme la valeur de bruit est soustraite à la valeur initiale pour les groupes H (cf. éq. (7.6) p. 137), la première valeur pour ces groupes reste égale à 0 après normalisation. Cette particularité est également vérifiée pour le premier groupe, d'où $\zeta = p^H$, $\max z_i = p^H/2$ et $\sum z_i = gp^H(p^H+2)/4$.

Rectangle et Sinus Contrairement au Triangle et à la Vague, les zéros des jeux de données générés pour le Rectangle et le Sinus dépendent de ν_v. Nous étudions d'abord le cas $\nu_v = 0$ pour le Rectangle puis pour le Sinus, puis le cas commun aux deux formes où $\nu_v > 0$.

Avec $\nu_v = 0$ et la forme Rectangle, les zones de zéros sont séparés de p^H points, donc $\zeta = p^H$ et $\sum z_i = gp^H(p^H + 2)/4$, à l'instar des formes Triangle et Vague. En revanche, le premier point du premier groupe n'est pas égal à 0, donc $\max z_i = p^H$.

Pour $\nu_v = 0$ avec la forme Sinus, les données ne contiennent qu'un zéro au milieu de chaque groupe L, d'où $\zeta = p = p^H + p^L$ et $\sum z_i = gp(p + 2)/16$ Concernant le premier point du premier groupe, il est séparé de $p^H + p^L/2$ point du zéro du groupe L, d'où $\max z_i = p^H + p^L/2$.

Lorsque $\nu_v > 0$, nous considérons que la complexité pour ces deux formes est similaire car elles ne disposent que d'un seul 0 dans tout le jeu de données. En effet, comme $\nu_v > 0$, une valeur aléatoire non nulle est ajoutée à chaque point des groupes L si bien que ces derniers sont composés de valeurs strictement positives et presque certainement différentes. Ainsi, après la troisième étape de normalisation dans [0,1], seule la valeur la plus faible à l'étape précédente est égale à 0.

Lorsqu'une seule valeur nulle est présente à l'indice i^* alors $\max z_i = \max(i^*, n - i^*)$ et $\sum z_i = \sum_{i=1}^{i^*} i + \sum_{i=1}^{n-i^*} i$. Le cas minimisant ces expressions correspond à $i^* = n/2$ tandis que celui les maximisant est $i^* = n$. Nous supposons donc le cas moyen $i^* = 3n/4$ pour ces formes lorsque $\nu_v > 0$. Ainsi, $\max z_i = 3n^2/4$ et $\sum z_i = (5n^2 + 8n)/16$.

Conformément aux résultats expérimentaux illustrés sur la figure 7.8 p. 153 et analytiques sur la figure 7.9 p. 153, nous vérifions qu'indépendamment des formes utilisées et de leur bruit, les expressions de complexité de la méthode par niveaux sont inférieures à celles de la méthode naïve : $(5n^2 + 8n)/16 < 3n^2/4$ dès que $n > 2$ et $gp^H\left(p^H + 2\right)/4$ est inférieur à $np^H/2$ car $n \gg g$ d'une manière générale et p^H et g évoluent de manière inverse, i.e. plus les groupes H sont grands moins il y en a pour n donné donc plus g est petit, et inversement.

Méthodes incrémentales Nous incluons dans notre étude analytique la complexité de la méthode incrémentale simple dont l'expression $O(\log^2 n)$ est donnée dans la section 6.2.4 p. 126, mais pas celle de la méthode incrémentale par niveaux, non établie.

Dans la mesure où cette étude a pour but de comparer les quatre approches de calcul du score d'érosion, l'absence de cette dernière n'est pas significative car, pour les tailles de jeux de données sur lesquelles les méthodes sont comparables, soit entre 1 000 et 10 000 points, les performances des deux méthodes incrémentales sont confondues, comme illustré sur la figure 7.8 p. 153.

Représentation graphique Afin de représenter graphiquement les complexités établies, nous les évaluons pour des valeurs de n entre 1 000 à 10 000 par pas de 1 000 et les pondérons par le nombre de cas où les combinaisons de paramètres spécifiques sont utilisées dans le scénario SP, indiqué dans la colonne « Nb » du tableau 7.2 p. 139. Sachant que

cinq pas de valeurs sont utilisés pour les paramètres évolutifs, il y a par exemple 80 cas où les formes Triangle ou Vague sont utilisées avec $(p^H, p^L) = (90, 10)$ ou $(p^H, p^L) = (50, 50)$ et $\nu_v > 0$ ou $\nu_v = 0$.

Nous représentons ainsi sur la figure 7.9 p. 153 le nombre d'opérations réalisées par les différentes méthodes pour le calcul du score d'érosion dans le cadre du scénario SP.

Cette figure montre que les expressions de complexité sont satisfaisantes car les tendances observées expérimentalement sur le graphique du haut de la figure 7.8 p. 153 sont vérifiées. Certaines différences sont toutefois visibles, en particulier concernant la méthode incrémentale qui apparaît nulle sur l'ensemble des tailles de jeux de données alors qu'elle atteint une valeur faible dans les résultats expérimentaux. Cette légère différence est explicable par le fait que l'expression utilisée pour la représenter ne prend en compte que le nombre de boucles réalisées sans tenir compte de la complexité de ces boucles, plus importante pour la méthode incrémentale simple que pour celles naïves ou par niveaux, comme l'illustrent les algorithmes 6.1 p. 125, 6.2 p. 125 et 6.3 p. 126. Le même argument pourrait également expliquer la différence entre la méthode par niveaux et la méthode naïve, plus importante dans le cas de la représentation analytique que dans les expériences réelles.

Utilisation mémoire

Comme mentionné dans la section 7.3.2 p. 151 l'occupation mémoire dépend de l'utilisation de la matrice λ pour stocker les indices clés des points. En pratique, les implémentations données dans la section 6.2 p. 123 montrent que seule la méthode incrémentale par niveaux l'utilise. Elle est donc la seule dont l'occupation mémoire soit variable, i.e. dépendante du jeu de données, car basée sur λ et n. Les autres méthodes ont une occupation mémoire constante fonction de n uniquement.

Plus précisément, ces dernières stockent un exemplaire du jeu de données et un exemplaire des scores d'érosion, soit $2n$ éléments.

La méthode incrémentale par niveaux stocke les valeurs λ_{il} en plus des données et des des scores d'érosion. Dans la solution développée, les λ_{il} sont stockés sous forme d'une liste de n listes de ω_i valeurs. L'utilisation mémoire de cette structure $\sum \omega_i = \overline{\omega}_i \times n$, où $\overline{\omega}_i$ est la valeur moyenne des ω_i, qui est la valeur renvoyée lors des expériences.

Les résultats obtenus sont $\overline{\omega}_i = 30$, $\min(\omega_i) = 2$ et $\max(\omega_i) = 129$, ainsi la taille moyenne de la structure est $30n$ valeurs. La consommation totale de mémoire est donc $\sum \omega_i + 2n$, soit 32 millions de valeurs pour les jeux de données d'un million de points.

Comme la structure stockant les λ contient les chaînes d'indices des valeurs inférieures les plus proches, sa taille est directement dépendante de la forme de données. Le pire cas est celui où les données sont strictement monotones. En ce cas, $\omega_i = i$ et la taille totale de la structure est $n(n+1)/2$, donc quadratique. Le meilleur cas correspond à une série de valeur constante avec une valeur 0. En ce cas, λ_{il} ne contient que l'indice de cette valeur, $\omega_i = 1$ et la taille de la structure est alors linéaire égale à n.

Le désavantage des méthodes par niveaux est donc leur consommation mémoire, qui toutefois n'a pas empêché leur bonne exécution sur une machine de bureau standard. De

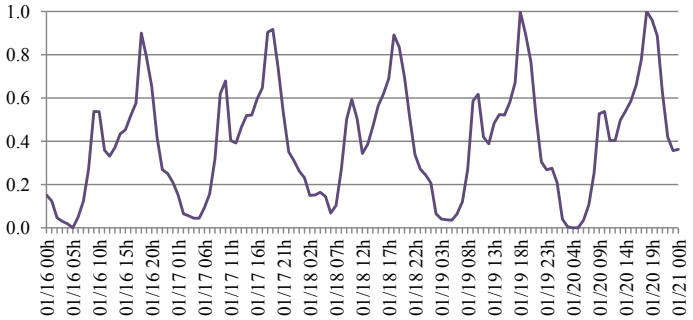


FIGURE 7.10 – Quantité de CO_2 par heure du 16/01/2012 au 21/01/2012 à la station Châtelet (RATP, 2012).

plus, l'utilisation des structures alternatives l et r mentionnée dans la section 6.1.3 p. 117 permet théoriquement de ramener leur occupation mémoire à $2n$. Le développement de ces approches est proposé en perspectives de cette thèse.

7.4 Application à des données réelles

Cette section décrit l'utilisation de DPE pour générer un résumé linguistique décrivant la périodicité de données réelles. Les données utilisées sont les mesures horaires de la quantité de CO_2 à la station de métro Châtelet, mises à disposition sur la plate-forme *open data* de la RATP [1], l'opérateur du métro parisien.

La figure 7.10 illustre ces données entre le 16 et le 21 janvier 2012. Un motif périodique quotidien présent du lundi au vendredi et composé de deux pics à 9h et à 18h est présent. Ces deux pics correspondent aux horaires de bureau en semaine.

DPE détecte la périodicité du motif, pas sa forme, et les deux pics sont simplement considérés comme un groupe de valeurs hautes. Leur identification spécifique pourrait être réalisée avec une des méthodes par actogramme présentée dans la section 4.2.3 p. 75, par transformée de Fourier présentée dans section 4.3.1 p. 78, auquel cas les phases des sinusoïdes de plus forte puissance devraient correspondre aux deux pics, ou encore à l'aide d'une des méthodes symboliques détaillées dans la section 4.5.2 p. 90 susceptibles de renvoyer un motif contenant des caractères correspondants aux pics.

La période de référence p_{ref} est 24h et le protocole du scénario S1 est utilisé, à savoir une comparaison entre les méthodes de regroupement γ_{es} et γ_{BL} avec les paramètres $t_v = 0,7$ et $t_m = 8\%$ et les cardinalités C et $\widetilde{X}$ pour le calcul de la taille des groupes. Le tableau 7.4 présente les résultats obtenus.

Résultats Conformément à nos attentes, le degré de périodicité est élevé, entre 0,72 et 0,86. La méthode de regroupement γ_{BL} renvoie des degrés de périodicité plus faibles que γ_{es}, confortant le fait que cette dernière est plus pertinente pour DPE car les données étudiées sont très périodiques et doivent être associées à une périodicité élevée.

1. `http://data.ratp.fr/`

Tableau 7.4 – Résultats obtenus sur les données réelles de la figure 7.10.

Méthode	**Cardinalité**	π	**Période**	**Phrase générée**
γ_{BL}	C	0,73	20,60 h	La période est environ 20 heures
γ_{BL}	$\widetilde{X}$	0,72	17,05 h	La période est exactement 17 heures
γ_{es}	C	0,82	24,20 h	La période est exactement 1 jour
γ_{es}	$\widetilde{X}$	0,86	17,05 h	La période est exactement 17 heures

De plus, la seule combinaison détectant la période de 24h est celle basée sur γ_{es} et la cardinalité C, ce qui confirme sa supériorité sur la cardinalité $\widetilde{X}$, en accord avec les résultats des expériences sur les données synthétiques. Il convient d'ailleurs de noter que l'évaluation de la période est très précise puisque l'erreur effectuée est de $|24,20-24|/24 = 0,8\%$, nettement moins élevée que celles réalisées par les autres méthodes qui s'échelonnent entre 14,2% et 29,0%.

Le fonctionnement du rendu linguistique est également visible dans l'approximation qui est faite des périodes calculées. Par exemple, 20,60h est représenté par « environ 20 heures » et 24,20h par « exactement 1 jour ». C'est cette dernière formulation qui est la plus conforme au résultat attendu.

7.5 Bilan

Ce chapitre présente deux études expérimentales permettant de valider la méthode DPE en termes de pertinence et de performance. La première étudie la méthode selon différents critères : décroissance régulière du degré de périodicité avec le bruit dans les données, évaluation juste de la période, robustesse de l'analyse pour des niveaux de bruits équivalents et étiquetage correct des groupes hauts et bas.

A l'aide de plusieurs scénarios permettant de comparer plusieurs variantes des différentes étapes de DPE détaillées dans les sections 5.2 p. 98 et 5.3 p. 105, nous avons établi que la plus efficace est basée sur la méthode de regroupement γ_{es} et utilise une cardinalité crisp pour évaluer la taille des groupes identifiés, une moyenne et une déviation absolue moyenne pour calculer leur régularité et une moyenne pour l'agréger et déterminer la périodicité de la série.

La seconde étude expérimentale est dédiée à la comparaison des performances des différentes méthodes de calcul du score d'érosion détaillées au chapitre 6. Cette étude montre que la méthode incrémentale par niveaux est la plus efficace et permet de calculer le score d'érosion d'un jeu de données d'un million de points en 1,5 seconde.

La méthode DPE est donc pertinente et efficace. Différents axes d'amélioration ont été identifiés au cours du chapitre comme l'évaluation par questionnaire du rendu linguistique, l'utilisation de la médiane dans les calculs de régularité ou la correction d'un biais de la méthode γ_{es} par rapport aux valeurs faibles érodées de nombreuses fois.

Chapitre 8

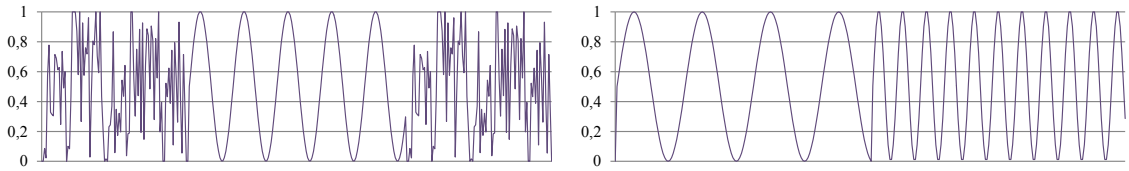
Contextualisation de la périodicité

> Notre histoire particulière dépend de nous encore, non le contexte dans lequel elle s'inscrit : ce qui ne signifie pas qu'elle nous échappe. Agir en tenant compte du contexte.
>
> —François Mauriac, *Le nouveau Bloc-Notes*

La méthode LDPE (*Local Detection of Periodic Events*) de détection locale des événements périodiques est une généralisation de la méthode DPE qui contextualise dans le temps la périodicité π, la période p et l'expression linguistique pour chaque partie du jeu de données qu'elle identifie comme localement périodique.

LDPE permet donc l'analyse de séries à périodicité locale comme les séries `(h)` et `(i)` de la figure 4.2 p. 71 rappelées sur la figure 8.1. Ces séries sont non stationnaires et leurs caractéristiques de périodicité évoluent dans le temps. Pour celle de gauche, la périodicité est faible aux extrémités et élevée en son centre tandis que celle de droite affiche une périodicité élevée dans l'ensemble mais avec deux périodes différentes.

Dans les deux cas, DPE renvoie un degré de périodicité faible et une période non significative car ces résultats sont calculés sur l'ensemble des groupes H et L identifiés lors de la première étape de la méthode détaillée dans la section 5.2 p. 98. Nous proposons donc avec LDPE d'identifier automatiquement les sous-ensembles de groupes ou *zone* de périodicité homogène puis de calculer leur période de la même manière qu'avec DPE. Comme illustré sur la figure 8.2, LDPE est une généralisation de DPE qui prend en entrée les groupes extraits de la première étape de DPE par la méthode de regroupement et renvoie un résultat équivalent, soit une période, une périodicité et une phrase, pour



Périodicité locale - 1 période Périodicité locale - 2 périodes

Figure 8.1 – Séries à périodicité locale

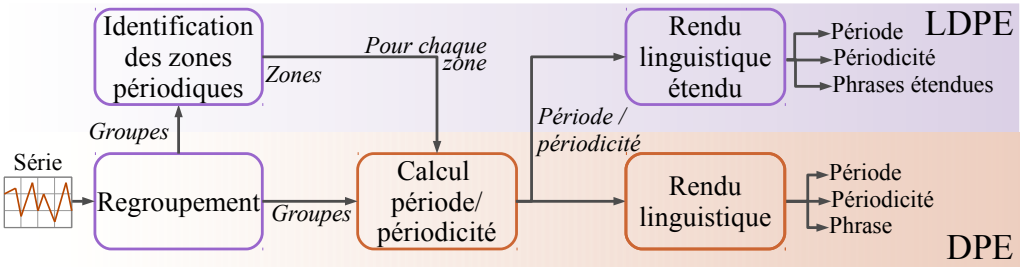


FIGURE 8.2 – Vue générale de la méthode LDPE englobant DPE

chaque zone périodique détectée. Un exemple de phrase générée est « Environ de Mars à Juin, les données sont périodiques de période exactement 2 semaines ». Dans le cas d'une série stationnaire pour laquelle la périodicité est constante, une seule zone est détectée et le résultat de LDPE est le même que celui de DPE.

La première phase de LDPE dédiée à l'identification des zones périodiques est réalisée en trois temps : d'abord, la *périodicité locale* de la série est calculée, comme détaillé dans la section 8.1, puis son *front de périodicité* présenté dans la section 8.2 p. 162 est déterminé, et les zones périodiques sont extraites selon l'approche décrite dans la section 8.3 p. 164.

La phase de calcul de la période, identique à celle de DPE, n'est pas représentée ici. En revanche, celle de génération linguistique, plus complexe du fait de la présence de l'information de contexte temporel, est détaillée dans la section 8.4 p. 167. Enfin, les nombreuses expériences réalisées sur des données artificielles et réelles pour valider notre approche sont présentées et discutées dans la section 8.5 p. 170.

Les travaux de ce chapitre ont fait l'objet de la publication (Moyse & Lesot, 2015).

8.1 Périodicité locale

La notion de périodicité locale définie dans LDPE correspond à la périodicité utilisée dans DPE calculée sur un sous-ensemble des groupes renvoyés par la méthode de regroupement. Dans les expériences de la section 8.5 p. 170 cette méthode est γ_{es}, mais LDPE ne fait aucune hypothèse sur la manière dont les groupes sont identifiés.

Dans cette section, nous donnons une définition précise de la périodicité locale ainsi qu'un test d'hypothèse permettant d'évaluer sa significativité.

8.1.1 Définition

Formellement, la méthode attend une liste ordonnée de g groupes $G = (G_j)_{j=1\ldots g}$ à partir de laquelle nous définissons π_j la *périodicité locale* du $j^{ème}$ groupe, calculée comme la périodicité du sous-ensemble de groupes contenant G_j évalué comme le plus pertinent. En posant $1 \leq j^- \leq j \leq j^+ \leq g$ et $j^+ - j^- > 1$, π_j est défini comme :

$$\pi_j = \pi\left(G, j^-, j^+\right) \tag{8.1}$$

avec $\pi(G, j^-, j^+)$ la périodicité calculée selon l'éq. (5.14) p. 108 en ne prenant en compte que les groupes dont les indices sont compris entre j^- et j^+ inclus. Avec cette notation, le résultat renvoyé par DPE s'écrit $\pi(G, 1, g)$. Lorsque le contexte est suffisamment clair $\pi(G, j^-, j^+)$ est noté $\pi(j^-, j^+)$.

La contrainte $j^+ - j^- > 1$ implique que deux groupes au minimum sont pris en compte pour le calcul de π_j et garantit qu'au moins un groupe H et un groupe L sont inclus.

La problématique du calcul de π_j est celle du choix des indices j^- et j^+ qui répond au compromis suivant : soit les indices sont proches afin de représenter localement la périodicité qui en ce cas cependant peut ne pas être significative car quelques groupes adjacents peuvent être à peu près de même taille et donc sembler périodiques par hasard, soit les indices sont distants auquel cas la périodicité calculée est significative mais la notion de localité est perdue.

L'utilisation d'un écart constant défini par l'utilisateur entre les indices j^- et j^+ n'est pas envisageable car il n'existe pas d'approche simple permettant de le définir tout en garantissant la significativité et la localité du résultat. De plus, la figure 8.3 p. 163 illustre l'inefficacité d'un écart constant vis-à-vis d'un écart adapté aux groupes considérés localement.

Ainsi, nous proposons pour LDPE une méthode permettant la résolution du compromis exposé ci-dessus, en évaluant la significativité de la périodicité calculée à l'aide d'un test statistique original et en faisant évoluer les bornes j^- et j^+ de manière progressive afin de maintenir une évaluation la plus locale possible. Nous introduisons le test statistique dans la sous-section suivante, le second aspect lié à l'évaluation des bornes j^-et j^+ étant détaillé dans la section 8.2.

8.1.2 Test de significativité de la périodicité locale

Nous supposons dans la suite du paragraphe que les groupes étudiés sont situés dans un *voisinage* de j, d'indices compris dans $[j^-, j^+]$. π désigne dans ce contexte la périodicité locale π_j. Afin de tester sa significativité, nous définissons l'hypothèse nulle H_0 comme celle correspondant au cas où la valeur prise par la variable aléatoire π sur le voisinage considéré a été obtenue par hasard. Nous associons un seuil α tel que H_0 est rejetée si la probabilité que π prenne cette valeur est inférieure à α, donc que la probabilité que π ait été obtenue par hasard soit trop faible pour qu'H_0 puisse être acceptée.

Avec τ le type des groupes dans $\{H, L\}$, notons g^τ le nombre de groupes de type τ dans le voisinage considéré, s_j^τ le nombre de points, ou la taille, du $j^{ème}$ groupe de type τ, $n^\tau = \sum_{j=j^-}^{j^+} s_j^\tau$ le nombre de points dans tous les groupes de type τ, $\mu^\tau = n^\tau / g^\tau$ leur taille moyenne et $d^\tau = \sum_{j=j^-}^{j^+} |s_j^\tau - \mu^\tau| / n^\tau$ leur déviation absolue moyenne (cf. éq. (5.13) p. 106).

Nous rappelons que $\pi = (\rho^H + \rho^L)/2$ et $\rho^\tau = 1 - \min(1, d^\tau/\mu^\tau)$, donc π dépend de μ^τ et d^τ (cf. éq. (5.14) p. 108) . Néanmoins, pour un voisinage donné, n^τ et g^τ sont fixés, donc $\mu^\tau = n^\tau / g^\tau$ l'est également. Ainsi, la seule composante variable dans l'expression de π est d^τ, à savoir la manière dont les points sont répartis dans les groupes. À titre

d'exemple, il existe pour le calcul de d^τ 2 380 façons de répartir 20 points dans 5 groupes sachant que chaque groupe doit contenir au moins un point et que d^τ n'est pas sensible à leur ordre. Parmi celles-ci, 820 ont une déviation d^τ égale à 2,40 soit une probabilité de 0,21 mais seulement 20 en ont une égale à 4,40, soit une probabilité de 0,01. Il est donc raisonnable de penser que le second cas n'est pas lié à une répartition aléatoire des points dans les groupes, qui correspond à une zone de bruit dans les données et donc à une valeur de déviation non significative, mais que le premier peut l'être.

Nous proposons par conséquent d'évaluer la probabilité que la périodicité calculée pour un voisinage donné soit non significative comme celle que la déviation des tailles groupes H et L soit non significative. En notant δ^H la valeur prise par la variable d^H et δ^L celle prise par d^L pour un voisinage donné donc pour n^H, n^L, g^H et g^L connus, nous rejetons H_0 avec un seuil de significativité α si :

$$\max\left(P\left(d^H = \delta^H\right), P\left(d^L = \delta^L\right)\right) < \alpha \tag{8.2}$$

Nous évaluons cette condition avec l'expression de $P(d = \delta)$ donnée dans le théorème 6.

Théorème 6. *Expression de* $P(d = \delta)$

Étant donné n points répartis dans g groupes, la probabilité que la déviation absolue moyenne de leur taille soit égale à δ est donnée par :

$$P\left(d = \delta\right) = \begin{cases} 0 & \text{si } n < g \text{ ou } g < 1 \text{ ou } g^2\delta \notin \mathbb{N} \\ \dfrac{\sum_{l \in \Lambda} \widetilde{N}\left(n, g, l, \delta\right)}{N\left(n, g, 1, +\infty\right)} & \text{sinon} \end{cases} \tag{8.3}$$

où N représente le nombre de répartitions possibles de n points dans g groupes et $\widetilde{N}$ le nombre de répartitions de n points dans g groupes dont la déviation absolue moyenne des tailles vaut δ et dont l contiennent un nombre de points inférieur ou égal à n/g.

Démonstration. La preuve est réalisée en deux temps. D'abord, une expression générale de d est établie, permettant son calcul par décomposition des groupes en deux, selon que leur taille est inférieure ou égale ou bien supérieure à la moyenne. La probabilité est ensuite calculée grâce à cette décomposition comme le rapport du nombre de combinaisons des n points dans les g groupes de déviation δ par le nombre total de combinaisons de n points dans g groupes. Le détail de la démonstration est donné dans l'annexe H p. 237. □

8.2 Fronts de périodicité

A l'aide du test d'hypothèse décrit ci-dessus, nous définissons les *fronts de périodicité* comme les séquences de périodicité locale π_j pour $j = 1...g$. Trois fronts de périodicité π^L, π^C et π^R sont définis en fonction de la stratégie utilisée pour calculer les bornes j^-et j^+pour chacune des valeurs de périodicité locale du front : π^L désigne un front de périodicité « vers la gauche », π^C « au centre » et π^R « vers la droite ».

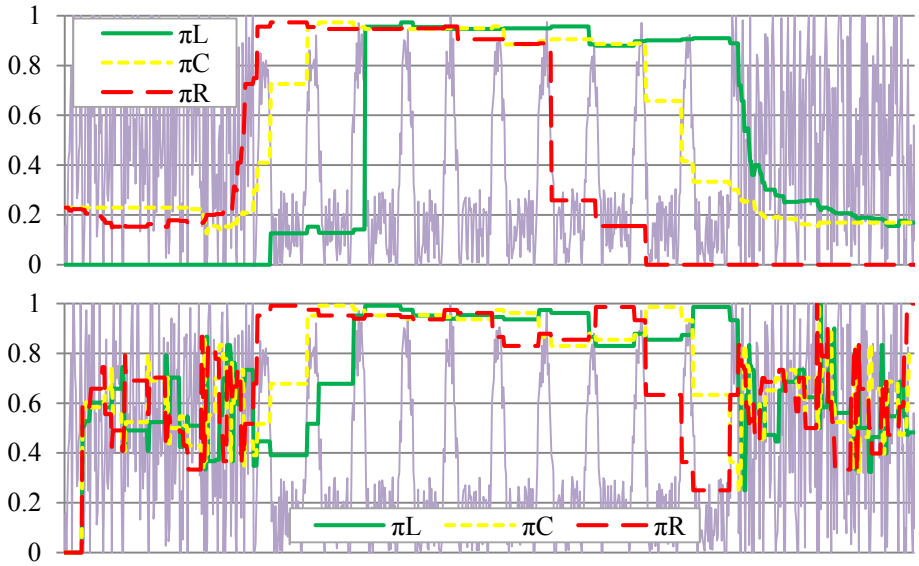


FIGURE 8.3 – Exemples de fronts de périodicité vers la gauche π^L, au centre π^C et vers la droite π^R. L'écart utilisé pour leur calcul entre les bornes j^- et j^+ est constant sur le graphique du bas, tandis qu'il est adaptatif et basé sur le test d'hypothèse présenté dans la section 8.1.2 p. 161 sur celui du haut.

Plus précisément, π_j^D la $j^{ème}$ valeur de périodicité locale d'un front de périodicité avec $D = \{L, C, R\}$ est définie par :

$$\pi_j^L = \arg\min_{k>0} \{\pi\,(j-k, j) \text{ tel que } H_0 \text{ est rejetée}\} \tag{8.4}$$

$$\pi_j^C = \arg\min_{k>0} \{\pi\,(j-\lfloor k/2\rfloor, j+\lceil k/2\rceil) \text{ tel que } H_0 \text{ est rejetée}\} \tag{8.5}$$

$$\pi_j^R = \arg\min_{k>0} \{\pi\,(j, j+k) \text{ tel que } H_0 \text{ est rejetée}\} \tag{8.6}$$

Ainsi, le calcul des fronts de périodicité débute avec $k = 1$, évalue π_j à gauche, autour, ou à droite du groupe j et retient cette valeur si elle permet de rejeter H_0, sinon incrémente k et recommence. Cette méthode s'adapte aux données considérées car le nombre de groupes pris en compte pour le calcul de la périodicité locale est variable en fonction du groupe considéré. De plus, sa complexité est faible car les valeurs successives des fronts peuvent être calculées de manière incrémentale et que les calculs de probabilités associés au test statistique sont souvent identiques et peuvent être accélérés avec un simple cache.

Nous avons représenté sur le graphique du haut de la figure 8.3 les fronts de périodicité déterminés de la sorte avec un seuil de significativité $\alpha = 1\%$. Les trois ont des valeurs élevées dans la partie centrale où les données sont le plus périodiques et des valeurs basses sur les bords du jeu de données. Le front gauche est bas au début mais élevé jusqu'à la fin de la zone périodique, le front droit l'est dès le début mais redevient bas avant sa fin, et le front central est élevé sur l'ensemble de la zone périodique à l'exception de ses bords : une baisse importante du front gauche indique une fin de zone périodique, une hausse du front droit son début et le front central représente un compromis entre les deux autres.

Afin d'illustrer l'intérêt de cette méthode, nous l'avons comparée avec une approche par écart constant entre j^- et j^+ où chaque valeur du front de périodicité est calculée pour une valeur de k donnée sans la faire évoluer jusqu'à vérification d'une condition. Le résultat de cette approche est illustré sur le graphique du haut de la figure 8.3 avec $k = 4$. En ce cas, la variabilité des valeurs des fronts est plus élevée, même dans la zone périodique au centre. De plus, leurs valeurs sur les zones bruitées semblent aléatoires, interdisant toute analyse ultérieure.

La supériorité de la première méthode sur la seconde est liée au fait que les groupes renvoyés dans les zones de bruit sont petits et nombreux. En effet, comme les valeurs du jeu de données y varient beaucoup les méthodes de regroupement y associent plusieurs petits groupes haut et bas. Or, comme montré par le théorème 6 p. 162, la probabilité que la déviation soit égale à la valeur calculée pour les groupes autour de j est d'autant plus élevée que le nombre de points est faible et que le nombre de groupes est élevé, ce qui est le cas pour les groupes petits et nombreux. Ainsi, la méthode par test statistique va augmenter automatiquement le nombre de groupes pris en compte dans ce cas jusqu'à obtenir une probabilité plus faible que le seuil fixé. Comme de nombreux groupes sont intégrés dans la calcul de la périodicité, la variabilité de leurs tailles est détectée et la valeur de la périodicité locale est faible.

8.3 Zones périodiques

Une *zone périodique* est un ensemble ordonné de groupes contigus identifiés comme périodiques. De la même manière que les points successifs de même type H ou L sont regroupés dans la première étape de DPE pour créer les groupes H et L (cf. section 5.2 p. 98), les groupes à leur tour sont étiquetés P ou N et regroupés en zones périodique ou non périodique respectivement.

Les fronts de périodicité définis dans la section précédente sont adéquats pour étiqueter les groupes car les caractéristiques de chacun d'entre eux permettent de détecter le début des zones périodiques, avec π^R, leur centre avec π^C et leur fin avec π^L. Nous détaillons dans la section 8.3.1 différentes méthodes d'agrégation de la valeur de ces fronts afin d'étiqueter les groupes puis dans la section 8.3.2 p. 166 une méthode de définition des zones périodiques sur cette base.

8.3.1 Étiquetage des groupes

Nous proposons de réaliser l'étiquetage des groupes périodiques par une simple technique de seuillage permettant de séparer les groupes de périodicité élevée étiquetés P des autres étiquetés N.

Cette approche permet d'identifier simplement les changements de valeurs significatifs dans les fronts de périodicité. L'étude de ces changements est associée aux domaines du *change point detection* présenté par Basseville & Nikiforov (1993) et du *concept drift*, déjà mentionné dans la section 6.3.1 p. 128. Ces deux familles d'approches nécessitent toutefois

la définition de modèles a priori pour les données, différentes en ce sens de l'approche retenue pour notre méthode qui n'en présuppose aucun.

D'autre part, l'utilisation du regroupement par score d'érosion afin de déterminer les valeurs élevées des fronts de périodicité n'est pas appropriée ici car il lisse les parties bruitées du signal, ce qui n'est pas souhaité pour les fronts de périodicité puisqu'ils le sont déjà par l'utilisation du test statistique. De plus, le score d'érosion est adapté à l'analyse de la périodicité du signal, qui n'est pas la propriété étudiée des fronts de périodicité.

Nous définissons la fonction d'étiquetage $m : G \to \{P, N\}$ qui assigne une étiquette P à un groupe j s'il appartient à une zone périodique et N sinon, le front de périodicité π^M, max de l'ensemble des fronts, défini par $\pi_j^M = \max(\pi_j^L, \pi_j^C, \pi_j^R)$, et les seuils moyens non pondéré $\overline{\pi}_j^d$ et pondéré $\hat{\pi}_j^d$ des fronts de périodicité de type $d \in \{L, C, R, M\}$ respectivement définis par :

$$\overline{\pi}^d = \max\left(\frac{1}{g}\sum_{j=1}^{g}\pi_j^d, \pi_{min}\right) \quad \text{et} \quad \hat{\pi}^d = \max\left(\frac{1}{n}\sum_{j=1}^{g} s_j \times \pi_j^d, \pi_{min}\right) \tag{8.7}$$

où s_j est la taille du $j^{ème}$ groupe et π_{min} la valeur minimale acceptable pour les seuils. La version pondérée de ces derniers permet de donner plus d'importance à la périodicité locale des grands groupes par rapport à celle des petits. Sur le graphique du haut de la figure 8.3 p. 163 par exemple, de nombreux groupes de périodicité faible sont identifiés sur les bords de la série et un grand groupe de périodicité élevée est détecté en son milieu. Sans pondération, le seuil moyen de périodicité est faible car les petits groupes apériodiques sont supérieurs en nombre. En pondérant par la taille, le seuil moyen est rehaussé par la périodicité importante du grand groupe central.

Les trois fonctions d'étiquetage m_1, m_2 et m_3 définies ci-dessous utilisent le seuil non pondéré, les versions utilisant le seuil pondéré étant notées m_{1w}, m_{2w} et m_{3w}.

La fonction m_1 considère qu'un groupe appartient à une zone périodique si le max des trois fronts de périodicité π^L, π^C et π^R du groupe j est supérieur ou égal à celui de leur moyenne, i.e. :

$$m_1(j) = \begin{cases} P & \text{si } \pi_j^M \geq \overline{\pi}^M \\ N & \text{sinon} \end{cases} \tag{8.8}$$

La méthode m_2 utilise les particularités des fronts de périodicité, à savoir que π^L indique les fins de zone périodique et π^R leur début. Ainsi, si l'un ou l'autre est supérieur à son seuil moyen et qu'en plus π^C l'est également, indiquant que les groupes sont périodiques autour de j, alors le groupe peut-être considéré comme appartenant à une zone périodique. Nous définissons donc m_2 par :

$$m_2(j) = \begin{cases} P & \text{si } \pi_j^C \geq \overline{\pi}^C \wedge \left(\pi_j^L \geq \overline{\pi}^L \vee \pi_j^R \geq \overline{\pi}^R\right) \\ N & \text{sinon} \end{cases} \tag{8.9}$$

La méthode m_3 considère qu'un groupe appartient à une zone périodique dès qu'un

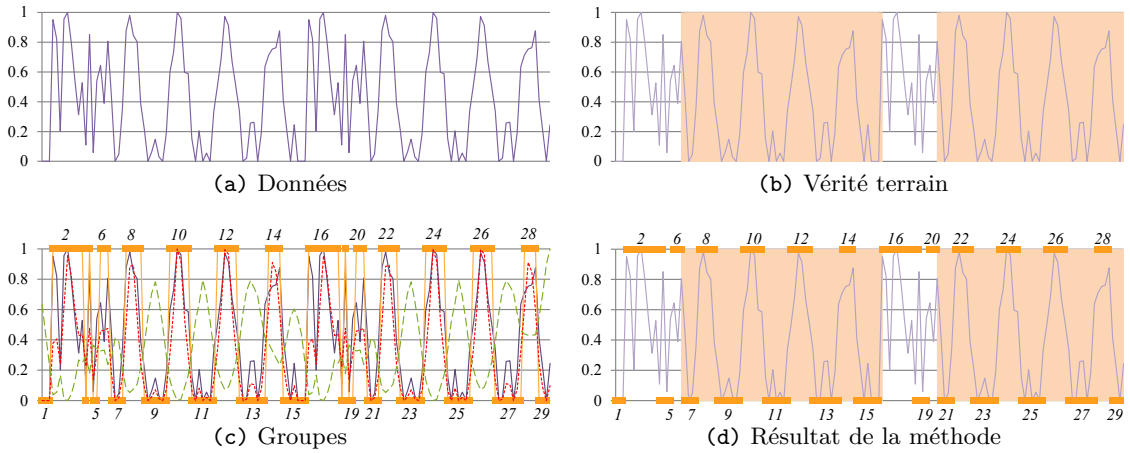


FIGURE 8.4 – Analyse des zones périodiques (fond coloré) d'une série de données illustrée sur le graphique (a). Le graphique (b) illustre la vérité terrain, le graphique (c) les groupes identifiés par la méthode de regroupement γ_{es} et le graphique (d) les zones périodiques détectées par LDPE

de ses fronts de périodicité est supérieur à son seuil moyen. La méthode est liée à m_1 qui considère le max des fronts de périodicité mais est plus optimiste que cette dernière car il suffit qu'un seul front soit supérieur à son seuil moyen pour que le groupe soit considéré comme périodique. m_3 est définie comme :

$$m_3(j) = \begin{cases} P & \text{si } (\pi_j^L \geq \overline{\pi}^L) \vee (\pi_j^C \geq \overline{\pi}^C) \vee (\pi_j^R \geq \overline{\pi}^R) \\ N & \text{sinon} \end{cases} \tag{8.10}$$

8.3.2 Définition des zones périodiques

A l'aide des étiquettes P ou N attribuées à chacun des groupes, les zones périodiques $Z = (Z_k)_{k=1\ldots z}$ sont constituées comme les ensembles ordonnés de groupes successifs étiquetés P.

Afin de permettre que deux zones très proches soient considérées comme une seule, celles séparées par moins de *minSep* groupes sont fusionnées. De même, pour ne pas renvoyer de zones trop petites, celles contenant moins de *minSize* groupes sont éliminées. Ce filtrage est réalisé à l'aide des opérateurs de fermeture et d'ouverture issus de la morphologie mathématique (Serra, 1983) : le premier fusionne les zones proches avec une dilatation puis une érosion de taille *minSep* et le second élimine les petites zones avec une érosion puis une dilatation de taille *minSize*. Les méthodes utilisant ce *filtrage post traitement* sont préfixées par un *f*, par exemple fm_1 ou fm_{3w}.

Les zones périodiques sont représentées par les indices de début et de fin du premier et du dernier groupe qu'elles contiennent et enrichies de leur période et de leur périodicité calculées avec l'approche DPE classique décrite dans la section 5.3 p. 105. Pour la figure 8.4

par exemple, deux zones sont déterminées :

$$Z_1 = ([23, 73], 0.83, 11.90)$$
$$Z_2 = ([93, 140], 0.78, 11.30)$$

signifiant que la première zone périodique Z_1 s'étend du point 23 au début du groupe 7 jusqu'au point 73 à la fin du groupe 15, que son degré de périodicité est $\pi = 0,83$ et que sa période candidate est $p_c = 11,90$. La deuxième zone périodique Z_2 s'étend du point 93 au début du groupe 21 jusqu'au point 140 à la fin du groupe 29, son degré de périodicité est $\pi = 0,78$ et sa période candidate est $p_c = 11,30$. La méthode LDPE renvoie bien en ce cas des résultats conformes à ceux attendus par l'utilisateur après inspection visuelle.

8.4 Rendu linguistique

Le rendu linguistique de LDPE permet de renvoyer une ou plusieurs phrases décrivant les zones périodiques identifiées lors des étapes détaillées dans les sections précédentes. En plus de permettre l'expression de la période, décrite pour DPE dans la section 5.4 p. 109, il intègre le contexte temporel, i.e. la localisation des zones périodiques, l'évaluation linguistique du degré de périodicité, et génère autant de phrases que de zones identifiées, contre une seule pour DPE.

Les sous-sections suivantes décrivent le protoforme utilisé dans LDPE ainsi que nos propositions concernant son rendu linguistique.

8.4.1 Protoforme utilisé

Nous proposons d'exprimer chaque zone périodique par le protoforme suivant :

$$\underbrace{\textit{Prec}_1\,\textit{CtxtTemp}}_{\textit{Contexte temporel}}\ \text{la série est}\ \underbrace{\textit{Pdté}\ (\pi)}_{\textit{Périodicité}}\Big[\underbrace{\text{de période}\ \textit{Prec}_2\,p\,\textit{unités}}_{\textit{Période}}\Big] \tag{8.11}$$

où $\textit{Prec}_1$ et $\textit{Prec}_2$ sont les adverbes de précision utilisés dans DPE et décrits par la variable linguistique illustrée sur la figure 5.9 p. 111, *CtxtTemp* est l'expression linguistique du contexte temporel comme « le premier trimestre » ou « durant l'été », *Pdté* est une expression linguistique de la périodicité, comme « très périodique », ou simplement « périodique » et π est la valeur de périodicité, p et *unités* une expression appropriée de la période p_c et de son unité respectivement. La partie *Période* entre crochets est facultative et intégrée dans la phrase résultat seulement si la périodicité π est suffisamment élevée.

Les phrases générées selon ce protoforme sont par exemple :

— « Les deux premiers mois, la série est très périodique (0,89) de période environ 1 semaine »

— « Durant le premier trimestre, la série est périodique (0,78) de période exactement 1 mois »

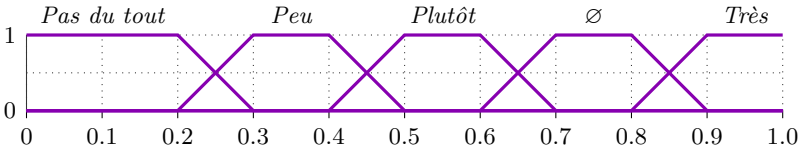


FIGURE 8.5 – Variable linguistique pour le degré de périodicité π

— « De septembre à novembre, la série est peu périodique »

Les deux exemples suivants caractérisent les deux zones identifiées à partir des données illustrées sur la figure 8.4 p. 166 dont la fréquence d'échantillonnage est le mois :

— « Environ le second quart de la série est périodique (0,83), de période environ 12 mois »

— « Environ le dernier quart de la série est périodique (0,78), de période environ 11 mois »

Nous décrivons dans les sous-sections suivantes les approches que nous avons développées pour rendre la périodicité et le contexte temporel de la zone périodique.

8.4.2 Rendu du degré de périodicité

Comme décrit dans la section 5.4 p. 109, la méthode DPE rend linguistiquement la période identifiée dans la série ainsi que la qualité de son approximation à l'unité la plus proche, mais pas son degré de périodicité π, simplement indiqué par sa valeur entre parenthèses à la fin de la phrase.

Afin de l'exprimer linguistiquement dans LDPE, nous proposons d'associer π à la variable linguistique illustrée sur la figure 8.5, utilisée de la même manière que la variable *Précision* détaillée dans la section 5.4.4 p. 110, i.e. la modalité retenue est celle dont la fonction d'appartenance est maximisée par π.

La modalité $\varnothing$ représente le quantificateur standard, i.e. la qualification d'une zone simplement « périodique », par opposition à d'autres zones « très périodiques » ou « plutôt périodiques » par exemple.

Les modalités « peu » et « pas du tout » ne sont pertinentes que dans le cas où la représentation linguistique souhaitée est exhaustive, i.e. incluant aussi les zones peu périodiques, pour lesquelles la partie *Période* entre crochets dans l'éq. (8.11) n'est pas générée.

Une fois la modalité sélectionnée, la partie *Périodicité* est instanciée en « *modalité* périodique (π) » comme par exemple « Plutôt périodique (0,51) ».

8.4.3 Rendu du contexte temporel

Le rendu du contexte temporel, spécifique à LDPE, permet de représenter linguistiquement la localisation dans le temps d'une zone périodique donnée. C'est sur la base de l'intervalle de la zone indiquant son début et sa fin en termes de points dans le jeu de données que le rendu linguistique est réalisé, de manière absolue ou relative, comme indiqué dans les deux paragraphes suivants.

Rendu absolu Le rendu absolu de la zone périodique fait référence aux unités du jeu de données et à un référentiel d'intervalles linguistiques fournis par l'utilisateur, afin de générer de phrases comme « les deux premiers trimestres » ou « la fin de l'année ».

Comme détaillé ci-dessous, les intervalles linguistiques fournis par l'utilisateur peuvent être ponctuels comme *Les vacances de printemps* ou *La semaine de Roland-Garros* ou bien rassemblés dans une liste, comme *Trimestres* ou *Mois*. Dans tous les cas, chaque intervalle est constitué d'une valeur de début et d'une valeur de fin dans l'unité de jeu de données, comme $[1, 90]$ pour l'intervalle *1er trimestre* lorsque le jeu de données est exprimé en jours.

Intervalles ponctuels Un intervalle ponctuel est un intervalle simple différents des listes d'intervalles décrites dans le paragraphe suivant. Sur le calendrier français 2016 par exemple, l'intervalle *les vacances de printemps* est représenté par $[99, 114]$ pour un jeu de données où l'unité est le jour, correspondant à l'intervalle du 9 avril ($99^{ème}$ jour de l'année) au 24 avril ($114^{ème}$ jour de l'année).

Listes d'intervalles D'autres intervalles sont définis sous formes de listes ordonnées, comme les trimestres, les mois ou les jours, contenant les intervalles ponctuels qui les constituent. Par exemple, la liste d'intervalles *Trimestres* est définie par :

$$([1, 90], [91, 181], [182, 273], [274, 365]) \tag{8.12}$$

où chaque intervalle correspond à un trimestre de l'année, par exemple le 2^{nd} intervalle fait référence au 2^{nd} trimestre qui s'étend du $91^{ème}$ au $181^{ème}$ jour de l'année. Le rendu linguistique du $i^{ème}$ intervalle est par exemple *Trimestre i* ou $i^{ème}$ *trimestre* ou encore *Dernier trimestre* pour le dernier intervalle.

Les listes d'intervalles peuvent aussi être organisées de manière hiérarchique, comme dans les travaux de Castillo-Ortega et al. (2011a) présentés dans la section 1.3.2 p. 17.

Lien entre zones périodiques et intervalles linguistiques Chaque zone périodique z est comparée aux intervalles linguistiques fournis par l'utilisateur. Les intervalles ponctuels sont analysés en premier lieu, car définis spécifiquement par l'utilisateur et habituellement moins standards que les listes d'intervalles.

La comparaison avec les intervalles ponctuels est basée sur le calcul de la distance Moore (Moore, 1963) entre intervalles : la distance d entre une zone dont l'intervalle est $z = [z^-, z^+]$ et un intervalle ponctuel $A = [a^-, a^+]$ est

$$d(z, A) = \max\left(\left|z^- - a^-\right|, \left|z^+ - a^+\right|\right) / \left(z^+ - z^-\right) \tag{8.13}$$

c'est-à-dire le plus grand écart entre les bornes inférieures et les bornes supérieures rapporté à la taille de z. Si cette distance est plus petite que le seuil ϵ de sélection d'une période approchée utilisé dans la section 5.4.3 p. 110, l'intervalle ponctuel est ajouté à une liste d'intervalles candidats.

La comparaison est ensuite poursuivie avec les listes d'intervalles selon une approche plus complexe liée au fait que z peut couvrir plusieurs intervalles consécutifs de la liste. Avec la liste des trimestres définie par l'éq. (8.12) et $z = [3, 184]$ par exemple, l'intervalle linguistique correspondant est *Les deux premiers trimestres*, soit l'union des deux premiers intervalles de la liste, qui est ajouté à la liste des intervalles candidats.

Afin de gérer ces cas de recouvrement sans pour autant avoir à les spécifier tous, nous proposons d'associer z à la représentation linguistique construite à partir du premier intervalle dont la distance relative de la borne inférieure à z^- est inférieure à ϵ jusqu'au dernier intervalle dont la distance relative de la borne supérieure à z^+ est inférieure à ϵ.

Par exemple, avec la zone $z = [3, 184]$, $\epsilon = 5\%$ et les intervalles de la liste *Trimestres*, l'intervalle dont la borne inférieure est à une distance relative inférieure à ϵ de z^- est $[1, 90]$ car $|3 - 1|/(90 - 1) = 0,022 < 5\%$ et celui dont la borne supérieure est à une distance relative inférieures à ϵ est $[182, 273]$ car $|184 - 182|/(273 - 182) = 0,022 < 5\%$.

Rendu linguistique des intervalles candidats Les intervalles candidats identifiés à l'étape précédente peuvent ensuite être convertis en phrases. Concernant les intervalles ponctuels, leur nom est utilisé directement, comme par exemple « Durant les vacances de Pâques, la série est... ».

Pour les listes d'intervalles si le premier intervalle de la liste est contenu dans le résultat, une phrase du type « Durant les n premiers trimestres, la série est... », si le dernier élément est contenu dans le résultat, une phrase du type « Durant les n derniers trimestres, la série est... », sinon la phrase est « Durant les trimestres a et b, la série est... ».

Rendu relatif Le rendu linguistique relatif est indépendant de l'unité utilisée dans le jeu de données initial et se base sur des fractions de ce dernier, permettant la génération de phrases comme « *Les deux premiers tiers* » ou « *La deuxième moitié* ». Le rendu de ces dernières est similaire au rendu absolu, basé dans ce cas sur des listes pré-existantes comme *Moitiés* = ([0, 0.5], [0.5 1]), ou *Tiers* = ([0, 0.33], [0.33, 0.66], [0.67, 1]) .

8.5 Expériences

Comme pour les expériences de la méthode DPE présentées au chapitre 7, celles visant à valider LDPE sont réalisées sur des données réelles et artificielles, ces dernières étant crées à l'aide du générateur détaillé dans la section 7.1 p. 134.

Les critères de qualité retenus sont introduits dans la section 8.5.1, puis le protocole expérimental utilisé dans la section 8.5.2, suivi des résultats obtenus et de leur discussion dans la section 8.5.3 p. 174, et enfin des conclusions sur données réelles dans la section 8.5.4 p. 177.

8.5.1 Critères de qualité

De la même manière que pour les tests sur DPE, l'utilisation de données artificielles permet la définition de critères de qualité simplement comparables avec les paramètres utilisés pour la génération des données.

Ici, les critères de qualité retenus sont le nombre de zones périodiques détectées ainsi que leur bonne localisation dans le temps.

Pour tester le premier critère, l'erreur zE sur le nombre de zones découvertes z est calculée par comparaison avec le nombre de zones z^T (T pour *truth)* utilisées pour générer le jeu de données, d'où :

$$zE = \begin{cases} \dfrac{\left|z - z^T\right|}{z} & \text{si } z > 0 \\ \mathbb{1}\left(z \neq z^T\right) & \text{sinon} \end{cases} \tag{8.14}$$

zE est donc une valeur positive à minimiser.

Pour le second critère, une comparaison groupe à groupe est réalisée entre les zones périodiques retournées par la méthode et celles définies à la génération afin de calculer le taux de bonne classification des groupes dans les zones périodiques de la même manière que le taux de bonne classification des points en groupes de valeurs hautes est utilisé pour tester la méthode DPE (cf. éq. (7.9) p. 140). En utilisant la fonction d'étiquetage des groupes m définie dans la section 8.3.1 p. 164 et m^T celle donnant les étiquettes générées, nous définissons pC le taux de bonne classification des groupes :

$$pC = \frac{1}{g}\sum_{j=1}^{g} \mathbb{1}\left(m(j) = m^T(j)\right) \tag{8.15}$$

$pC \in [0, 1]$ doit être maximisé.

Pour les mêmes raisons que pour les expériences sur DPE (cf. section 7.2 p. 139), l'évaluation du rendu linguistique n'est pas réalisée ici.

8.5.2 Protocole

Douze variantes de LDPE sont comparées, construites selon la méthode utilisée, m_1, m_2 ou m_3 définies dans les éq. (8.8) à (8.10) p. 166, l'utilisation ou non de seuils moyens pondérés ou non (cf. éq. (8.7) p. 165) et le filtrage post traitement ou non des zones périodiques décrit dans la section 8.3.2 p. 166. Ces variantes ainsi que leurs désignations sont résumées dans le tableau 8.1.

Chaque variante est testée avec 256 combinaisons de paramètres, chacun des quatre paramètres étant testé avec quatre valeurs :

— le niveau de signification $\alpha = \{1\%, 5\%, 10\%, 15\%\}$ utilisé pour le test statistique défini dans la section 8.1.2 p. 161,

— la valeur de périodicité minimale $\pi_{min} = \{0.2, 0.4, 0.6, 0.8\}$ utilisée pour la détermination des seuils moyens (cf. éq. (8.7) p. 165),

Tableau 8.1 – Désignation des douze méthodes comparées

Seuils moyens (cf. éq. (8.7) p. 165)		✓		✓
Filtrage post traitement (cf. section 8.3.2 p. 166)			✓	✓
Méthode 1	m_1	m_{1w}	fm_1	fm_{1w}
Méthode 2	m_2	m_{2w}	fm_2	fm_{2w}
Méthode 3	m_3	m_{3w}	fm_3	fm_{3w}

— l'écart minimal entre deux zones périodique successives *minSep*= $\{2, 4, 6, 8\}$ et la taille minimale d'une zone *minSize*= $\{2, 4, 6, 8\}$ utilisées lors de l'étape de filtrage post traitement détaillée dans la section 8.3.2 p. 166.

Enfin, chaque variante avec chaque combinaison de paramètres est calculée pour des jeux de données créés à l'aide du générateur décrit dans la section 7.1 p. 134 selon six scénarios déclinés selon cinq configurations, chacune répétée 20 fois.

Ainsi, le nombre de jeux de données généré est égal à 256 paramètres fois 6 scénarios fois 5 configurations fois 20 répétitions soit 153 600.

zE et pC sont évalués pour chaque variante, chaque jeu de paramètre et chaque jeu de donnée et leur moyenne et leur écart-type sont calculés à plusieurs niveaux : d'abord par configuration sur les 20 répétitions, puis par scénario sur les cinq configurations, puis par jeu de paramètre sur les six scénarios, puis par variante sur les 256 paramètres afin d'obtenir un classement des meilleures méthodes et une vue de l'influence des paramètres.

Scénarios Les scénarios, numérotés S1 à S6, représentent six cas de périodicité locale, i.e. de séries de données composées de zones périodiques et apériodiques, illustrées sur les figures 8.6 et 8.7 p. 173. Chaque scénario se décline en cinq configurations qui déterminent la taille de chacune de ces zones.

Le jeu de données issu de S1 (cf. figure 8.6) est composé de trois zones, apériodiques pour la première et la dernière et périodique pour la seconde, désignées par le code *npn*. Ce scénario constitue un test standard pour la méthode LDPE et permet de vérifier sa capacité à identifier une zone périodique au milieu de données bruitées. Les configurations associées à ce scénario entraînent la génération de jeux de données dont la zone centrale périodique est de plus en plus large au détriment des zones extérieures qui s'amenuisent. Dans la première configuration, les trois zones occupent respectivement 40%, 20% et 40% du jeu de données. Ces ratios évoluent linéairement vers ceux de la cinquième et dernière configuration où leurs valeurs sont respectivement 10%, 80% et 10%.

Les cinq autres scénarios illustrés sur la figure 8.7 sont définis selon le même principe de juxtaposition de zones périodiques et non périodiques dont la taille évolue. S2 permet de tester un cas plus complexe que S1 avec une zone périodique supplémentaire ajoutée à la $3^{ème}$ non périodique de S1. S2 est donc noté *npnp*. S3, noté *pnpnp*, ajoute une zone périodique au début du jeu de données. La vocation de S4 est le test de la méthode sur

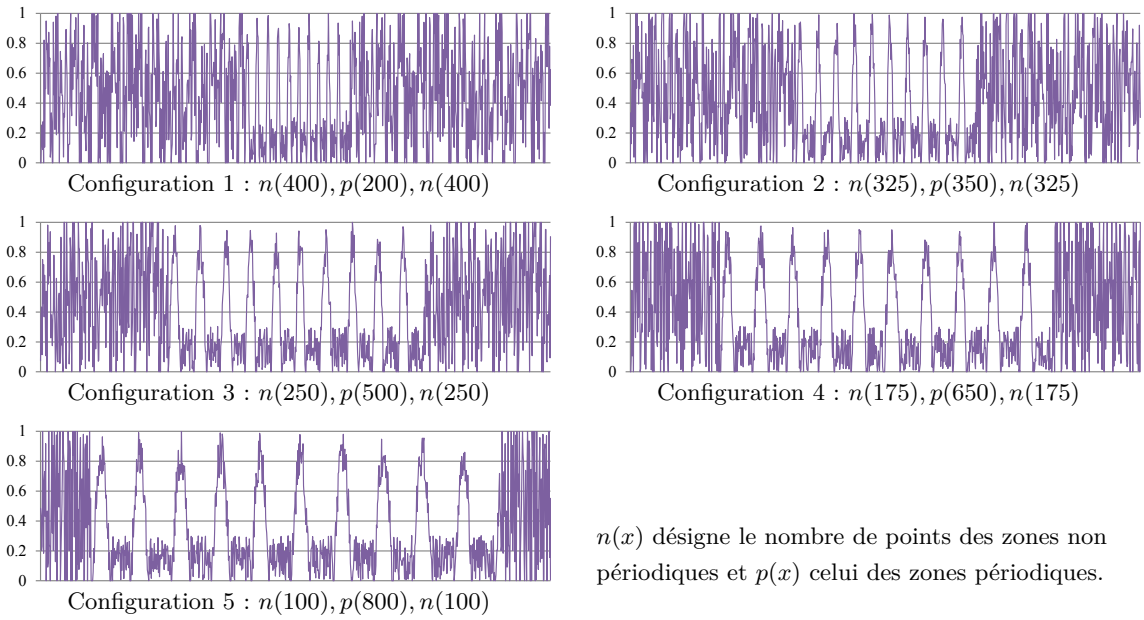


FIGURE 8.6 – Exemples de jeux de données générés par les cinq configurations de S1

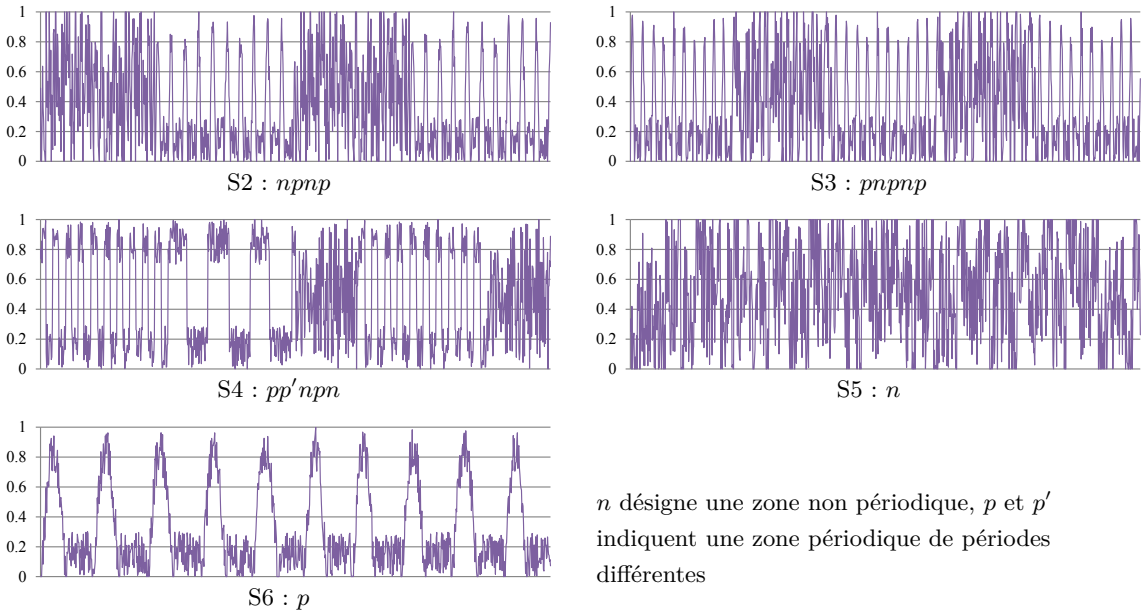


Figure 8.7 – Exemples de jeux de données générés pour les scénarios 2 à 6

des zones successives de périodes différentes, notées $pp'npn$. Enfin, S5 et S6 sont des cas triviaux destinés à vérifier que la méthode LDPE renvoie le même résultat que la méthode DPE dans le cas de séries complètement apériodiques ou complètement périodiques respectivement.

Les autres paramètres du générateur de données sont les suivants : pour toutes les zones, les groupes haut et bas ont la même taille et 10 groupes hauts et 10 groupes bas sont générés. Pour les zones périodiques, la forme Vague est utilisée à l'exception de S4 qui se base sur la forme Rectangle, le bruit sur la taille des groupes est $\nu_s^H = \nu_s^L = 0,2$ et celui en valeur est $\nu_v = 0,3$. Pour les zones non périodiques, la forme Sinus est utilisée

à l'exception de S4 qui se base sur la forme Rectangle, le bruit sur la taille des groupes est $\nu_s^H = \nu_s^L = 1$ et celui en valeur est $\nu_v = 1$.

8.5.3 Résultats et discussion

Du fait de leur nombre important, l'ensemble des résultats obtenus ne sont pas détaillés ici, mais rassemblés dans les tableaux de l'annexe I p. 243. Ils montrent que la méthode LDPE fonctionne bien pour l'identification des zones périodiques sur les différents scénarios. La meilleure méthode au regard des critères de qualité mesurés par zE et pC est fm_{2w}, basée sur le filtrage post traitement et les seuils moyens : son taux d'erreur $zE = 21\%$ dans le décompte des zones périodiques et son taux de bonne classification des groupes est $pC = 91\%$ en moyenne.

Les paragraphes suivants discutent plus en détail de l'influence des différents paramètres et des variantes de LDPE.

Influence des paramètres

Nous étudions dans ce paragraphe l'influence des paramètres α, π_{min}, *minSep* et *minSize* sur chacune des douze variantes testées. Cette influence est calculée comme le nombre de fois où une valeur de paramètre donnée est utilisée dans l'un des 30 meilleurs résultats obtenus. Ce nombre est pondéré par la position du résultat, i.e. la $30^{ème}$ position rapporte 1 point, la $29^{ème}$ 2 et ainsi de suite jusqu'à la $1^{ère}$ qui en rapporte 30. Ces scores sont disponibles dans les tableaux de l'annexe I p. 243 et discutés ci-dessous.

Résultats sur l'identification des zones zE Le paramètre le plus important dans l'identification des zones est π_{min} qui, lorsqu'il est égal à 0,8, la plus grande valeurs testée, renvoie entre 70% et 90% des 30 meilleurs résultats observés pour l'ensemble des méthodes à l'exception de fm_{2w} pour laquelle 60% des 30 meilleurs résultats sont obtenus avec $\pi_{min} = 0.6$.

Ce résultat est important car il implique que la valeur moyenne des fronts de périodicité (cf. éq. (8.7) p. 165) est un seuil trop faible puisque pour l'identification correcte des groupes appartenant à des zones périodiques puisque les résultats sont meilleurs à mesure que ce seuil est rehaussé par des valeurs de π_{min} plus grandes. L'usage de fractiles proches de 1, comme le troisième quartile par exemple, pourrait être envisagé afin de définir des seuils plus robustes aux valeurs faibles des fronts de périodicité.

D'autre part, l'influence du niveau de signification α *dans le cas des méthodes non filtrées* apparaît aussi clairement, puisque 80% des meilleurs résultats pour l'ensemble de ces méthodes sont obtenus avec $\alpha = 1\%$ ou 5%, i.e. les deux plus petites valeurs testées. Ce résultat indique l'efficacité du test d'hypothèse décrit dans la section 8.1.2 p. 161. En effet, plus α est petit et plus la fenêtre prise en compte pour le calcul de la périodicité locale est large, permettant un lissage des fronts de périodicité et compensant ainsi l'absence de filtrage ultérieur.

Cette analyse est corroborée par l'absence d'influence claire du paramètre α pour les *méthodes avec filtrage post traitement* (préfixées par f). En ce cas, les meilleurs résultats sont obtenus indépendamment des valeurs de *minSep* et *minSize*, indiquant une bonne robustesse de la méthode aux paramètres de filtrage utilisés.

Résultats sur la classification des points pC π_{min} est ici aussi le paramètre le plus influent pour la classification des points, avec un effet similaire à celui qu'il a pour l'identification des zones, excepté pour les méthodes m_2 et fm_2.

Au contraire, l'influence de α est nettement moins importante pour ce second critère. Pour expliquer ce résultat, nous rappelons les caractère *pessimiste* de zE et *optimiste* de pC : en effet, la mauvaise classification d'un groupe pour zE entraîne une augmentation de $1/z^T$ de l'erreur, tandis qu'elle entraîne une diminution de $1/g$ pour pC. Étant donné que $z^T = 1...3$ et $g \approx 100$ avec les paramètres du scénario S4 décrits dans la section 8.5.2 p. 172, l'impact est donc de 33% au minimum pour zE et 1% pour pC. Ainsi, les score zE sont moins bon que pC.

Ainsi, le pouvoir de filtrage des faux positifs de α est moins sensible pour pC car les quelques erreurs de classements évitées avec une faible valeur du paramètre sont peu sensibles sur pC. Il semblerait même que l'effet de α soit légèrement antagoniste à celui observé pour zE car les meilleurs scores sont obtenus pour les valeurs testées les plus élevées ($\alpha = 5\%$ ou 10% pour les méthodes non filtrées et $\alpha = 5\%$ à 15% pour les méthodes filtrées) tandis que les meilleurs scores de zE surviennent pour ses valeurs les plus faibles. Cet effet est dû à l'élargissement de la fenêtre auto-adaptative dans le cas des valeurs basses de α conduisant à des périodicités locales plus faibles sur les bords d'une zone périodique puisque les groupes des zones apériodiques environnantes sont aussi pris en compte. Si cette diminution n'a pas d'effet négatif sur l'identification du nombre de zones elle en a en revanche sur la bonne classification des groupes mesurée par pC, d'où un effet contraire de α pour ce critère.

Enfin, pour les méthodes utilisant le filtrage post traitement, les meilleurs scores de pC sont obtenus pour les plus grandes valeurs de *minSize* (80% des meilleurs résultats avec *minSize*=6 ou 8) et les plus petites de *minSep* (80% des meilleurs résultats avec *minSep*=2 ou 4). Ce résultat semble indiquer que des zones de taille assez importante séparées par des écarts de faible taille peuvent être identifiées par erreur. Ce phénomène découle de l'utilisation d'un seuil fixe pour l'identification des valeurs hautes et basses des fronts de périodicité, ce que nous avons déjà constaté dans la section 7.2.3 p. 145 avec le seuil global de la méthode de regroupement γ_{BL}.

Influence de la méthode de classification

Le tableau 8.2 présente les résultats obtenus sur les critères zE et pC par les meilleures méthodes sur l'ensemble des paramètres et des scénarios. Le détail des scores est donné dans l'annexe I p. 243.

Pour les deux critères zE et pC, la méthode fm_{2w} utilisant la fonction d'étiquetage m_2

Tableau 8.2 – Moyenne et écart-type de zE et pC des trois meilleures méthodes pour tous les paramètres et tous les scénarios

	1^{er}	2^{nd}	$3^{ème}$
zE (μ, σ)	fm_{2w} $(21\%, 11\%)$	fm_{1w} $(25\%, 13\%)$	fm_2 $(26\%, 7\%)$
pC (μ, σ)	fm_{2w} $(91\%, 5\%)$	m_{2w} $(89\%, 4\%)$	fm_1 $(89\%, 5\%)$

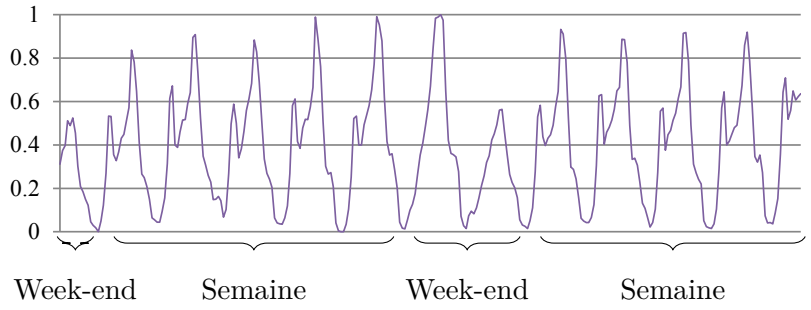


Figure 8.8 – Deux semaines de mesures du CO_2 toutes les heures à la station Châtelet.

définie par l'éq. (8.9) p. 165 avec un filtrage post traitement et un seuil moyen pondéré arrive en première position : la méthode a le plus petite erreur d'identification du nombre de zones et le meilleur de taux de bonne classification des points pour l'ensemble des scénarios et l'ensemble des paramètres.

La méthode d'étiquetage m_2 fournit les meilleurs résultats indépendamment des seuils et du filtrage utilisés, car elle est présente à quatre reprises sur six dans le tableau 8.2. La méthode m_1 semble la deuxième plus pertinente en apparaissant deux fois dans le tableau. La méthode m_3, définie comme la plus optimiste des trois (cf. section 8.3.1 p. 164) paraît donc la moins appropriée dans ce contexte.

En ce qui concerne les paramètres des méthodes, l'utilisation du filtrage est manifestement efficace car représentée cinq fois dans le tableau. De même, la pondération de la valeur de référence renvoie de bons résultats, avec quatre représentants aux deux meilleures positions dans le tableau.

Remarque sur les résultats du scénario S4 Si les scénarios 1, 2 et 3 visent à tester la capacité de LDPE à distinguer les zones périodiques de celles non périodiques, le scénario 4 a pour particularité de tester son aptitude à différencier deux zones périodiques consécutives de période différente.

Bien que LDPE ne permette pas de calculer l'évolution de la période de manière aussi précise que les méthodes temps-fréquence conçues dans ce but (cf. section 4.4 p. 83 et Mallat, 1999), les assez bons résultats obtenus pour le scénario 4 détaillés en annexe I p. 243 montrent toutefois que la méthode est robuste aux changements de période.

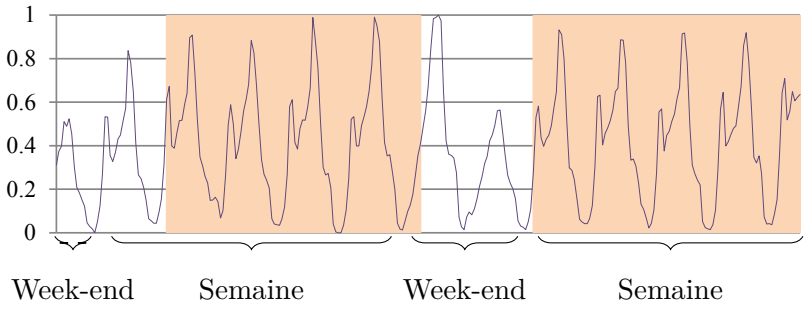


FIGURE 8.9 – Résultat de LDPE sur le jeu de données réel de la figure 8.8

8.5.4 Données réelles

Nous donnons dans ce paragraphe un exemple d'utilisation de la méthode LDPE sur les données réelles présentée dans la section 7.4 p. 157. L'objectif ici est d'observer comment la méthode parvient à discriminer les deux motifs présents dans la série et illustrés sur la figure 8.8, l'un correspondant aux mesures de CO_2 durant la semaine et l'autre à celles du week-end.

Résultats Le résultat obtenu sur ces données à l'aide de la méthode fm_{2w} avec les paramètres *minSep*=2, *minSize* =2, $\alpha = 10\%$ et $\pi_{min} = 0.8$ est illustré sur la figure 8.9. Il montre que la périodicité des groupes identifiés pour la semaine est suffisamment différente de celles du week-end pour les dissocier automatiquement.

Le résultat n'est pas parfait pour autant car le premier pic de la première semaine n'est pas inclus dans la première zone périodique. Cette erreur est due à la méthode de regroupement γ_{es} détaillée dans la section 5.2 p. 98 qui est influencée par les valeurs élevées atteintes par les pics centraux. Ce biais du score d'érosion est également relevé dans la section 7.2.3 p. 148 où les résultats des expériences menées pour la comparaison des méthodes de regroupement sont discutés.

Les zones renvoyées par la méthode sont Z_1=([60,141], 0,92, 24,00) et Z_2=([202,291], 0,83, 22,90). Avec la méthode de rendu linguistique décrite dans la section 8.4 p. 167 et les listes d'intervalle *Moitiés*, *Tiers* et *Cinquièmes* et un seuil d'erreur $\epsilon = 5\%$, les phrases renvoyées sont :

— « Environ du cinquième à la moitié, la série est très périodique (0,92) de période exactement 1 jour »

— « Environ du deuxième tiers à la fin, la série est très périodique (0,83) de période environ 1 jour »

Ces deux phrases illustrent l'efficacité du rendu linguistique de la contextualisation temporelle des zones. En effet, la première zone périodique commence bien au premier cinquième environ de la série et la seconde au second tiers.

Concernant la seconde phrase, il est intéressant de constater qu'elle mentionne une période d'« environ 1 jour » contre celle de la phrase précédente qui est « exactement 1

jour ». La périodicité de la seconde (0,83) est également plus faible que celle de la première (0,92). Cet écart est dû à la présence d'un groupe haut non terminé à la fin de la série, qui introduit une déviation dans le calcul de la régularité des groupes hauts et donc une baisse de la périodicité et une période un peu plus éloignée de 24h. Sur des séries plus grandes, ces effets de bords deviennent négligeables.

8.6 Bilan

Nous avons présenté dans ce chapitre la méthode LDPE qui est une généralisation de la méthode DPE permettant d'en contextualiser les résultats dans le temps. Ainsi, LDPE renvoie une phrase comme *« Environ du deuxième tiers à la fin, la série est très périodique (0,83) de période environ 1 jour »* sur une base de données réelles.

Pour ce faire, nous introduisons la notion de *périodicité locale* qui est une mesure de périodicité comparable à celle utilisée dans DPE mais localisée dans le temps. Cette dernière utilise un test statistique original qui permet la définition de fenêtre dont la largeur s'adapte automatiquement aux données afin d'analyser des sous-parties du jeu de données.

Nous introduisons également les *fronts de périodicité*, qui permettent la mesure des périodicités locales vers la gauche, vers la droite et au centre. A l'aide de ces trois types de fronts de périodicité, nous segmentons la série originale en *zones* périodiques ou non qui sont ensuite rendues linguistiquement en intégrant la localisation de la zone périodique dans le temps en plus des informations de périodicité et de période.

Nous avons testé la méthode sur un grand nombre de jeux de données artificielles illustrant différents cas d'usage de LDPE et montré que la méthode permettait de distinguer les zones périodiques de celles ne l'étant pas ainsi que les zones périodiques adjacentes de périodes différentes. Dans le cas de données réelles liées aux transports parisiens, LDPE identifie automatiquement de zones périodiques liées à la semaine et non périodiques liées au week-end.

Conclusion et perspectives

> Qu'est-ce qu'il y aurait à la fin si tout était au commencement ?
>
> —Victor Hugo, *Notre-Dame de Paris*

Nous présentons dans cette conclusion les différentes contributions de cette thèse ainsi que les perspectives qu'elle ouvre.

Contributions

Nos contributions s'articulent autour des deux niveaux d'analyse retenus pour étudier les résumés linguistiques, détaillés dans les deux parties de la thèse, et concernant respectivement l'interprétabilité des RLF et les résumés de périodicité pour les séries temporelles.

Résumés linguistiques flous

Le premier apport de cette thèse pour les résumés linguistiques flous est la mise en avant de l'importance de leur analyse à un niveau global. En effet, l'approche utilisée pour leur génération, présentée dans le chapitre 1, est basée sur l'instanciation indépendante de phrases dont la somme constitue le résumé renvoyé, sans tenir compte des liens entre ces phrases et de la cohérence du résultat final.

De la même manière, les différentes mesures de qualité présentées dans le chapitre 2 portent principalement sur le vocabulaire utilisé pour créer le résumé et sur les phrases générées. Les seules mesures destinées au résumé sont construites par agrégation de mesures définies pour les phrases et appliquées à chacune de celles du résumé.

Nous avons donc mis en avant l'importance de la mesure de la qualité d'un résumé vu comme un tout et non plus seulement comme un agrégat d'éléments indépendants. Nous avons également montré que les relations entre phrases étaient déterminantes pour l'interprétation du résumé et particulièrement complexes car non définies en amont par l'utilisateur. En effet, si le vocabulaire et les protoformes sont spécifiés manuellement et ainsi assurés d'une certaine cohérence, le processus de génération des phrases est standard et indépendant du vocabulaire et des données : des garanties de cohérence doivent y être intégrées.

Nous nous sommes donc concentrés dans le chapitre 3 sur les propriétés de cohérence d'un résumé, permettant de garantir que les phrases « Beaucoup de jeunes sont grands » et « Pas beaucoup de jeunes sont grands » ne peuvent apparaître au sein du même résumé, ou bien que la phrase « Peu de jeunes ne sont pas grands » a la même valeur de vérité que « La plupart des jeunes sont grands ».

Pour établir le cadre théorique permettant de garantir ces propriétés, nous avons mis en perspective les différentes structures d'opposition existant entre des phrases de complexité croissante : d'abord les phrases simples, puis celles quantifiées avec les quantificateurs classiques, puis avec les quantificateurs généralisés, et enfin utilisant des négations floues.

Cette approche nous a permis d'identifier les 16 cas d'oppositions entre les phrases d'un résumé linguistique flou et de les mettre en relation dans un cube en 4 dimensions décrivant les oppositions qu'elles entretiennent mutuellement parmi l'antonyme, le complément, l'antonyme complément et la dualité. Ce cube a en outre la propriété de généraliser un certain nombre de structures logiques d'opposition existantes.

De plus, la formalisation mise en œuvre pour déterminer ce cube nous a également permis d'étendre la portée des relations de cohérence aux phrases issues des protoformes de type « QRx sont P », qui n'étaient jusqu'alors vérifiées que pour celles basées sur les protoformes « Qx sont P ».

Périodicité des séries temporelles

Le deuxième axe d'analyse des RLF porte plus spécifiquement sur les séries temporelles et leur caractère périodique. Nous avons proposé dans le chapitre 4 un état de l'art des différentes méthodes permettant le calcul de leur période, selon le domaine de représentation sur lequel elles se basent, temporel, fréquentiel, temporo-fréquentiel, symbolique ou autre. L'originalité de cette présentation repose sur la multiplicité des approches qu'elle décrit et qui ne sont que rarement présentées ensemble et de manière synthétique du fait des nombreux domaines desquels elles sont issues.

Cet état de l'art nous a permis de proposer au chapitre 5 une nouvelle approche intuitive et générale pour le calcul de la période appelée DPE, basée sur le fait qu'une série est *périodique si elle alterne de manière régulière des groupes de valeurs hautes et basses, où la régularité est fonction de leurs tailles respectives.*

Cette méthode se distingue des autres sur trois points au moins. D'une part, elle identifie en plus de la période un degré de périodicité pour la série indiquant son caractère périodique ou non et fournit en outre une représentation linguistique de ces deux valeurs. D'autre part, elle réalise l'identification des groupes hauts et bas à l'aide d'une méthode sans paramètre que nous avons proposée. Enfin, elle est basée sur une hypothèse simple et intuitive qui n'utilise aucun modèle a priori.

Afin de rendre la méthode DPE utilisable sur de grands jeux de données, nous l'avons étudiée d'un point de vue algorithmique dans le chapitre 6. Nous nous sommes concentré en particulier sur le calcul du score d'érosion dont la complexité est quadratique dans son implémentation naïve. Nous avons proposé trois autres modes de calcul pour ce dernier,

par niveaux, incrémental, et incrémental par niveaux, et établi les théorèmes montrant qu'ils permettent un calcul exact du score d'érosion. Nous avons également montré que le mode incrémental par niveaux est particulièrement efficace car autorisant le calcul du score d'érosion d'un million de points en 1,5 seconde.

En plus de son efficacité algorithmique, nous avons également illustré dans le chapitre 7 la pertinence de DPE sur des données réelles et artificielles : elle est robuste au bruit, détecte correctement les séries strictement périodiques, retourne un degré de périodicité régulièrement décroissant pour des séries de moins en moins périodiques, évalue leur période avec précision et renvoie la phrase « La période est exactement 1 jour » pour des données réelles contenant un motif quotidien.

Nous avons enfin proposé dans le chapitre 8 la méthode LDPE qui est une extension de la méthode DPE permettant une analyse contextuelle de détection et de rendu linguistique de la période. LDPE génère des phrases comme « Environ du deuxième tiers à la fin, la série est très périodique (0,83) de période environ 1 jour ». Elle fonctionne en appliquant DPE à des sous-séquences de la série d'origine dont les bornes sont déterminées automatiquement à l'aide d'un nouveau test statistique.

Perspectives

Les différentes contributions de cette thèse ouvrent un ensemble de perspectives, détaillées dans les paragraphes suivants et organisées autour des deux axes déjà présentés ainsi qu'un troisième associé à la question de leur mise en œuvre.

Résumés linguistiques flous

Nous avons constaté que les résumés linguistiques flous peuvent à la fois être simplifiés dans leur utilisation et augmentés dans leur interprétabilité, ces deux aspects pouvant être développés simultanément. Concernant leur simplification, nous envisageons différentes approches de génération automatique du vocabulaire, présentées dans le premier paragraphe ci-dessous. Pour leur interprétabilité, les deux paragraphes suivants présentent des pistes pour son amélioration à l'aide de règles de génération et de méthodes de contextualisation.

Définition automatique du vocabulaire La définition de chaque modalité de chaque variable linguistique puis celle des quantificateurs utilisés peut être fastidieuse. Nous pensons qu'il pourrait être intéressant de définir automatiquement les modalités des variables linguistiques comme des partitions de Ruspini construites à l'aide des valeurs min et max des attributs concernés, d'utiliser systématiquement une t-norme probabiliste dont il est montré au chapitre 3 qu'elle permet de garantir les propriétés de cohérences et de générer automatiquement les quantificateurs en fonction du résultat de la fonction de comptage relative, i.e. prenant en compte la taille du jeu de données.

Les quantificateurs définis pourraient être par exemple « Environ 80% » ou « Moins

de 10% » dans le cas de fonctions de comptage renvoyant 0,79 et 0,07 et créés automatiquement sur la base de travaux spécifiques comme ceux de **?**.

En plus de simplifier l'étape de définition du vocabulaire, ils permettent aussi d'éviter les résumés du type « Peu de jeunes sont grands », « Peu de jeunes sont de taille moyenne » et « Peu de jeunes sont petits » qui peuvent prêter à confusion car en ce cas les trois phrases donnent l'impression que tous les individus n'ont pas été pris en compte. Un résumé contenant les phrases « Environ 1/3 des jeunes sont grands », « Environ 1/3 des jeunes sont de taille moyenne » et « Environ 1/3 des jeunes sont petits » est plus clair à cet égard.

Règles de génération L'utilisation de ces quantificateurs avec des partitions de Ruspini permet également de garantir que la somme des décomptes sur l'ensemble des modalités de chaque variable linguistique est bien égale à 1, ce qui est aussi suggéré par Mencar & Fanelli (2008); Gacto et al. (2011). Il est de plus possible de réduire drastiquement le nombre de phrases générées en n'en créant qu'une par attribut tenant compte de ses autres modalités, par exemple « Environ 75% des jeunes sont grands, les autres sont plus petits ». D'autres règles de génération plus sophistiquées peuvent également être envisagées en fonction des seuils à partir desquels une modalité l'emporte sur les autres.

Par ailleurs, certaines phrases peuvent parfois être porteuses d'interprétations induites. Par exemple, « La plupart des gens petits sont mal payés » laisse entendre que les personnes qui ne sont pas petites sont correctement payées. Or il est possible que toutes les personnes, dont les petites, soient mal payées. Une règle commandant d'ignorer l'évaluation des phrases du type « QRx sont P » pour l'ensemble des R si la fonction de comptage de celle basée sur « Qx sont P » est supérieure à un seuil permet d'éviter ce biais.

Enfin, l'utilisation de règles issues des méthodes de génération automatique du langage présentées dans le chapitre 1 peut être mise à profit afin de réduire la taille du résumé et d'en faciliter la lecture. Si par exemple les deux phrases « La plupart des jeunes sont grands » et « La plupart des jeunes sont bien payés » sont générées avec un degré de vérité suffisant, alors elles peuvent être remplacées par la seule phrase « La plupart des jeunes sont grands et bien payés ».

Contextualisation Un vocabulaire défini par l'utilisateur lui est plus simple d'interprétation puisqu'il connaît les valeurs couvertes par les variables linguistiques et les quantificateurs qu'il a définis. Dans le cas d'un vocabulaire créé automatiquement, nous pourrions générer des phrases le décrivant afin de le porter à la connaissance de l'utilisateur. Ainsi, la phrase « Un individu grand mesure au moins 1,80m » pourrait être ajoutée à « La plupart des jeunes sont grands » afin de l'expliciter. Une autre alternative serait de générer une seule phrase du type « La plupart des jeunes sont grands (environ 1,80m et plus) ».

Résumés linguistiques de périodicité

Un grand nombre de perspectives sont également envisageables pour les résumés linguistiques de périodicité. Nous présentons dans les paragraphes suivants celles liées à l'utilisation de DPE pour des séries avec tendances, pour leur segmentation et leur prédiction, puis pour des analyses multirésolution.

Séries avec tendances Nous avons montré dans les expériences du chapitre 7 que DPE a des difficultés à traiter les séries contenant une tendance, car le score d'érosion peut être erroné pour les séries dont les groupes hauts et bas ne contiennent pas au moins une valeur égale à 1 et à 0 respectivement. Cet effet est lié au fait que la dimension temporelle a un poids plus important que la dimension des valeurs lors du calcul du score d'érosion. L'utilisation de fenêtres temporelles et/ou d'une pondération différente permettrait de résoudre cette difficulté.

Exploitation de DPE L'utilisation du score d'érosion pour l'identification des groupes hauts et bas pourrait également être étendue afin de distinguer des niveaux supplémentaires. Ainsi, au lieu de tester $es > \overline{es}$ pour l'attribution des étiquettes H et L uniquement, l'étude de leur différence permettrait d'associer des étiquettes comme *TrèsHaut*, *Similaire* ou *TrèsBas* par exemple, selon que la valeur d'es est très supérieure, environ égale ou très inférieure à celle d'$\overline{es}$.

D'autre part, l'identification des groupes hauts et bas peut également être utilisée pour réaliser des prédictions dans le cas où la périodicité calculée est élevée. En ce cas en effet, l'hypothèse de stationnarité de la série peut être retenue et la suite de la série peut être construite par juxtaposition de groupes hauts et bas calculés comme les valeurs moyennes des groupes déjà identifiés.

Multirésolution La décomposition de la série en groupes hauts et bas permet aussi son analyse multirésolution, selon plusieurs échelles. Par exemple, la série composée de la *suite des tailles de groupes* identifiés peut à son tour être traitée par DPE pour permettre une analyse à une résolution plus large. Dans le cas d'une série composée de la répétition d'un groupe haut large, d'un groupe bas large, d'un groupe haut fin et d'un groupe bas fin, DPE ne détecte pas de périodicité car les groupes hauts et bas ne sont pas de mêmes tailles. En revanche, la suite de leurs tailles est périodique car celles des groupes larges sont des valeurs hautes et celles des groupes fins des valeurs basses.

De manière analogue, une étude sur des résolutions plus précises peut être menée en appliquant DPE à chacun des groupes identifiés. Si par exemple les groupes hauts extraits sont eux-mêmes composés d'une alternance de groupes haut et bas, leur périodicité interne peut être détectée. Supposons une série rectangulaire où chaque plateau est en fait une sinusoïde de faible amplitude. En ce cas, les groupes hauts sont identifiés comme tels mais l'application de DPE sur chacun de ces groupes permettrait de déterminer la fréquence de ces sinusoïdes qui apparaîtraient comme une succession de groupes hauts et bas.

Enfin, une analyse de la forme de chacun des groupes peut également être menée pour déterminer la période de motifs spécifiés et non plus uniquement de groupes hauts et bas. Si la série est composée d'une suite de triangles, son résumé pourrait être alors « La série est constituée de triangles successifs d'amplitude 1 jour environ ».

Mise en œuvre des approches de résumés

Les résumés linguistiques, généraux ou dédiés à la périodicité, pourraient être implémentés en flux. En ce cas, seules les phrases dont le degré de vérité ou tout autre mesure de qualité sont les plus élevées seraient affichées à l'utilisateur au fur et à mesure de la réception des données, provoquant un réordonnancement des phrases déjà présentées.

Dans le cas particulier de DPE, la mise en œuvre en flux fenêtré permettrait potentiellement de résoudre le problème lié au score d'érosion sur les séries avec tendance. Dans ce contexte également, l'étiquette de certains groupes pourrait être mise à jour entre une date et la suivante, permettant ainsi de détecter les changements de « régime » dans les données, de manière analogue aux analyses de type *concept drift*.

Bibliographie

Adalbjornsson, S., Sward, J., Wallin, J., & Jakobsson, A. (2015). Estimating periodicities in symbolic sequences using sparse modeling. *IEEE Trans. Signal Process.*, 63(8), 2142–2150.

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient similarity search in sequence databases. In *Proc. of FODO'93* (pp. 1–15).

Agrawal, R., Lin, K. I., Sawhney, H. S., & Shim, K. (1995). Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-series Databases. In *Proc. of VLDB'95* (pp. 490–501).

Almeida, R. J., Lesot, M.-J., Bouchon-Meunier, B., Kaymak, U., & Moyse, G. (2013). Linguistic summaries of categorical time series patient data. In *Proc. of FUZZ-IEEE'13*.

Alonso, J., Magdalena, L., & González-Rodríguez, G. (2009). Looking for a good fuzzy system interpretability index: An experimental approach. *Int. J. Approx. Reason.*, 51(1), 115–134.

Amini, M. R., Usunier, N., & Gallinari, P. (2005). Automatic text summarization based on wordcClusters and ranking algorithms. In *Proc. of ECIR'05* (pp. 142–156).

Andre-Jonsson, H. & Badal, D. Z. (1997). Using Signature Files for Querying Time-Series Data. In *Proc. of PKDD'97* (pp. 211–220).

Androulakis, I. P. (2005). New approaches for representing, analyzing and visualizing complex kinetic mechanisms. In *Proc. of ESCAPE'05* (pp. 235–240).

Aref, W. G., Elfeky, M. G., & Elmagarmid, A. K. (2004). Incremental, online, and merge mining of partial periodic patterns in time-series databases. *IEEE Trans. Knowl. Data Eng.*, 16(3), 335–345.

Argon, O., Shavitt, Y., & Weinsberg, U. (2013). Inferring the periodicity in large-scale Internet measurements. In *Proc. of INFOCOM'13* (pp. 1672–1680).

Arguelles, L. & Triviño, G. (2013). I-struve: Automatic linguistic descriptions of visual double stars. *Eng. Appl. Artif. Intell.*, 26(9), 2083–2092.

Arora, R., Sethares, W. A., & Bucklew, J. A. (2008). Latent periodicities in genome sequences. *IEEE J. Sel. Top. Signal Process.*, 2(3), 332–342.

Auger, F., Flandrin, P., Lin, Y. T., McLaughlin, S., Meignen, S., Oberlin, T., & Wu, H. T. (2013). Time-Frequency reassignment and synchrosqueezing. *IEEE Signal Process. Mag.*, 30(6), 32–41.

Backmutsky, V., Blaska, J., & Sedlacek, M. (2000). Methods of finding actual signal period time. In *Proc. of IMEKO'00* (pp. 243–248).

Bagnall, A., Ratanamahatana, C., Keogh, E. J., Lonardi, S., & Janacek, G. (2006). A bit level representation for time series data mining with shape based similarity. *Data Min. Knowl. Discov.*, 13(1), 11–40.

Baier, N. U. (2005). *Approximately Periodic Time Series and Nonlinear Structures.* Thèse, EPFL.

Bangham, A. & Marshall, S. (1998). Image and signal processing with mathematical morphology. *Electron. Commun. Eng. J.*, 10(3), 117–128.

Barber, D. (2012). *Bayesian reasoning and machine learning.* Cambridge University Press.

Barro, S., Bugarín, A., Cariñena, P., & Díaz-Hermida, F. (2003). A framework for fuzzy quantification models analysis. *IEEE Trans. Fuzzy Syst.*, 11(1), 89–99.

Barwise, J. & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguist. Philos.*, 4(2), 159–219.

Basseville, M. & Nikiforov, I. (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs.

Berberidis, C., Aref, W. G., Atallah, M. J., Vlahavas, I. P., & Elmagarmid, A. K. (2002). Multiple and partial periodicity mining in time series databases. In *Proc. of ECAI'02* (pp. 370–374).

Bernd, B., Ligges, U., & Weihs, C. (2009). *Frequency estimation by DFT interpolation: a comparison of methods.* Rapport technique, Technische Universität Dortmund.

Blanché, R. (1966). *Les structures intellectuelles, essai sur l'organisation systématique des concepts.* J.Vrin.

Blanco, I., Delgado, M., Martín-Bautista, M., Sánchez, D., & Vila, M. (2002). Quantifier guided aggregation of fuzzy criteria with associated importances. In *Aggreg. Oper.* (pp. 272–287).

Blum, H. (1967). A transformation for extracting new descriptors of shape. *Model. Percept. speech Vis. form*, 19(5), 362–380.

Bodenhofer, U. & Bauer, P. (2005). Interpretability of linguistic variables: a formal account. *Kybernetika*, 41(2), 227–248.

Borda, M., Nafornita, I., Isar, D., & Isar, A. (2005). New instantaneous frequency estimation method based on image processing techniques. In *Proc. of SPIE'05* (pp. 1–11).

Bosc, P., Liétard, L., & Pivert, O. (1998). Extended functional dependencies as a basis for linguistic summaries. In *Proc. of PKDD'98* (pp. 255–263).

Bosc, P., Pivert, O., & Ughetto, L. (1999). On data summaries based on gradual rules. In *Comput. Intell.* (pp. 512–521).

Bouchon-Meunier, B. (2007). *La Logique Floue*. PUF.

Bouchon-Meunier, B. & Moyse, G. (2012). Fuzzy linguistic summaries: where are we, where can we go ? In *Proc. of IEEE CIFEr'12* (pp. 317–324).

Bradford, S. C. (2007). *Time-frequency analysis of systems with changing dynamic properties*. Thèse, California Institute of Technology.

Brazier, K. T. S. (1994). Confidence Intervals from the Rayleigh Test. *Mon. Not. R. Astron. Soc.*, 268(3), 709–712.

Bretthorst, G. L. (1997). *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag.

Brockwell, P. J. & Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Brown, M. (1984). Generalized quantifiers and the square of opposition. *Notre Dame J. Form. Log.*, 25(4), 303–322.

Buccheri, R. (1988). The problem of period detection in sources of hard gamma-ray emission. *Space Sci. Rev.*, 49, 197–206.

Cariñena, P., Bugarín, A., Mucientes, M., & Barro, S. (1999). A language for expressing expert knowledge using fuzzy temporal rules. In *Proc. of EUSFLAT'99* (pp. 171–174).

Cariñena, P., Bugarín, A., Mucientes, M., & Barro, S. (2000). A language for expressing fuzzy temporal rules. *Mathw. Soft Comput.*, 7(2-3), 213–227.

Casasnovas, J. & Torrens, J. (2003). An axiomatic approach to fuzzy cardinalities of finite fuzzy sets. *Fuzzy Sets Syst.*, 133(2), 193–209.

Casillas, J., Cordón, O., Herrera, F., & Magdalena, L. (2003). Interpretability Improvements to Find the Balance Interpretability-Accuracy in Fuzzy Modeling: An Overview. In *Interpret. Issues Fuzzy Model.*, volume 128 (pp. 3–22). Springer Berlin Heidelberg.

Castelltort, A. & Laurent, A. (2015). Extracting fuzzy summaries from NoSQL graph databases. In *Proc. of FQAS'15* (pp. 189–200).

Castillo-Ortega, R., Marín, N., & Sánchez, D. (2011a). Linguistic local change comparison of time series. In *Proc. of FUZZ-IEEE'11* (pp. 2909–2915).

Castillo-Ortega, R., Marín, N., & Sánchez, D. (2011b). Linguistic query answering on data cubes with time dimension. *Int. J. Intell. Syst.*, 26(10), 1002–1021.

Castillo-Ortega, R., Marín, N., Sánchez, D., Corchado, E., & Yin, H. (2009). Fuzzy quantification-based linguistic summaries in data cubes with hierarchical fuzzy partition of time dimension. In *Intell. Data Eng. Autom. Learn.*, volume 5788 (pp. 578–585).

Castillo-Ortega, R., Marín, N., Sánchez, D., & Tettamanzi, A. (2012). Quality assessment in linguistic summaries of data. In *Proc. of IPMU'12* (pp. 285–294).

Castro, P., Almeida, R. J., & Calda Pinto, J. R. (2007). Restoration of Double-Sided Ancient Music Documents with Bleed-Through. In *Prog. Pattern Recognition, Image Anal. Appl.*, volume 4756 (pp. 940–949).

Chassande-Motin, E., Flandrin, P., & Auger, F. (1998). On the statistics of spectrogram reassignment vectors. *Multimed. Syst. Signal Process.*, 9, 335–362.

Chatfield, C. (1996). *The analysis of time series: an introduction, 5th ed.* Chapman & Hall.

Chen, S. & Haralick, R. M. (1995). Recursive erosion, dilation, opening, and closing transforms. *IEEE Trans. Image Process.*, 4(3), 335–345.

Chicharo, J. F. (1996). A new algorithm for improving the accuracy of periodic signal analysis. *IEEE Trans. Instrum. Meas.*, 45(4), 827–831.

Cooley, J. W. & Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Math. Comput.*, 19(90), 297.

Costa, M. J., Finkenstädt, B., Roche, V., Lévi, F., Gould, P. D., Foreman, J., Halliday, K., Hall, A., & Rand, D. A. (2013). Inference on periodicity of circadian time series. *Biostatistics*, 14(4), 792–806.

Cubero, J., Medina, J., Pons, O., & Vila, M. (1999). Data summarization in relational databases through fuzzy dependencies. *Inf. Sci. (Ny).*, 121(3-4), 233–270.

Dale, R., Scott, D., & Di Eugenio, B. (1998). Introduction to the special issue on natural language generation. *Comput. Linguist.*, 24(3), 346–353.

Danlos, L. & El Ghali, A. (2002). A complete integrated NLG system using AI and NLU tools. In *Proc. of COLING'02* (pp. 1–7).

Danlos, L., Meunier, F., & Combet, V. (2011). EasyText: an operational NLG system. In *Proc. of ENLG'11* (pp. 139–144).

Daubechies, I., Lu, J., & Wu, H.-T. (2011). Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool. *Appl. Comput. Harmon. Anal.*, 30(2), 243–261.

Daw, C. S., Finney, C. E. A., & Tracy, E. R. (2003). A review of symbolic analysis of experimental data. *Rev. Sci. Instrum.*, 74(2), 915–930.

De Cheveigné, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4), 1917–1930.

De Soto, A. R. & Trillas, E. (1999). On antonym and negate in fuzzy logic. *Int. J. Intell. Syst.*, 14(3), 295–303.

Delgado, M., Ruiz, M., Sánchez, D., & Vila, M. (2014). Fuzzy quantification: a state of the art. *Fuzzy Sets Syst.*, 242, 1–30.

Delgado, M., Sánchez, D., & Vila, M. (2000). Fuzzy cardinality based evaluation of quantified sentences. *Int. J. Approx. Reason.*, 23(1), 23–66.

Deluca, A. & Termini, S. (1972). A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Inf. Control*, 20(4), 301–312.

Demars, C. (2005). *Représentations bidimensionnelles d'un signal de parole - éléments de monographie.* Rapport technique, LIMSI.

Detyniecki, M. & Marsala, C. (2007). Video rushes summarization by adaptive acceleration and stacking of shots. In *Proc. of TVS'07* (pp. 65–69).

Di-Jorio, L., Laurent, A., & Teisseire, M. (2009). Mining Frequent Gradual Itemsets from Large Databases. In *Proc. of IDA'09* (pp. 297–308).

Díaz-Hermida, F. & Bugarín, A. (2010). Linguistic summarization of data with probabilistic fuzzy quantifiers. In *Proc. of ESTYLF'10* (pp. 255–260).

Dokládal, P. & Dokládalová, E. (2011). Computationally efficient, one-pass algorithm for morphological filters. *J. Vis. Commun. Image Represent.*, 22(5), 411–420.

Dragomiretskiy, K. & Zosso, D. (2014). Variational Mode Decomposition. *IEEE Trans. Signal Process.*, 62(3), 531–544.

Dubois, D. & Prade, H. (1985a). A review of fuzzy set aggregation connectives. *Inf. Sci. (Ny).*, 36(1-2), 85–121.

Dubois, D. & Prade, H. (1985b). Fuzzy cardinality and the modeling of imprecise quantification. *Fuzzy Sets Syst.*, 16(3), 199–230.

Dubois, D. & Prade, H. (2002). Bipolarity in Flexible Querying. In *Proc. of FQAS'02* (pp. 174–182).

Dubois, D. & Prade, H. (2008). Gradual elements in a fuzzy set. *Soft Comput.*, 12(2), 165–175.

Dubois, D. & Prade, H. (2012). From Blanché's hexagonal organization of concepts to formal concept analysis and possibility theory. *Log. Universalis*, 6(1-2), 149–169.

Dubois, D., Prade, H., & Rico, A. (2015). The cube of opposition. A structure underlying many knowledge representation formalisms. In *Proc. of IJCAI'15* (pp. 25–31).

Durnerin, M. (1999). *Une stratégie pour l'interpétation en analyse spectrale. Détection et caractérisation des composantes d'un spectre.* Thèse.

Durrande, N., Hensman, J., Rattray, M., & Lawrence, N. D. (2013). Gaussian process models for periodicity detection. *arXiv:1303.7090 [math.ST].*

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B., & Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. In *Proc. of ICML'13* (pp. 1166–1174).

Dziedzic, M., Kacprzyk, J., & Zadrozny, S. (2013). On some quality criteria of bipolar linguistic summaries. In *Proc. Fed. Conf. Comput. Sci. Inf. Syst.* (pp. 643–646).

Eciolaza, L., Triviño, G., Delgado, B., Rojas, J., & Sevillano, M. (2011). Fuzzy linguistic reporting in driving simulators. In *Proc. IEEE CIVTS'11* (pp. 30–37).: IEEE.

Elfeky, M. G., Aref, W. G., & Elmagarmid, A. K. (2005a). Periodicity detection in time series databases. *IEEE Trans. Knowl. Data Eng.*, 17(7), 875–887.

Elfeky, M. G., Aref, W. G., & Elmagarmid, A. K. (2005b). WARP: time warping for periodicity detection. In *Proc. of ICDM'05* (pp. 138–145).

Elfeky, M. G., Aref, W. G., & Elmagarmid, A. K. (2006). STAGGER: Periodicity Mining of Data Streams Using Expanding Sliding Windows. In *Proc. of ICDM'06* (pp. 188–199).

Emrani, S., Chintakunta, H., & Krim, H. (2014). Real time detection of harmonic structure: a case for topological signal analysis. In *Proc. of ICASSP'14* (pp. 3445–3449).

Enright, J. T. (1965). The search for rhythmicity in biological time-series. *J. Theor. Biol.*, 8(3), 426–468.

Ergün, F., Jowhari, H., & Saglam, M. (2010). Periodicity in Streams. In *Proc. of APPROX-RANDOM'10* (pp. 545–559).

Evans, N. W. D., Mason, J. S., & Roach, M. J. (2002). Noise compensation using spectrogram morphological filtering. In *Proc. of IASTED'02* (pp. 157–161).

Fahlman, G. G. & Ulrych, T. J. (1982). A new method for estimating the power spectrum of gapped data. *Mon. Not. R. Astron. Soc.*, 199(1), 53–65.

Feller, W. (1967). *An introduction to probability theory and its application, 3rd Ed.*, volume 1. John Wiley & Sons, Inc.

Ferreira, L. N. & Zhao, L. (2014). Detecting time series periodicity using complex networks. In *Proc. of BRACIS'14* (pp. 402–407).

Flandrin, P. (1998). *Time-Frequency/Time-Scale Analysis*. Academic press.

Flandrin, P., Auger, F., & Chassande-Motin, E. (2002). Time-Frequency reassignment from principles to algorithms. In *Appl. Time-Frequency Signal Process.* (pp. 179–203).

Flandrin, P., Rilling, G., & Gonçalves, P. (2004). Empirical Mode Decomposition as a Filter Bank. *IEEE Signal Process. Lett.*, 11(2), 112–114.

Foster, G. (1996). Wavelets for period analysis of unevenly sampled time series. *Astron. J.*, 112, 1709–1729.

Frei, M. G. & Osorio, I. (2007). Intrinsic time-scale decomposition: time-frequency-energy analysis and real-time filtering of non-stationary signals. *Proc. R. Soc.*, 463, 321–342.

Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005). Mining data streams: a review. *ACM SIGMOD Rec.*, 34(2), 18–26.

Gabor, D. (1946). Theory of communication. Part 1: The analysis of information. *J. Inst. Electr. Eng. III Radio Commun. Eng.*, 93(26), 439–441.

Gacto, M., Alcalá, R., & Herrera, F. (2011). Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Inf. Sci. (Ny).*, 181(20), 4340–4360.

Gama, J., Žliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), 1–37.

Gamut, L. (1991). *Logic, Language, and Meaning, volume 1: Introduction to Logic*. University of Chicago Press.

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. In *Proc. of SPECOM'05* (pp. 191–194).

Gerhard, D. (2003). *Pitch extraction and fundamental frequency: history and current techniques*. Rapport technique, University of Regina.

Gilles, J. (2013). Empirical Wavelet Transform. *IEEE Trans. Signal Process.*, 61(16), 3999–4010.

Glöckner, I. (1997). *DFS - An axiomatic approach to fuzzy quantification.* Rapport technique, Universität Bielefeld.

Glöckner, I. & Knoll, A. (2001). A formal theory of fuzzy natural language quantification and its role in granular computing. In *Granul. Comput.* (pp. 215–256). Physica-Verlag HD.

Goldberg, E., Driedger, N., & Kittredge, R. I. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2), 45–53.

Goldblum, C. E., Ritter, R. C., & Gillies, G. T. (1988). Using the fast Fourier transform to determine the period of a physical oscillator with precision. *Rev. Sci. Instrum.*, 59(5), 778–782.

Gonçalves, P., Flandrin, P., & Chassande-Motin, E. (1997). Time-frequency methods in time-series data analysis. In *Proc. of GWDAW'97* (pp. 35–46).

Gorard, S. (2005). Revisiting a 90-year-old debate: the advantages of the mean deviation. *Br. J. Educ. Stud.*, 53(4), 417–430.

Grice, H. (1970). Logic and conversation. *Syntax Semant.*, 3, 41–58.

Han, J., Dong, G., & Yin, Y. (1999). Efficient mining of partial periodic patterns in time series database. In *Proc. of ICDE'99* (pp. 106–115).

Han, J., Gong, W., & Yin, Y. (1998). Mining Segment-Wise Periodic Patterns in Time-Related Databases. In *Proc. of KDD'98* (pp. 214 – 218).

Heck, A., Manfroid, J., & Mersch, G. (1985). On period determination methods. *Astron. Astrophys. Suppl. Ser.*, 59, 63–72.

Horn, L. R. (2002). *A Natural History of Negation.* CSLI Publications.

Huang, K.-Y. & Chang, C.-H. (2005). SMCA: a general model for mining asynchronous periodic patterns in temporal databases. *IEEE Trans. Knowl. Data Eng.*, 17(6), 774–785.

Huang, N. E., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N.-C., Tung, C. C., & Liu, H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London A Math. Phys. Eng. Sci.*, 454(1971), 903–995.

Huang, N. E. & Wu, Z. (2008). A review on Hilbert-Huang transform: method and its applications to geophysical studies. *Rev. Geophys.*, 46(2), 1–23.

Hugueney, B. (2006). Cadre général et algorithmes de constructions pour des représentations symboliques adaptatives de séries temporelles. *Rev. Modul.*, 34, 1–12.

Huijse Heise, P., Estevez, P. A., Protopapas, P., Zegers, P., & Principe, J. C. (2012). An information theoretic algorithm for finding periodicities in stellar light curves. *IEEE Trans. Signal Process.*, 60(10), 5135 – 5145.

Hüllermeier, E. (2002). Association rules for expressing gradual dependencies. In *Proc. of PKDD'02* (pp. 200–211).

Hüllermeier, E. (2015). Does machine learning need fuzzy logic ? *Fuzzy Sets Syst.*, 281, 292–299.

Indyk, P., Koudas, N., & Muthukrishnan, S. (2000). Identifying Representative Trends in Massive Time Series Data Sets Using Sketches. In *Proc. of VLDB'00* (pp. 363–372).

Indyk, P. & Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proc. of STOC'98* (pp. 604–613).

Jones, R. H. & Brelsford, W. M. (1967). Time series with periodic structure. *Biometrika*, 54(3), 403–408.

Kacprzyk, J. & Wilbik, A. (2009). Towards an efficient generation of linguistic summaries of time series using a degree of focus. In *Proc. of NAFIPS'09* (pp. 1–6).

Kacprzyk, J., Wilbik, A., & Zadrozny, S. (2008). Linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Fuzzy Sets Syst.*, 159(12), 1485–1499.

Kacprzyk, J. & Yager, R. R. (2001). Linguistic summaries of data using fuzzy logic. *Int. J. Gen. Syst.*, 30(2), 133–154.

Kacprzyk, J., Yager, R. R., & Zadrozny, S. (2000). A fuzzy logic based approach to linguistic summaries of databases. *Int. J. Appl. Math. Comput. Sci.*, 100(4), 813–834.

Kacprzyk, J. & Zadrozny, S. (1994). Fuzzy querying for Microsoft Access. In *Proc. of FUZZ-IEEE'94* (pp. 167–171).

Kacprzyk, J. & Zadrozny, S. (2002). Protoforms of linguistic data summaries: towards more general natural-language-based data mining tools. In *Soft Comput. Syst.* (pp. 417–425).

Kacprzyk, J. & Zadrozny, S. (2005a). Fuzzy linguistic data summaries as a human consistent, user adaptable solution to data mining. In *Do Smart Adapt. Syst. Exist.*, volume 173 (pp. 321–340).

Kacprzyk, J. & Zadrozny, S. (2005b). Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Inf. Sci. (Ny).*, 173(4), 281–304.

Kacprzyk, J. & Zadrozny, S. (2010). Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation. *IEEE Trans. Fuzzy Syst.*, 18(3), 461–472.

Kacprzyk, J. & Zadrozny, S. (2013a). Comprehensiveness and interpretability of linguistic data summaries: A natural language focused perspective. In *Proc. of IEEE SSCI CIHLI'13* (pp. 33–40).

Kacprzyk, J. & Zadrozny, S. (2013b). Grasping the content of web servers logs: a linguistic summarization approach. In *Synerg. Soft Comput. Stat. Intell. Data Anal.*, volume 190 of *Advances in Intelligent Systems and Computing* (pp. 449–457). Springer Berlin Heidelberg.

Kay, S. M. & Marple, L. (1981). Spectrum analysis - A modern perspective. *Proc. IEEE*, 69(11), 1380–1419.

Ke, H.-R., Wang, K.-C., Yang, C.-I., & Chang, K.-F. (2014). Wavelet and Hilbert-Huang transform based on predicting stock forecasting in second-order autoregressive mode. *Int. J. Appl. Phys. Math.*, 4(1), 9–14.

Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings. *Proc. of the IEEE*, 74(11), 1477–1493.

Keogh, E. J., Chu, S., Hart, D., & Pazzani, M. J. (2001). An online algorithm for segmenting time series. In *Proc. of ICDM'01* (pp. 289–296).

Keogh, E. J. & Ratanamahatana, C. (2005). Exact indexing of dynamic time warping. *Knowl. Inf. Syst.*, 7(3), 358–386.

Klement, E. P., Mesiar, R., & Pap, E. (2000). *Triangular norms*. Springer Netherlands.

Knuth, D. E. (1992). Two notes on notation. *Am. Math. Mon.*, 99(5), 403–422.

Kootsookos, P. J. (1999). *A review of the frequency estimation and tracking problems*. Rapport technique, Australian National University.

Kou, K.-I. & Xu, R.-H. (2012). Windowed linear canonical transform and its applications. *Signal Processing*, 92(1), 179–188.

Krawczyk, A. & Krapiec, M. (2010). The permutation test for testing the statistical significance of the power spectrum estimation in dendrochronological analysis. *Geochronometria*, 36(1), 23–29.

Krempl, G., Žliobaite, I., Brzezinski, D., Hüllermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M., & Stefanowski, J. (2014). Open challenges for data stream mining research. *SIGKDD Explor.*, 16(1), 1–10.

Lantuejoul, C. & Maisonneuve, F. (1984). Geodesic methods in quantitative image analysis. *Pattern Recognit.*, 17(2), 177–187.

Larsson, S. (1996). Parameter estimation in epoch folding analysis. *Astron. Astrophys. Suppl.*, 117, 197–201.

Laurent, A., Tijus, C., & Bouchon-Meunier, B. (2004). Fuzzy cognitive quantification. In *Proc. of CogSci'04* (pp. 1–4).

Law, A., Freer, Y., Hunter, J., Logie, R., McIntosh, N., & Quinn, J. (2005). A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *J. Clin. Monit. Comput.*, 19(3), 183–94.

Lefèvre, S. & Claveau, V. (2011). Topic segmentation: application of mathematical morphology to textual data. In *Proc. of ISMM'11* (pp. 472–481).

Leise, T. L., Indic, P., Paul, J. M., & Schwartz, W. J. (2013). Wavelet Meets Actogram. *J. Biol. Rhythms*, 28, 62–68.

Lemaire, V. (2014). Data processing and analytics. In *Proc. of eBISS'14*.

Lemaire, V., Salperwyck, C., & Bondu, A. (2015). A survey on supervised classification on data streams. *Lect. Notes Bus. Inf. Process.*, 205, 88–125.

Lemire, D. (2006). Streaming maximum-minimum filter using no more than three comparisons per element. *Nord. J. Comput.*, 13(4), 328–339.

Leonowicz, Z. (2006). Parametric methods for time-frequency analysis of electric signals. *Pr. Nauk. Inst. Pod. Elektrotechniki i Elektrotechnologii Politech. Wroclawskiej. Monogr.*, 45(15), 1–108.

Leonowicz, Z., Lobos, T., & Schegner, P. (2002). Modern spectral analysis of non-stationary signals in electrical power systems. In *Proc. of PSCC'02* (pp. 1–6).

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining Data Streams. In *Min. massive datasets* (pp. 131–162). Cambridge University Press.

Lesot, M.-J., Moyse, G., & Bouchon-Meunier, B. (2016). Interpretability of fuzzy linguistic summaries. *Fuzzy Sets Syst.*, 292, 307–317.

Lesot, M.-J., Smits, G., & Pivert, O. (2013). Adequacy of a user-defined vocabulary to the data structure. In *Proc. of FUZZ-IEEE'13* (pp. 1–8).

Li, Y., Lin, J., & Oates, T. (2012). Visualizing variable-length time series motifs. In *Proc. of SDM'12* (pp. 895–906).

Li, Z. (2013). *Mining periodicity and object relationship in spatial and temporal data.* Thèse, University of Illinois at Urbana-Champaign.

Li, Z., Wang, J., & Han, J. (2015). ePeriodicity: mining event periodicity from incomplete observations. *IEEE Trans. Knowl. Data Eng.*, 27(5), 1219–1232.

Liétard, L. (2008). A new definition for linguistic summaries of data. In *Proc. of FUZZ-IEEE'08* (pp. 506–511).

Lin, J., Keogh, E. J., Lonardi, S., & Patel, P. (2002). Finding motifs in time series. In *Proc. Work. Temporal Data Mining, KDD'02* (pp. 53–68).

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: an enabling technique. *Data Min. Knowl. Discov.*, 6(4), 393–423.

Liu, Y. & Kerre, E. E. (1998). An overview of fuzzy quantifiers. (I). Interpretations. *Fuzzy Sets Syst.*, 95(1), 1–21.

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., & Ghahramani, Z. (2014). Automatic construction and natural-language description of nonparametric regression models. In *Proc. of AAAI'14* (pp. 1–12).

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *J. Mach. Learn. Res.*, 2, 419–444.

Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci.*, 39(2), 447–462.

Ma, S. & Hellerstein, J. L. (2001). Mining partially periodic event patterns with unknown periods. In *Proc. of ICDE'01* (pp. 205–214).

Maes, S. (1994). *The wavelet transform in signal processing, with application to the extraction of the speech modulation model features.* Thèse, Université Catholique de Louvain.

Mahmood, M. K., Allos, J. E., & Abdul-Karim, M. A. H. (1985). Microprocessor implementation of a fast and simultaneous amplitude and frequency detector for sinusoidal signals. *IEEE Trans. Instrum. Meas.*, 34(3), 413–417.

Mallat, S. (1999). *A wavelet tour of signal processing, 2nd edition.* Academic press.

Manku, G. S., Rajagopalan, S., & Lindsay, B. G. (1998). Approximate medians and other quantiles in one pass and with limited memory. In *Proc. of SIGMOD'98* (pp. 426–435).

Mann, S. & Haykin, S. (1992). Time-frequency perspectives: the 'chirplet' transform. *Proc. of ICASSP'92*, 3, 417–420.

Mannila, H., Toivonen, H., & Inkeri Verkamo, A. (1997). Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, 1(3), 259–289.

Marín, N. & Sánchez, D. (2016). On generating linguistic descriptions of time series. *Fuzzy Sets Syst.*, 285, 6–30.

Marsala, C. & Bouchon-Meunier, B. (2003). Choice of a method for the construction of fuzzy decision trees. In *Proc. of FUZZ-IEEE'03*, volume 1 (pp. 584–589).

Martin, T. P. (2013). The X-mu representation of fuzzy sets - Regaining the excluded middle. In *Proc. of UKCI'13* (pp. 67–73).

Mattioli, J. & Schmitt, M. (1992). Inverse problems for granulometries by erosion. *J. Math. Imaging Vis.*, 2(2-3), 217–232.

Mencar, C. & Fanelli, A. (2008). Interpretability constraints for fuzzy information granulation. *Inf. Sci. (Ny).*, 178(24), 4585–4618.

Mendel, J. M. (2000). Uncertainty, fuzzy logic, and signal processing. *Signal Processing*, 80, 913–933.

Méndez Núñez, S. & Triviño, G. (2010). Combining semantic web technologies and computational theory of perceptions for text generation in financial analysis. In *Int. Conf. Fuzzy Syst.* (pp. 1–8).

Mesiar, R. & Stupnanová, A. (2015). Open problems from the 12th international conference on fuzzy set theory and its applications. *Fuzzy Sets Syst.*, 261, 112–123.

Millioz, F., Huillery, J., & Martin, N. (2006). Short time Fourier transform probability distribution for time-frequency segmentation. In *Proc. of ICASSP'06* (pp. 448–451).

Minnen, D., Starner, T., Essa, I. A., & Isbell Jr, C. R. (2007). Improving activity discovery with automatic neighborhood estimation. In *Proc. of IJCAI'07* (pp. 2814–2819).

Moore, P., Carranza, R., & Johns, A. (1994). A new numeric technique for high-speed evaluation of power system frequency. *IEE Proc. - Gener. Transm. Distrib.*, 141(5), 529–536.

Moore, R. (1963). *Interval arithmetic and automatic error analysis in digital computing.* Thèse, Stanford University.

Morard, V., Dokládal, P., & Decencière, E. (2012). One-dimensional openings, granulometries and component trees in O(1) per pixel. *IEEE J. Sel. Top. Signal Process.*, 6(7), 840–848.

Mörchen, F. & Ultsch, A. (2005). Optimizing time series discretization for knowledge discovery. In *Proc. of KDD'05* (pp. 1–6).

Moretti, A. (2011). From the "logical square" to the "logical poly-simplexes". A quick survey of what happened in between. In *New Perspect. Sq. Oppos.* (pp. 1–21).

Morinaka, Y., Yoshikawa, M., Amagasa, T., & Uemura, S. (2001). The L-index: An indexing Structure for Efficient Subsequence Matching in Time Sequence Databases. In *Proc. of PAKDD'01* (pp. 51–60).

Mostowski, A. (1957). On a generalization of quantifiers. *Fundam. Math.*, 44(1), 12–36.

Moyse, G. & Lesot, M.-J. (2014). Fast and incremental erosion score computation. In *Proc. of IPMU'14* (pp. 376–385).

Moyse, G. & Lesot, M.-J. (2015). Linguistic summaries of locally periodic time series. *Fuzzy Sets Syst.*, 285, 94–117.

Moyse, G., Lesot, M.-J., & Bouchon-Meunier, B. (2013a). Linguistic summaries for periodicity detection based on mathematical morphology. In *Proc. of IEEE SSCI FOCI'13* (pp. 106–113).

Moyse, G., Lesot, M.-J., & Bouchon-Meunier, B. (2013b). Mathematical morphology tools to evaluate periodic linguistic summaries. In *Proc. of FQAS'13* (pp. 257–268).

Moyse, G., Lesot, M.-J., & Bouchon-Meunier, B. (2015). Oppositions in fuzzy linguistic summaries (best student paper). In *Proc. of FUZZ-IEEE'15*.

Murinová, P. & Novák, V. (2014). Analysis of generalized square of opposition with intermediate quantifiers. *Fuzzy Sets Syst.*, 242, 89–113.

Najman, L. & Talbot, H. (2013). *Mathematical Morphology*. John Wiley & Sons.

Nason, G. P. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, 75, 879–904.

Newstead, S. E., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Appl. Ergon.*, 18(3), 178–182.

Novák, V. (2008). A formal theory of intermediate quantifiers. *Fuzzy Sets Syst.*, 159(10), 1229–1246.

Novák, V. (2015). Mining information from time series in the form of natural language expressions. In *Proc. of IFSA-EUSFLAT'15* (pp. 1–7).

Novák, V., Štěpnička, M., Perfilieva, I., & Pavliska, V. (2008). Analysis of periodical time series using soft computing methods. In *Proc. of FLINS'08*.

Nussbaumer Knaflic, C. (2015). *Storytelling with data: a data visualization guide for business professionals*. Wiley.

Oppenheim, A. V. & Schafer, R. W. (2004). From frequency to quefrency: a history of the cepstrum. *IEEE Signal Process. Mag.*, 21(5), 95–106.

Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999). *Discrete-time signal processing, 2nd edition*. Prentice Hall.

Otunba, R. & Lin, J. (2014). APT: approximate period detection in time series. In *Proc. of SEKE'14* (pp. 490–494).

Otunba, R., Lin, J., & Senin, P. (2014). MBPD: motif-based period detection. In *Proc. of PAKDD'14* (pp. 793–804).

Oudni, A., Lesot, M.-J., & Rifqi, M. (2013). Characterisation of gradual itemsets based on mathematical morphology tools. In *Proc. of EUSFLAT'13* (pp. 826–833).

Ozden, B., Ramaswamy, S., & Silberschatz, A. (1998). Cyclic association rules. In *Proc. of ICDE'98* (pp. 412–421).

Paetz, J., Erz, K., Arlt, B., & Hanisch, E. (2003). The MEDAN database: patients with abdominal septic shock. *Zentralbl. Chir.*, 128(4), 298–303.

Papadimitriou, S., Brockwell, A., & Faloutsos, C. (2003). Adaptive, hands-off stream mining. In *Proc. of VLDB'03* (pp. 560–571).

Pardo-Igúzquizaa, E. & Rodriguez-Tovarb, F. J. (2000). The permutation test as a non-parametric method for testing the statistical significance of power spectrum estimation in cyclostratigraphic research. *Earth Planet. Sci. Lett.*, 181(1-2), 175–189.

Pecht, J. (1985). Speeding-up successive Minkowski operations with bit-plane computers. *Pattern Recognit. Lett.*, 3(2), 113–117.

Peeters, G., La Burthe, A., & Rodet, X. (2002). Toward automatic music audio summary generation from signal analysis. In *Proc. of ISMIR'02* (pp. 1–7).

Percival, D. B. & Walden, A. T. (1998). *Spectral analysis for physical applications.*

Peterson, P. (1979). On the logic of "Few", "Many" and "Most". *Notre Dame J. Form. Log.*, 20(1), 155–179.

Pilarski, D. (2010). Linguistic summarization of databases with Quantirius: a reduction algorithm for generated summaries. *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, 18(3), 305–331.

Plautz, J. D., Straume, M., Stanewsky, R., Jamison, C. F., Brandes, C., Dowse, H. B., Hall, J. C., & Kay, S. A. (1997). Quantitative analysis of Drosophila period gene transcription in living animals. *J. Biol. Rhythms*, 12(3), 204–217.

Portet, F., Reiter, E., Hunter, J., & Sripada, S. (2007). Automatic generation of textual summaries from neonatal intensive care data. In *Artif. Intell. Med.* (pp. 227–236).

Poulimenos, A. G. & Fassois, S. D. (2006). Parametric time-domain methods for non-stationary random vibration modelling and analysis - A critical survey and comparison. *Mech. Syst. Signal Process.*, 20(4), 763–816.

Preotiuc-Pietro, D. & Cohn, T. (2013). A temporal model of text periodicities using Gaussian Processes. In *Proc. of EMNLP'13* (pp. 977–988).

Pulkkinen, P. & Koivisto, H. (2010). A dynamically constrained multiobjective genetic fuzzy system for regression problems. *IEEE Trans. Fuzzy Syst.*, 18(1), 161–177.

Qu, Y., Wang, C., & Sean Wang, X. (1998). Supporting fast search in time series for movement patterns in multiple scales. In *Proc. of CIKM'98* (pp. 251–258).

Quinn, B. (1994). Estimating frequency by interpolation using Fourier coefficients. *IEEE Trans. Signal Process.*, 42(5), 1264–1268.

Ralescu, D. (1995). Cardinality, quantifiers, and the aggregation of fuzzy criteria. *Fuzzy Sets Syst.*, 69(3), 355–365.

Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2016). Data discretization: taxonomy and big data challenge. *WIREs Data Min. Knowl. Discov.*, 6(1), 5–21.

Ramos-Soto, A., Bugarín, A., & Barro, S. (2016). On the role of linguistic descriptions of data in the building of natural language generation systems. *Fuzzy Sets Syst.*, 285, 31–51.

Raschia, G. & Mouaddib, N. (2000). A fuzzy-based heuristic measure evaluating quality of a concept partition: application to SaintEtiQ, a database summarization system. In *Proc. of FUZZ-IEEE'00*, volume 2 (pp. 957–960).

Raschia, G. & Mouaddib, N. (2002). StEtiQ: a fuzzy set-based approach to database summarization. *Fuzzy Sets Syst.*, 129(2), 137–162.

Rashidul Hasan, M. A. F. M. & Shimamura, T. (2012). A fundamental frequency extraction method based on windowless and normalized autocorrelation functions. In *Proc. of WSSEAS'12* (pp. 305–309).

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian processes for Machine Learning*. MIT Press.

Rasmussen, D. & Yager, R. R. (1997). Summary SQL - A fuzzy tool for data mining. *Intell. Data Anal.*, 1(1), 49–58.

Rasmussen, D. & Yager, R. R. (1999). Finding fuzzy and gradual functional dependencies with SummarySQL. *Fuzzy Sets Syst.*, 106(2), 131–142.

Ratanamahatana, C., Keogh, E. J., Bagnall, A., & Lonardi, S. (2005). A novel bit level time series representation with implications for similarity search and clustering. In *Proc. of PAKDD'05* (pp. 771–777).

RATP (2012). Qualité de l'air mesurée dans nos stations - http://data.ratp.fr/.

Refinetti, R., Cornélissen, G., & Halberg, F. (2007). Procedures for numerical analysis of circadian rhythms. *Biol. Rhythm Res.*, 38, 275–325.

Rehman, N. & Mandic, D. P. (2010). Multivariate empirical mode decomposition. *Proc. R. Soc. London A Math. Phys. Eng. Sci.*, 466, 1291–1302.

Reiter, E. (1996). Building Natural-Language Generation Systems. *arXiv:cmp-lg/9605002*.

Reiter, E. & Dale, R. (1997). Building applied natural language generation systems. *J. Nat. Lang. Eng.*, 3(1), 57–87.

Reiter, E. & Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press.

Renard, X., Rifqi, M., Erray, W., & Detyniecki, M. (2015). Random-shapelet: an algorithm for fast shapelet discovery. In *Proc. of DSAA'15* (pp. 1–10).

Rocacher, D. & Bosc, P. (2005). The set of fuzzy rational numbers and flexible querying. *Fuzzy Sets Syst.*, 155(3), 317–339.

Ros, M., Molina-Solana, M., Delgado, M., Fajardo, W., & Vila, M. A. (2016). Transcribing Debussy's Syrinx dynamics through linguistic description: the MUDELD algorithm. *Fuzzy Sets Syst.*, 285, 199–216.

Rosenblum, M. & Kurths, J. (1995). A simple test for hidden periodicity in time series data. *Int. J. Bifurc. Chaos*, 5(1), 265–269.

Ross, M., Shaffer, H., Cohen, A., Freudberg, R., & Manley, H. (1974). Average magnitude difference function pitch extractor. *IEEE Trans. Acoust.*, 22(5), 353–362.

Rousseeuw, P. & Croux, C. (1993). Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.*, 88(424), 1273–1283.

Ruspini, E. H. (1969). A new approach to clustering. *Inf. Control*, 15(1), 22–32.

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H., & Cousens, D. R. (1988). SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instruments Methods Phys. Res. Sect. B*, 34(3), 396–402.

Saint-Paul, R. & Raschia, G. (2002). Mining a commercial banking data set: the SaintEtiQ approach. In *Proc. IEEE SMC'02*, volume 2 (pp. 488 – 493).

Sánchez, D., Delgado, M., & Vila, M. (2009). Fuzzy Quantification Using Restriction Levels. In *Proc. Int. Work. Fuzzy Log. Appl.* (pp. 28–35).

Sanchez-Valdes, D., Alvarez-Alvarez, A., & Triviño, G. (2016). Dynamic linguistic descriptions of time series applied to self-track the physical activity. *Fuzzy Sets Syst.*, 285, 162–181.

Sanchez-Valdes, D. & Triviño, G. (2013). Computational perceptions of uninterpretable data. A case study on the linguistic modeling of human gait as a quasi-periodic phenomenon. *Fuzzy Sets Syst.*, 253, 101–121.

Santamaria, I., Pokharel, P. P., & Principe, J. C. (2006). Generalized correlation function: definition, properties, and application to blind equalization. *IEEE Trans. Signal Process.*, 54(6), 2187–2197.

Sant'Anna, A. & Wickstrom, N. (2011). Symbolization of time-series: an evaluation of SAX, Persist, and ACA. In *Proc. of CISP'11* (pp. 2223–2228).

Sayeed, A. M. & Jones, D. L. (1995). Optimal detection using bilinear time-frequency and time-scale representations. *IEEE Trans. Signal Process.*, 43(12), 2872–2883.

Scargle, J. D. (1982). Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J.*, 263, 835–853.

Schwarzenberg-Czerny, A. (1989). On the advantage of using analysis of variance for period search. *Mon. Not. R. Astron. Soc.*, 241, 153–165.

Sejdić, E., Djurović, I., & Stanković, L. (2011). Fractional Fourier transform as a signal processing tool: An overview of recent developments. *Signal Processing*, 91(6), 1351–1369.

Serra, J. (1983). *Image Analysis and Mathematical Morphology.* Academic Press.

Serra, J. (1986). Introduction to mathematical morphology. *Comput. Vision, Graph. Image Process.*, 35(3), 283–305.

Sethares, W. A. & Staley, T. W. (1999). Periodicity transforms. *IEEE Trans. Signal Process.*, 47(11), 2953–2964.

Shaker, A. & Hüllermeier, E. (2015). Recovery analysis for adaptive learning from non-stationary data streams: Experimental design and case study. *Neurocomputing*, 150, 250–264.

Shin, K. & Hammond, J. K. (2008). *Fundamentals of signal processing for sound and vibration engineers.* John Wiley & Sons.

Silverman, B. W. (1998). Wavelets in statistics: some recent developments. In *Proc. of COMPSTAT'98* (pp. 15–26).

Sokolove, P. G. & Bushell, W. N. (1978). The chi square periodogram: Its utility for analysis of circadian rhythms. *J. Theor. Biol.*, 72(1), 131–160.

Sripada, S., Reiter, E., & Davy, I. (2003). SumTime-Mousam: configurable marine weather forecast generator. *Expert Updat.*, 6(3), 4–10.

Štěpnička, M., Dvorák, A., Pavliska, V., & Vavríčková, L. (2010). Linguistic approach to time series analysis and forecasts. In *Proc. of FUZZ-IEEE'10* (pp. 1–9).

Štěpnička, M., Dvorák, A., Pavliska, V., & Vavríčková, L. (2011). A linguistic approach to time series modeling with the help of F-transform. *Fuzzy Sets Syst.*, 180(1), 164–184.

Stoica, P. (1993). List of references on spectral line analysis. *Signal Processing*, 31(3), 329–340.

Stoica, P. & Moses, R. (2005). *Spectral analysis of signals.* Prentice Hall.

Stopa, J. E. & Cheung, K. F. (2014). Periodicity and patterns of ocean wind and wave climate. *J. Geophys. Res. Ocean.*, 119(8), 5563–5584.

Straume, M. (2004). DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. *Methods Enzymol.*, 383, 149–166.

Sun, Y., Chan, K. L., & Krishnan, S. M. (2005). Characteristic wave detection in ECG signal using morphological transform. *BMC Cardiovasc. Disord.*, 5(1), 28.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In Elsevier (Ed.), *Speech coding Synth.* (pp. 495–518).

Tanaka, Y., Iwamoto, K., & Uehara, K. (2005). Discovery of Time-series motif from multi-dimensional data based on MDL principle. *Mach. Learn.*, 58, 269–300.

Thomas, K. E., Sripada, S., & Noordzij, M. L. (2012). Atlas.txt: exploring linguistic grounding techniques for communicating spatial information to blind users. *Univers. Access Inf. Soc.*, 11(1), 85–98.

Thomson, D. (1982). Spectrum estimation and harmonic analysis. *Proc. IEEE*, 70(9), 1055–1096.

Torrence, C. & Compo, G. P. (1998). A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.*, 79(1), 61–78.

Triviño, G. & Sanchez-Valdes, D. (2015). Generation of linguistic advices for saving energy: architecture. In *Proc. of TPNC'15* (pp. 83–94).

Triviño, G. & Sugeno, M. (2013). Towards linguistic descriptions of phenomena. *Int. J. Approx. Reason.*, 54(1), 22–34.

Tsuji, M. & Yamada, E. (2001). A wavelet approach to real time estimation of power system frequency. In *Proc. of SICE'01* (pp. 58–65).

Turpin-Dhilly, S. & Botte-Lecoq, C. (1998). Application of fuzzy mathematical morphology for pattern classification. In *Adv. Data Sci. Classif.* (pp. 125–130).

Vaidyanathan, P. P. (2014). Ramanujan-sum expansions for finite duration (FIR) sequences. In *Proc. of ICASSP'14* (pp. 4933–4937).

Vaidyanathan, P. P. & Pal, P. (2014). The Farey-dictionary for sparse representation of periodic signals. In *Proc. of ICASSP'14* (pp. 360–364).

Van der Heide, A. & Triviño, G. (2009). Automatically generated linguistic summaries of energy consumption data. In *Proc. of ISDA'09* (pp. 553–559).

Van Droogenbroeck, M. & Buckley, M. (2005). Morphological erosions and openings: fast algorithms based on anchors. *J. Math. Imaging Vis.*, 22(2-3), 121–142.

Vaughan, S. (2005). A simple test for periodic signals in red noise. *Astron. Astrophys.*, 431(1), 391–403.

Vaughan, S. (2010). A Bayesian test for periodic signals in red noise. *Mon. Not. R. Astron. Soc.*, 402, 307–320.

Vincent, L. (1992). Morphological algorithms. In *Math. Morphol. Image Process.* (pp. 255–288).

Vincent, L. & Dougherty, E. R. (1994). Morphological segmentation for textures and particles. In *Digit. Image Process. Methods* (pp. 43–102).

Vincent, L. & Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. PAMI*, 13(6), 583–598.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: a neural image caption generator. In *Proc. of CVPR'15* (pp. 3156–3164).

Vlachos, M., Yu, P. S., & Castelli, V. (2005). On periodicity detection and structural periodic similarity. In *Proc. of SIAM'05*, volume 119 (pp. 449–460).

Wang, Q. & Megalooikonomou, V. (2008). A dimensionality reduction technique for efficient time series similarity analysis. *Inf. Syst.*, 33(1), 115–132.

Wang, X., Tang, H., & Zhao, X. (2005). Noisy Speech Pitch Detection Based on Mathematical Morphology and Weighted MACF. In *Proc. of Sinobiometrics'05*, volume 3338 (pp. 594–601).

Westerstahl, D. (2012). Classical vs. modern squares of opposition, and beyond. In *Sq. Oppos. a Gen. Framew. Cogn.* (pp. 195–229). Peter Lang.

Wilbik, A. (2010). *Linguistic summaries of time series using fuzzy sets and their application for performance analysis of mutual funds.* Thèse, Polish Academy of Sciences.

Wilbik, A. & Keller, J. M. (2012). A distance metric for a space of linguistic summaries. *Fuzzy Sets Syst.*, 208, 79–94.

Wu, Z. & Huang, N. E. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data Anal.*, 1(1), 1–41.

Wygralak, M. (1986). Fuzzy cardinals based on the generalized equality of fuzzy subsets. *Fuzzy Sets Syst.*, 18(2), 143–158.

Wygralak, M. (1997). On the best scalar approximation of cardinality of a fuzzy set. *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, 5(6), 681–687.

Yager, R. R. (1982). A new approach to the summarization of data. *Inf. Sci. (Ny).*, 28(1), 69–86.

Yang, J., Wang, W., & Yu, P. S. (2000). Mining asynchronous periodic patterns in time series data. In *Proc. of KDD'00* (pp. 275–279).

Yang, R. & Su, Z. (2010). Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*, 26, 168–174.

Yseop (2011). Faire parler les chiffres automatiquement - http://www.yseop.com.

Yu, J., Reiter, E., Hunter, J., & Sripada, S. (2003). SumTime-Turbine: a knowledge-based system to communicate gas turbine time-series data. In *Proc. IEA/AIE '03* (pp. 379–384).

Zadeh, L. A. (1965). Fuzzy sets. *Inf. Control*, 8(3), 338–353.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning - I. *Inf. Sci. (Ny).*, 8(3), 199–249.

Zadeh, L. A. (1979). A Theory of Approximate Reasoning. *Mach. Intell.*, 9, 149–194.

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Comput. Math. with Appl.*, 9(1), 149–184.

Zadeh, L. A. (1985). Syllogistic reasoning in fuzzy logic and its application to usuality and reasoning with dispositions. *IEEE Trans. Syst. Man, Cybern.*, SMC-15(6), 754–763.

Zadeh, L. A. (2002). From computing with numbers to computing with words: from manipulation of measurements to manipulation of perceptions. *Int. J. Appl. Math. Comput. Sci.*, 12(3), 307–324.

Zadrozny, S. & Kacprzyk, J. (2007). Summarizing the contents of web server logs: a fuzzy linguistic approach. In *Proc. of FUZZ-IEEE'07* (pp. 1–6).

Zahorian, S. A. & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *J. Acoust. Soc. Am.*, 123(6), 4559–4571.

Zayezdny, A., Adler, Y., & Druckmann, I. (1992). Short time measurement of frequency and amplitude in the presence of noise. *IEEE Trans. Instrum. Meas.*, 41(3), 397–402.

Zhou, F., Torre, F., & Hodgins, J. K. (2008). Aligned cluster analysis for temporal segmentation of human motion. In *Proc. of FG'08* (pp. 1–7).

Zhou, S.-M. & Gan, J. Q. (2008). Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling. *Fuzzy Sets Syst.*, 159(23), 3091–3131.

Zielinski, T., Moore, A. M., Troup, E., Halliday, K. J., & Millar, A. J. (2014). Strengths and limitations of period estimation methods for circadian data. *PLoS One*, 9(5), 1–26.

Zucker, S. (2015). Detection of periodicity based on serial dependence of phase-folded data. *Mon. Not. R. Astron. Soc.*, 449(3), 2723–2733.

Annexes

Annexe A

Systèmes de génération de résumés linguistiques

Les deux tableaux ci-dessous listent des systèmes de génération de résumés linguistiques, dans un contexte de GAT pour le premier et de RLF pour le second (cf. section 1.1 p. 7). Ces tableaux ont été établis en 2012 pour l'article (Bouchon-Meunier & Moyse, 2012). Ramos-Soto et al. (2016) en donnent un état de l'art plus récent.

	Description	**Application**	**Réf.**
Yseop	Génération automatique de comptes-rendus basés sur des règles métiers.	Rapport de la santé financière d'une entreprise sur la base de son bilan et de son compte de résultats.	Yseop (2011)
SumTime Mousam	Techniques de production de texte en langage naturel sur la base de textes à trous.	Prédictions météorologiques pour des plates-formes d'extraction pétrolière en mer du Nord.	Sripada et al. (2003)
SumTime Turbine		Résumés textuels de l'activité des turbines d'une usine.	Yu et al. (2003)
EasyText	Traitement de données numériques dédié à la génération d'analyses.	Utilisé dans une compagnie de sondage d'opinions.	Danlos et al. (2011)
FOG	Générateur de rapports en français et en anglais.	Génération de bulletins météorologiques.	Goldberg et al. (1994)
BT-45	Projet BabyTalk, dédié à la génération de description textuelle de mesures en milieu médical.	Utilisé sur 45 minutes d'enregistrement de différentes machines dans un bloc de réanimation néonatale.	Portet et al. (2007)

	Description	**Application**	**Réf.**
FQuery	Calcul de la valeur de vérité d'un protoforme.	Résumés linguistiques de fichiers log d'un serveur Web. Module MS Access.	Kacprzyk & Zadrozny (1994); Zadrozny & Kacprzyk (2007)
Quantirius	Système interactif de fouille et d'évaluation de résumés linguistiques dans une base de données.	Mise à jour de relations entre la durée des abonnements, la fréquence d'achats et de vente et les opinions d'utilisateurs sur un site d'enchères. Résumé de l'évolution d'un indice boursier sur la place boursière de Varsovie.	Pilarski (2010)
Summary SQL	Langage de requête destiné à identifier les dépendances fonctionnelles floues et graduelles.	/	Rasmussen & Yager (1999)
SaintEtiQ	Résumé des données par la génération d'un arbre de concepts, le plus général à la racine.	Étude du comportement des clients d'une banque.	Saint-Paul & Raschia (2002)
/	Protoformes étendus de résumés flous avec prise en compte de tendances identifiées dans les données.	Analyse de 8 années de cotations journalières d'un fond d'investissement	Kacprzyk et al. (2008)
/	Utilisation de l'approche GLMP pour la production d'un compte-rendu.	Analyse des données financières de plusieurs compagnies énergétiques espagnoles entre 2005 et 2009.	Méndez Núñez & Triviño (2010)

Annexe B

Exemple d'application de génération de RLF

B.1 Contexte

A l'occasion de l'UE PLDAC 2015, nous avons encadré deux étudiants de M1 qui ont réalisé un outil de génération de RLF. Comme détaillé au chapitre 1, ce dernier prend en entrée des données, des variables linguistiques ainsi qu'un ensemble de paramètres et construit toutes les phrases possibles basées sur le protoforme « QRx sont P ».

Le jeu de données est issu d'un logiciel réalisé en début de thèse qui collecte, toutes les semaines, les classements des meilleures ventes de livres sur les sites Amazon en France, au Royaume-Uni, aux États-Unis et au Canada (Bouchon-Meunier & Moyse, 2012).

Les données sont analysées afin de déterminer s'il existe une corrélation entre le nombre de pages d'un livre et ses classements dans les meilleures ventes : les petits livres se vendent-ils mieux ?

Données Le nombre de pages et les classements de 13 500 livres ont été collectés et un score par livre a été attribué de sorte que chaque fois qu'il apparaît dans un classement hebdomadaire, mensuel ou annuel, son score est augmenté de 100 moins sa position (les classements contiennent 100 livres). Par exemple, un livre qui apparaît à la 4ème position et à la 23ème dans deux classements aura donc un score de $100 - 4 + 100 - 23 = 173$.

Variables linguistiques Les VL *Score* et *Pages* ont été définies comme illustré sur la figure B.1. Les quantificateurs utilisés pour l'expérience sont donnés sur la figure B.2.

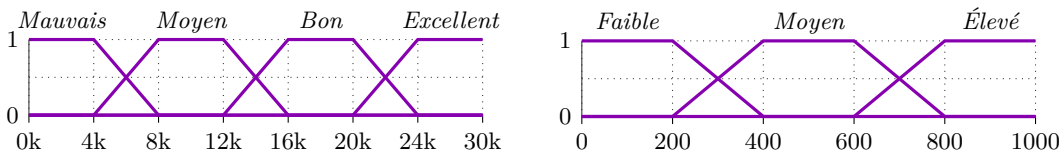


FIGURE B.1 – Les VL *Score* (à gauche) et nombre de *Pages* (à droite)

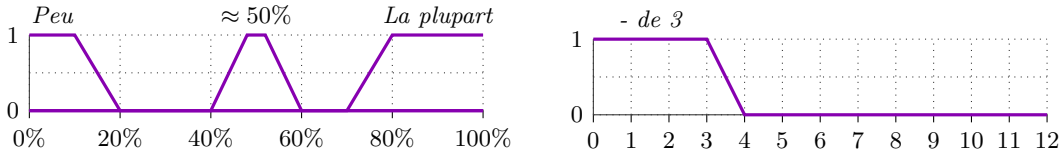


FIGURE B.2 – Quantificateurs relatifs (à gauche) et absolu (à droite)

Paramètres Les résumés retenus ont une valeur de vérité supérieure ou égale à 1/2, calculée avec la formule donnée dans l'éq. (1.2) p. 14 et la t-norme de Zadeh. De plus, leur degré de focus est supérieur ou égal à 10%.

B.2 Résultats

Les résumés suivants ont été générés après exécution du programme :

```
1. La plupart des livres ayant un nombre de pages faible sont mauvais (1,00)
2. Peu de livres ayant un nombre de pages faible sont moyens (1,00)
3. Peu de livres ayant un nombre de pages faible sont bons (1,00)
4. Peu de livres ayant un nombre de pages faible sont excellents (1,00)
5. Moins de 3 livres ayant un nombre de pages faible sont excellents (1,00)
6. Moins de 3 livres ayant un nombre de pages faible sont bons (1,00)
7. Moins de 3 livres ayant un nombre de pages faible sont moyens (1,00)

8. La plupart des livres ayant un nombre de pages moyen sont mauvais (1,00)
9. Peu de livres ayant un nombre de pages moyen sont moyens (1,00)
10. Peu de livres ayant un nombre de pages moyen sont bons (1,00)
11. Peu de livres ayant un nombre de pages moyen sont excellents (1,00)
12. Moins de 3 livres ayant un nombre de pages moyen sont moyens (1,00)
13. Moins de 3 livres ayant un nombre de pages moyen sont bons (1,00)
14. Moins de 3 livres ayant un nombre de pages moyen sont excellents (1,00)

15. Environ la moitié des livres mauvais ont un nombre de pages faible (0,78)
16. Environ la moitié des livres mauvais ont un nombre de page moyen (1,00)
17. Peu de  livres mauvais ont un nombre de page élevé (1,00)

18. Moins de 3 livres ayant un nombre de pages élevé sont mauvais (1,00)
19. Moins de 3 livres ayant un nombre de pages élevé sont moyens (1,00)
20. Moins de 3 livres ayant un nombre de pages élevé sont bons (1,00)
21. Moins de 3 livres ayant un nombre de pages élevé sont excellents (1,00)
```

B.3 Analyse

De nombreuses remarques peuvent être réalisées sur la base de ce test simple :

— le nombre de phrases générées nuit considérablement à l'interprétabilité du résumé dans son ensemble. Les techniques de GAT liées à la lexicalisation et à l'agrégation de phrases permettraient d'améliorer le rendu textuel (cf. section 1.1.2 p. 9).

— le choix des termes retenus pour les modalités des VL est critique. Même s'ils semblent clairs au niveau de la VL, ils ne le sont plus nécessairement une fois utilisés dans des phrases. Ils peuvent également porter un sens différents de celui visé initialement. La phrase 1 par exemple peut être interprétée comme faisant référence à la qualité du livre et non à son classement dans les meilleures ventes.

— L'utilisation du quantificateur absolu *Moins de 3* n'apporte pas à la compréhension des résultats dans le cas d'un jeu de données de plusieurs milliers d'éléments comme c'est la cas ici. Le quantificateur est systématiquement redondant vis-à-vis de *Peu*.

— De plus, le seuil sur le degré de focus, qui ne s'applique qu'aux quantificateurs relatifs, n'est pas utilisé pour le quantificateur absolu *Moins de 3*. Ainsi, dans le cadre d'un jeu de données important, il est toujours possible de trouver un faible nombre d'éléments vérifiant la propriété testée, comme illustré par les phrases 5 à 7 ou 18 à 21, privant donc ces phrases non spécifiques de leur portée informative.

— Un nombre important de phrases pourraient être éliminées en définissant les antonymes des modalités des VL et des quantificateurs. Par exemple, les phrases 2 à 4 sont des conséquences directes de la phrase 1, puisque si la plupart des livres ayant peu de pages sont mauvais, alors peu de livres ayant peu de pages sont moyens, bons ou excellents. La même remarque s'applique aux phrases 8 à 11.
Ainsi, les principes de cohérence décrits au chapitre 3 améliorent l'interprétabilité du résumé en évitant les contradictions logiques mais également en permettant de ne pas générer toutes les phrases possibles puisque certaines découlent naturellement d'autres lorsque la cohérence est respectée.

— Les phrases 15 à 17 ne concernent que les livres mauvais. Il manque donc celles traitant des livres moyens, bons et excellents. De même, les seules phrases concernant les livres ayant un nombre de pages élevé utilisent le quantificateur *Moins de 3* et non les quantificateurs relatifs. Les degrés de focus et/ou les valeurs de vérité de ces phrases sont-ils trop faibles ? Pourquoi ?

Annexe C

Étude sur les cardinalités

Cette annexe présente l'étude approfondie que nous avons réalisée sur le cardinal utilisé dans l'évaluation du degré de vérité des protoformes flous

Une partie de ces travaux ont été publiés dans (Bouchon-Meunier & Moyse, 2012).

C.1 Cardinalités utilisées dans le calcul des valeurs de vérité

Le calcul de la valeur de vérité présenté dans la section 1.2 p. 10 fait intervenir une cardinalité et une t-norme. Dans la définition initiale des résumés linguistiques flous (Yager, 1982), la cardinalité $\sigma Count$ (Deluca & Termini, 1972) est utilisée. C'est la cardinalité la plus courante de mesure de la taille d'un sous-ensemble flou (sef). Elle consiste en la somme des degrés d'appartenance des éléments du sef. Toutefois, d'autres définitions détaillées ci-dessous peuvent être utilisées : la cardinalité peut être représentée sous forme d'un scalaire dans le cas d'une cardinalité *crisp* ou d'un sef dans le cas d'une cardinalité *floue*. Nous proposons dans les parties suivantes l'utilisation de ces différents types de cardinalités afin d'appréhender leur effet sur l'évaluation du degré de vérité des protoformes flous.

C.1.1 Cardinalités crisp

Une cardinalité crisp est la mesure du nombre d'éléments présents dans un sef sous forme d'un scalaire, entier ou réel.

nCard

Étant donné un ensemble de points $\{x_j, j = 1...n\}$ et A un sef de fonction d'appartenance μ_A, en notant μ_i^A les $\mu_A(x_i)$ ordonnés de manière décroissante avec $\mu_0^A = 1$ et $\mu_{n+1}^A = 0$ et $j = \max\{s \in \{1, ..., n\} \mid \mu_{s-1}^A + \mu_s^A > 1\}$, la cardinalité $nCard$ (Ralescu, 1995) est définie comme :

$$nCard(A) = \begin{cases} 0 & \text{si } A = \emptyset \\ j & \text{si } \mu_j^A \geq 0,5 \\ j-1 & \text{si } \mu_j^A < 0,5 \end{cases} \tag{C.1}$$

La notation μ_i est utilisée pour μ_i^A lorsque le contexte est suffisamment précis.

Nous avons montré (preuve omise ici) que cette cardinalité peut être calculée comme :

$$nCard(A) = \begin{cases} z+1 & \text{si } \mu_{z+1} = 0,5 \\ z & \text{sinon} \end{cases} \tag{C.2}$$

avec $z = |A_{0,5^+}|$ et $A_{0,5^+}$ est l'α-coupe stricte de A : $A_{0,5^+} = \{x/\mu_A(x) > 0,5\}$. z mesure donc, à une unité près, le nombre d'éléments de A dont le degré d'appartenance est strictement supérieur à 0,5. Wygralak (1997) propose également des cardinalités entières basées sur $|A_{0,5}|$.

Comparaison entre σCount et nCard

La cardinalité $\sigma Count$ est plus intuitive et plus simple à calculer que $nCard$. Cependant, elle renvoie un nombre réel comme estimation du nombre d'éléments d'un sef, ce qui peut paraître surprenant pour une cardinalité. En effet, la sémantique de « il y a 3,6 personnes grandes » n'est pas simple à appréhender. L'intérêt de la cardinalité $nCard$ est donc sa capacité à estimer la taille d'un sef sous forme d'un nombre entier.

De plus, σ-count souffre d'un effet de bord dans le cas d'éléments nombreux appartenant à un degré faible à un sef. Par exemple, une cardinalité de 10 pour un sef composé de 1000 éléments avec un degré d'appartenance de 0,01 peut sembler contre-intuitive.

$nCard$ évite cet écueil en ne comptant que les éléments ayant un degré d'appartenance supérieur à 0,5. Ce mode d'évaluation à cependant l'inconvénient d'entraîner des discontinuités dans un certain nombre de cas. Plus précisément, il est possible de trouver deux sef dont la différence de taille mesurée par $\sigma Count$ est aussi petite que l'on veut mais dont la différence de taille mesurée par $nCard$ est égale au nombre d'éléments n du sef :

$$\begin{array}{lllll} \forall \epsilon \in]0;0,5], \exists A, B & \text{tq} & \sigma Count(A) - \sigma Count(B) & = & \epsilon \\ & \text{et} & nCard(A) - nCard(B) & = & n \end{array} \tag{C.3}$$

Le manque de robustesse de $nCard$ par rapport $\sigma Count$ plaide en faveur de l'utilisation de ce dernier dans le cadre des résumés linguistiques flous.

C.1.2 Cardinalités floues

Une cardinalité floue est un nombre flou défini sur l'univers $\{1, ..., n\}$ représentant la cardinalité du sef considéré. Cette cardinalité n'est pas directement utilisable avec les quantificateurs flous qui, en tant que sef eux-mêmes, associent à une valeur un degré d'appartenance.

Avec une cardinalité floue, la quantification est réalisée par comparaison entre le sef du quantificateur et celui de la cardinalité. Différentes approches en ce sens sont présentées dans la section 2.3.2 p. 34. Nous en proposons une nouvelle ici basée sur la similarité de Jaccard.

Cardinalité Z

Zadeh (1979) propose de définir le degré d'appartenance de k au cardinal flou d'un sef A comme $Z(A,k) = \sup\{\alpha \mid |A_\alpha| = k\}$ pour chaque valeur de $k = 0...|supp\,(A)\,|$. Cette expression provient directement de la définition de la cardinalité crisp transposée au flou par le principe d'extension et peut être interprétée comme la possibilité que le sef A ait k éléments. Cependant, cette cardinalité est non convexe : il peut exister des valeurs de k telles que $\nexists\alpha$ tel que $|A_\alpha| = k$, auquel cas la valeur 0 est employée, créant des « trous » dans la cardinalité floue.

FECount

Pour pallier cet inconvénient, Zadeh (1983) introduit :

$$FGCount\,(A,k) = \sup\{\alpha \mid |A_\alpha| \geq k\} \text{ et } FLCount\,(A,k) = \sup\{\alpha \mid |A_\alpha| \leq k\}$$

qui représentent respectivement la possibilité que A contiennent au moins k éléments et la possibilité que A contiennent au plus k éléments. Sur cette base, $FECount$ est définie comme la possibilité que A contienne exactement k éléments, i.e. qu'il en contienne au plus k *et* au moins k, soit, en utilisant le min comme opérateur de conjonction :

$$FECount\,(A,k) = \min\left(FLCount\,(A,k)\,, FGCount\,(A,k)\right) \tag{C.4}$$

$FECount(A,k)$ est convexe contrairement à $Z(A,k)$.

fCard

Wygralak (1986); Ralescu (1995); Casasnovas & Torrens (2003) proposent :

$$fCard\,(A,k) = \min\left(\mu_k, 1-\mu_{k+1}\right) \tag{C.5}$$

où les μ_k sont les degrés d'appartenance du sef A triés par ordre décroissant comme dans l'éq. (C.1) p. 215.

Toutefois, cette cardinalité floue est équivalente à la précédente : Dubois & Prade (1985b) montrent que $FGCount(A,k) = \mu_k$ et $FLCount(A,k) = 1-\mu_{k+1}$. Ainsi, $fCard(A,k) = FECount(A,k)$ par les éq. (C.4) et éq. (C.5).

Exploitation des cardinalités floues pour le degré de vérité des résumés linguistiques flous

La quantification de la cardinalité floue est réalisée en comparant les sef de la cardinalité et du quantificateur. Cette comparaison peut être réalisée de différentes manières, comme détaillé dans la section 2.3.2 p. 34.

Nous avons proposé de réaliser cette comparaison par le calcul d'une similarité de

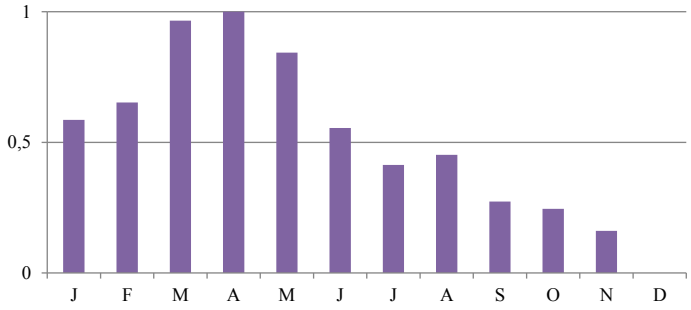


FIGURE C.1 – Scores des classements par mois des livres de régime sur 10 ans

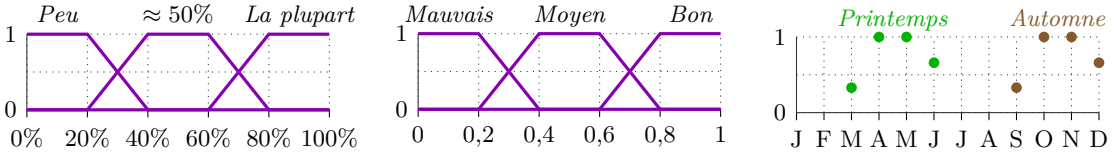


FIGURE C.2 – Quantificateurs et variables linguistiques *Score* et *Calendrier* pour l'étude des classements de livres

Jaccard entre le quantificateur Q et la cardinalité floue du sef P, conduisant à définir :

$$t\left(Qx \text{ sont } P\right) = \frac{\sum_{i=1\ldots n} \min\left(\mu_Q\left(x_i\right), card\left(P\right)\right)}{\sum_{i=1\ldots n} \max\left(\mu_Q\left(x_i\right), card\left(P\right)\right)} \tag{C.6}$$

Il est à noter que ce calcul de la valeur de vérité ne s'applique qu'aux protoformes « Qx sont P » et non « QRx sont P ».

C.1.3 Expériences

Données

Nous avons réalisé des expériences avec les données réelles collectées sur le site Amazon décrites dans l'annexe B p. 211. Nous avons calculé un score compris entre 0 et 1 sur les livres de régime alimentaire en fonction de leurs positions dans chacun des classements, puis nous avons sommé ces scores par mois depuis l'année 2002 jusqu'à l'année 2011 afin d'évaluer la saisonnalité de leurs ventes. La figure C.1 illustre ces données, et montre clairement que les pics de vente de livres de régimes sont au printemps.

Protocole

Les résumés linguistiques générés sont basés sur les quantificateurs et les modalités illustrés sur la figure C.2.

Afin de comparer les différentes cardinalités, la valeur de la vérité phrase « Moins de la moitié des classements de vente pour les livres de régimes sont bons », instanciée à partir du protoforme « Qx sont P », est calculée avec les cardinalités $\sigma Count$, $nCard$ et $fCard$.

La valeur de vérité de « Environ la moitié des bons classements de ventes pour les livres de régime apparaissent au printemps », instanciée à partir du protoforme « QRx sont P »,

TABLEAU C.1 – Résultats expérimentaux

Valeurs de vérité de « Qx sont P »

Q	P	$\sigma Count$	$nCard$	$fCard$
Peu	Bon	0,24	0,32	0,11
	Mauvais	1,00	1,00	0,14
≈ la moitié	Bon	0,76	0,68	0,18
	Mauvais	0,00	0,00	0,01
La plupart	Bon	0,00	0,00	0,00
	Mauvais	0,00	0,00	0,00

Valeurs de vérité de « QRx sont P »

Q	R	P	$\sigma Count$	$nCard$
Peu	Bon	Aut.	1,00	1,.00
		Pr.	0,00	0,00
	Mauvais	Aut.	0,00	0,00
		Pr.	1,00	1,00
≈ la moitié	Bon	Aut.	0,00	0,00
		Pr.	0,00	0,00
	Mauvais	Aut.	0,62	0,00
		Pr.	0,00	0,00
La plupart	Bon	Aut.	0,00	0,00
		Pr.	0,00	0,00
	Mauvais	Aut.	0,38	1,00
		Pr.	0,00	0,00

est calculée avec les cardinalités $\sigma Count$ et $nCard$.

Résultats

Les résultats des expériences sont donnés dans le tableau C.1. Les valeurs de vérité obtenues sur les protoformes du type « Qx sont P » sont indiquées sur le tableau de gauche, celles obtenues sur les protoformes du type « QRx sont P » sont reportées sur le tableau de droite.

Cardinalité crisp Sur les données considérées, on observe que les cardinalités crisp $\sigma Count$ et $nCard$ renvoient des valeurs de vérités similaires pour les protoformes du type « Qx sont P ». Pour les protoformes du type « QRx sont P » en revanche, les résultats sont assez différents et soulignent les interprétations différentes à prêter à $\sigma Count$ et $nCard$, ce dernier ne comptant que les éléments dont le degré d'appartenance est supérieur à 0,5 (cf. éq. (C.2) p. 216).

D'une manière générale, $\sigma Count$ est préférable à $nCard$ du fait du manque de robustesse de ce dernier. En fait, $nCard$ est utile lorsqu'un nombre entier est requis pour décrire la cardinalité d'un sef, comme par exemple dans un protoforme très simple « A contient $nCard(A)$ éléments », mais pas dans les protoformes classiques.

Cardinalité floue Concernant le calcul des « Qx sont P » avec la cardinalité floue $fCard$, le mode de calcul basé sur la similarité entre le quantificateur et la cardinalité floue ne renvoie que des valeurs très faibles, entre 0 et 0,18. Ces valeurs faibles s'expliquent par le fait que les nombres flous renvoyés par $fCard$ ne sont presque jamais normaux (i.e. au moins une de leurs valeurs a un degré d'appartenance égal à 1) et que leur similarité par rapport à un quantificateur normal est toujours faible. Un exemple du calcul de cette

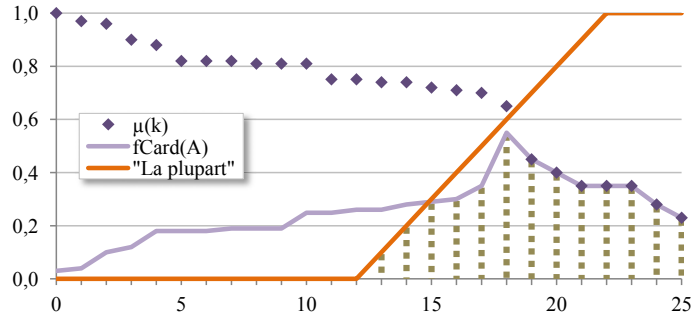


FIGURE C.3 – Similarité (petits carrés gris) entre une cardinalité floue et un quantificateur

similarité est représenté par la zone remplie par les carrés gris sur la figure C.3, résultat de la comparaison de Jaccard entre $fCard$ et le quantificateur *La plupart*.

Nous avons mené par la suite un travail non détaillé ici concernant les propriétés de $fCard$ visant à déterminer quel sef renverrait une cardinalité qui aurait un degré de similarité important avec un quantificateur. Les résultats de ce travail ont montré que $fCard$ n'est normale que si le sef considéré est crisp et que si le le sef est flou, alors une seule valeur de $fCard$ est supérieure à 0,5, sans atteindre 1.

Annexe D

Borne supérieure pour CV à partir de d

Cette annexe présente un résultat utilisé pour déterminer la mesure de régularité ρ utilisée dans DPE et détaillée dans la section 5.3.2 p. 106.

La démonstration présentée ici repose sur le fait que la définition de ρ implique que CV soit compris dans $[0, 1]$. Les notations utilisées sont celles de l éq. (5.13) p. 106 (indépendamment de τ) pour μ la moyenne du jeu de données, d sa déviation absolue moyenne, CV son coefficient de variation et de l'annexe H p. 237 pour L l'ensemble des points dont les valeurs sont inférieures ou égales à la moyenne, l la cardinalité de L et U le complémentaire de L dans X, donc l'ensemble des points dont les valeurs sont strictement supérieures à la moyenne. De plus, on note :

$$\theta = \sum_{i=1}^{l} x_i \text{ et } \theta' = \sum_{i=l+1}^{n} x_i$$

Nous avons donc proposé dans un premier temps de déterminer η la borne supérieure de CV pour un nombre n de données dans $[0, 1]$, l'idée étant d'utiliser η pour normaliser CV.

Pour ce faire, nous transformons l'expression de CV :

$$\begin{aligned}
CV &= \frac{d}{\mu} \\
&= \frac{\sum_{i=1}^{n} |x_i - \mu|}{\sum_{i=1}^{n} x_i} \\
&= \frac{\sum_{i=1}^{l} (\mu - x_i) + \sum_{i=l+1}^{n} (x_i - \mu)}{\theta + \theta'} \\
&= \frac{l\mu - \theta + \theta' - (n-l)\,\mu}{\theta + \theta'} \\
&= \frac{2l\mu - \theta + \theta' - \theta - \theta'}{\theta + \theta'} \text{ car } n\mu = \theta + \theta' \\
&= 2\frac{l\mu - \theta}{\theta + \theta'} \\
&= 2\left(\frac{l}{n} - \frac{\theta}{\theta + \theta'}\right)
\end{aligned}$$

De plus, en augmentant la valeur de θ et en diminuant d'autant celle de θ', la valeur de μ reste constante et celle de d augmente, donc pour l et μ donnés, la plus grande valeur de CV est atteinte pour la plus petite valeur possible de θ et la plus grande de θ'. Dans le meilleur des cas, $\theta = 0$ si les l valeurs sont égales à 0 et $\theta' = n - l$ si les $n - l$ valeurs sont égales à 1.

Supposons désormais que les jeux de données considérés soient composés de l valeurs 0 et de $n - l$ valeurs 1. On a alors $\theta = 0$, $\theta' = n - l$ et donc :

$$CV = \frac{2l}{n}$$

En ce cas, la plus grande valeur de l, à savoir $l = n - 1$, maximise CV, qui est alors égal à $2(n-1)/n$ et qui tend donc vers 2 quand n est grand.

Annexe E

Généralisation du score d'érosion pour des données dans $\mathbb{R}$

Cette annexe présente un ensemble d'expressions plus générales pour le score d'érosion afin de permettre son calcul pour des données dans $\mathbb{R}$ et plus dans $[0, 1]$, tel que présenté dans la section 6.3 p. 127. La définition de γ_{es} (cf. éq. (5.9) p. 102) pour le regroupement par score d'érosion est également mise à jour.

x_i et $\widehat{x}_i$ Nous considérons que les $\widehat{x}_i$ reçus sont dans $\mathbb{R}$ mais que les valeurs considérées à un instant t sont dans $[m, M]$. A l'aide de ces notations, $x_i \in [0, 1]$ est défini ainsi :

$$x_i = \frac{\widehat{x}_i - m}{\Delta} \tag{E.1}$$

avec $\Delta = M - m$.

es_i et $\widehat{es}_i$ Définissions $\widehat{es}_i$ le score d'érosion calculé avec les $\hat{x}_i$:

$$\widehat{es}_i = \sum_{j=0}^{\widehat{z}_i} \widehat{\epsilon}_i^j = \sum_{j=0}^{\widehat{z}_i} \min\left(\hat{\epsilon}_{i-1}^{j-1}, \hat{\epsilon}_i^{j-1}, \hat{\epsilon}_{i+1}^{j-1}\right)$$

Notons d'abord que $z_i = \hat{z}_i$ car :

$$\begin{aligned}
\widehat{z}_i &= \arg\min_{j>0} \widehat{\epsilon}_i^j = m \\
&= \arg\min_{j>0} \{\min(\widehat{x}_{i-j}, ..., \widehat{x}_{i+j}) = m\} \\
&= \arg\min_{j>0} \{\Delta \min(x_{i-j}, ..., x_{i+j}) + m = m\} \\
&= \arg\min_{j>0} \{\min(x_{i-j}, ..., x_{i+j}) = 0\} \\
&= z_i \text{ (Eq. 5.7)}
\end{aligned}$$

es_i et $\widehat{es}_i$ sont alors liés par la relation suivante :

$$\begin{aligned} es_i &= \sum_{j=0}^{z_i} \epsilon_i^j = \sum_{j=0}^{z_i} \min\left(x_{i-j}, ..., x_{i+j}\right) \text{ par l'Eq. (6.1)} \\ &= \sum_{j=0}^{z_i} \min\left(\frac{\widehat{x}_{i-j} - m}{\Delta}, ..., \frac{\widehat{x}_{i+j} - m}{\Delta}\right) \\ &= \frac{\sum_{j=0}^{z_i} \min\left(\widehat{x}_{i-j}, ..., \widehat{x}_{i+j}\right) - m \times z_i}{\Delta} = \frac{\sum_{j=0}^{z_i} \widehat{\epsilon}_i^j - m \times z_i}{\Delta} \\ &= \frac{\widehat{es}_i - m \times z_i}{\Delta} \end{aligned}$$

es_i et $\overline{es}_i$ Concernant le lien entre es_i et $\overline{es}_i$, nous définissons dans un premier temps l'opérateur de dilatation δ, opération adjointe de l'érosion ϵ (cf. éq. (5.4) p. 101) :

$$\delta_i^j = \max\left(\delta_{i-1}^{j-1}, \delta_i^{j-1}, \delta_{i+1}^{j-1}\right) \qquad \delta_i^0 = \epsilon_i^0 = x_i \tag{E.2}$$

et le score de dilatation non normalisé :

$$ds_i = \sum_{j=0}^{zd_i} \delta_i^j \tag{E.3}$$

avec :

$$zd_i = \arg\min_{j>0} \delta_i^j = 1 \tag{E.4}$$

La relation $\overline{z}_i = zd_i$ est immédiate :

$$\begin{aligned} \overline{z}_i &= \arg\min_{j>0} \overline{\epsilon}_i^j = 0 \\ &= \arg\min_{j>0} \left\{\min\left(1 - x_{i-j}, ..., 1 - x_{i+j}\right) = 0\right\} \\ &= \arg\min_{j>0} \left\{\max\left(x_{i-j}, ..., x_{i+j}\right) = 1\right\} \\ &= zd_i \end{aligned}$$

Ainsi, la relation entre $\overline{es}_i$ et ds_i s'exprime :

$$\begin{aligned}
\overline{es}_i &= \sum_{j=0}^{\overline{z}_i} \overline{\epsilon}_i^j = \sum_{j=0}^{\overline{z}_i} \min\left(\overline{x}_{i-j}, ..., \overline{x}_{i+j}\right) \\
&= \sum_{j=0}^{\overline{z}_i} \min\left(1 - x_{i-j}, ..., 1 - x_{i+j}\right) \\
&= \sum_{j=0}^{\overline{z}_i} \min\left(\frac{\Delta - \widehat{x}_{i-j} + m}{\Delta}, ..., \frac{\Delta - \widehat{x}_{i+j} + m}{\Delta}\right) \\
&= \frac{M \times \overline{z}_i - \sum_{j=0}^{\overline{z}_i} \max\left(\widehat{x}_{i-j}, ..., \widehat{x}_{i+j}\right)}{\Delta} \\
&= \frac{M \times \overline{z}_i - \hat{ds}_i}{\Delta}
\end{aligned}$$

Ces écritures permettent d'évaluer la fonction γ_{es} à partir des valeurs dans $\mathbb{R}$ du jeu de données initial. En effet, la condition $es_i^* > \overline{es}_i^*$ (cf. éq. (5.9) p. 102) s'écrit alors :

$$\begin{aligned}
es_i^* > \overline{es}_i^* &\Leftrightarrow \frac{\widehat{es}_i - m \times z_i}{\Delta \times M_{es}} > \frac{M \times \overline{z}_i - \hat{ds}_i}{\Delta \times \overline{M}_{es}} \\
&\Leftrightarrow \overline{M}_{es}\left(\widehat{es}_i - m \times z_i\right) > M_{es}\left(M \times \overline{z}_i - \widehat{ds}_i\right) \qquad \text{(E.5)}
\end{aligned}$$

Annexe F

Théorèmes liés aux calculs incrémentaux

Cette annexe présente les démonstrations des théorèmes de calcul par niveaux, incrémental et incrémental par niveaux du score d'érosion, référencés dans la section 6.1 p. 113.

F.1 Mise à jour des érosions $\epsilon_i^j(t+1)$

Rappel du théorème 2 p. 120 *Mise à jour des érosions successives à l'arrivée de* x_{n+1}
En notant $q = \underset{k=1...n}{\arg\max}\, x_k \leq x_{n+1}$ *et* $m = (n+1+q)/2$, *on a :*

$$\epsilon_i^j(t+1) = \begin{cases} \epsilon_i^j(t) & \text{si } i \leq m \quad \text{(F.1)} \\ \epsilon_i^j(t) & \text{si } i > m \text{ et } j < n+1-i \quad \text{(F.2)} \\ x_{n+1} & \text{si } i > m \text{ et } n+1-i \leq j < i-q \quad \text{(F.3)} \\ \epsilon_q^{j-(i-q)} & \text{si } i > m \text{ et } j \geq i-q \quad \text{(F.4)} \end{cases}$$

Démonstration. La démonstration repose sur le calcul de l'érosion de chaque *donnée* à chaque *étape*, i.e. le calcul de ϵ_i^j pour tous les i et tous les j.

Il faut remarquer dans un premier temps que si i et j sont tels que x_{n+1} n'est pas impliqué dans le calcul de $\epsilon_i^j(t+1)$, alors $\epsilon_i^j(t) = \epsilon_i^j(t+1)$. L'implication de x_{n+1} dans le calcul de ϵ_i^j est simple à vérifier en utilisant l'écriture non récursive du score d'érosion (cf. éq. (6.1) p. 115) qui permet d'écrire :

$$\forall i \forall j \quad (n+1) \notin \{i-j, ..., i+j\} \Rightarrow \epsilon_i^j(t+1) = \epsilon_i^j(t) \quad \text{(F.5)}$$

Si x_{n+1} est impliqué dans le calcul de $\epsilon_i^j(t+1)$, on distingue deux cas selon que x_q l'est aussi ou non.

Cas $0 \leq i \leq m$

Si j est tel que $n+1 \notin \{i-j, ..., i+j\}$, alors $\epsilon_i^j(t+1) = \epsilon_i^j(t)$ selon l'éq. (F.5).

Si j est tel que $(n+1) \in \{i-j, ..., i+j\}$, alors $q \in \{i-j, ..., i+j\}$. En effet, $q \leq n+1$ par définition donc $q \leq i+j$ et :

$$\begin{aligned} m &= (n+1+q)/2 \\ \text{donc } 2m - q &= n+1 \\ \text{comme } 2m - q \leq i+j &\text{ puisque } i+j \geq n+1 \\ &\text{on a } q \geq 2m-(i+j) \geq 2i-(i+j) \text{ car } i \leq m \\ &\text{d'où } q \geq i-j \end{aligned}$$

On a donc $\epsilon_i^j(t+1) = \min(x_{i-j}, ..., x_q, ...x_{n+1}, ..., x_{i+j}) \leq x_q < x_{n+1}$ donc x_{n+1} n'est pas impliqué dans le calcul de $\epsilon_i^j(t+1)$, d'où $\epsilon_i^j(t+1) = \epsilon_i^j(t)$, ce qui prouve l'éq. (F.1).

Cas $i > m$

On distingue trois cas selon la valeur de j (cf. éq. (F.2) à (F.4)) :

Cas 1 : $j < n+1-i \Rightarrow n+1 \notin \{i-j, ..., i+j\}$ ce qui montre l'éq. (F.2) par l'éq. (F.5).

Cas 2 : $n+1-i \leq j < i-q$, donc $q < i-j$ et $i+j \geq n+1$, donc x_{n+1} est impliqué dans le calcul de $\epsilon_i^j(t+1)$ mais pas x_q. Donc $\epsilon_i^j(t+1) = x_{n+1}$ car $x_{n+1} = \min(x_{q+1}, ..., x_{n+1})$ par définition de q, ce qui prouve l'éq. (F.3).

Cas 3 : $j \geq i-q \Leftrightarrow i-j \leq q$ donc :

$$\begin{aligned} j \geq i-q &\Leftrightarrow i+j \geq 2i-q \\ &\Rightarrow i+j \geq 2m-q \text{ car } i > m \\ &\Leftrightarrow i+j \geq n+1 \text{ par déf. de } m \end{aligned}$$

Ainsi :

$$\begin{aligned} x_i^j(t+1) &= \min\left(x_{i-j}(t+1), ..., x_q(t+1), \underbrace{..., x_{n+1}(t+1)}_{\geq x_q \text{ déf. de } q}, \underbrace{..., x_{i+j}(t+1)}_{=+\infty \text{ car éq. (6.9)}}\right) \\ &= \min(x_{i-j}(t), ..., x_q(t)) \end{aligned}$$

En notant $\beta = j-(i-q)$, la condition $j \geq i-q$ devient $\beta \geq 0$ et :

$$\begin{aligned} \min(x_{i-j}(t), ..., x_q(t)) &= \min(x_{q-\beta}(t), ..., x_q(t)) \\ &= \min(x_{q-\beta}(t), ..., x_q(t), ..., x_{q+\beta}(t)) \text{ car éq. (6.9) et par déf. de } q \\ &= x_q^\beta(t) \text{ par l'éq. (6.1)} \\ &= x_q^{j-(i-q)}(t) \text{ qui prouve l'éq. } (F.4) \end{aligned}$$

□

F.2 Mise à jour des indices des valeurs clés $\lambda_{il}(t+1)$

Rappel du théorème 4 p. 122 *Mise à jour incrémentale de λ_{il}*

En notant $q = \arg\max_{k=1...n} x_k \leq x_{n+1}$, $m = (n+1+q)/2$, et k_i tel que $d_{i,k_i-1}(t) < n+1-i \leq d_{ik_i}(t)$ avec $k_{n+1}=0$, on a :

$$\forall i, \forall l,\ \lambda_{il}(t+1) = \begin{cases} \lambda_{il}(t) & \text{si } i \leq m \quad \text{(F.6)} \\ \lambda_{il}(t) & \text{si } i > m \text{ et } l < k_i \quad \text{(F.7)} \\ n+1 & \text{si } i > m \text{ et } l = k_i \quad \text{(F.8)} \\ \lambda_{q,l-k_i-1}(t) & \text{si } i > m \text{ et } l > k_i \quad \text{(F.9)} \end{cases}$$

$$\forall i,\ \omega_i(t+1) = \begin{cases} \omega_i(t) & \text{si } i \leq m \quad \text{(F.10)} \\ k_i + \omega_q(t) & \text{si } i > m \quad \text{(F.11)} \end{cases}$$

Démonstration. Les étapes de cette preuve sont similaires à celles de la démonstration de la mise à jour des érosions en incrémental : plusieurs cas sont identifiés, selon que les x_i considérés sont avant ou après m et dans ce deuxième cas, selon que l'étape d'érosion est avant ou après k_i.

Cas $i \leq m$ pour λ_{il} et ω_i

Comme vu dans la démonstration du théorème 2 p. 120, pour tout $i \leq m$, les érosions en $t+1$ sont identiques à celles en t. Les érosions clés en particulier sont inchangées, donc $\lambda_{il}(t) = \lambda_{il}(t+1)$ et $\omega_i(t) = \omega_i(t+1)$, ce qui prouve les éq. (F.6) et (F.10).

Cas $i > m$ et $l < k_i$ pour λ_{il}

Toujours selon le théorème 2 p. 120, les érosions sont également inchangées si $i > m$ et $j < n+1-i$. Or par définition de k_i, $d_{i,k_i-1}(t) < n+1-i$ donc pour $l < k_i$ $d_{il}(t) < n+1-i$ donc $\epsilon_i^{d_{il}(t)}(t+1) = \epsilon_i^{d_{il}(t)}(t)$ soit $\chi_{il}(t+1) = \chi_{il}(t)$.

Ce qui est valable pour χ_{il} l'est aussi pour λ_{il} puisque ce dernier est l'indice de la valeur χ_{il}. Donc $\lambda_{il}(t+1) = \lambda_{il}(t)$, ce qui prouve l'éq. (F.7).

Cas $i > m$ et $l = k_i$ pour λ_{il}

Par l'éq. (F.3) p. 227 nous savons que $\epsilon_i^j(t+1) = x_{n+1}$ si $n+1-i \leq j < i-q$. De plus, $\epsilon_i^{n+1-i} < \epsilon_i^{n-i}$ par l'éq. (6.2) p. 116, donc $n+1-i \in D_i$ par définition (cf. éq. (6.3) p. 116).

Comme les valeurs précédentes de D_i ont pour indices $1...k_i-1$, $n+1-i$ est ajouté en $k_i^{ème}$ position. Or la $n+1-i^{ème}$ érosion est égale x_{n+1} donc $\lambda_{il}(t+1) = n+1$ si $l = k_i$ ce qui prouve l'éq. (F.8)

Cas $i > m$ et $l > k_i$ pour λ_{il}

La prochaine érosion différente de x_{n+1} est x_q par définition de q. De plus, l'éq. (F.4) p. 227 montre que les érosions suivantes sont les érosions de x_q. Donc les érosions clés suivantes sont les érosions clés de x_q. Donc les éléments de D_i d'indice supérieur à k_i sont les éléments de D_q, ce qui montre l'éq. (F.9).

Cas $i > m$ pour ω_i

L'éq. (F.11) est démontrée à partir des points ci-dessus, puisque le nombre d'éléments de D_i après mise à jour est égal à k_i auquel s'ajoutent les éléments de D_q au nombre

de ω_q. □

F.3 Mise à jour du score d'érosion incrémental par niveaux

Rappel du théorème 5 p. 123 *Mise à jour incrémentale par niveaux de es_i*

En notant $q = \underset{k=1\ldots n}{\arg\max}\, x_k \leq x_{n+1}$, $m = (n+1+q)/2$, et k_i tel que $d_{i,k_i-1}(t) < n+1-i \leq d_{ik_i}(t)$ avec $k_{n+1}=0$ et p_i défini pour $i > m$ tel que $\lambda_{ip_i}(t) = q$, on a :

$$\forall i,\, es_i(t+1) = \begin{cases} es_i(t) & \text{si } i \leq m \\ \chi_{i,k_i-1}(t)(n+1-i-d_{ik_i}(t)) + 2x_{n+1}(i-m) & \text{si } m < i < n+1 \\ \quad - \sum_{j=k_i}^{p_i-1} \chi_{ij}(t)(d_{i,j+1}(t) - d_{ij}(t)) + es_i(t) & \\ 2x_{n+1}(n+1-m) + es_q(t) & \text{si } i = n+1 \end{cases}$$

Démonstration. La mise à jour du du score d'érosion par niveaux repose sur le fait que les valeurs clés avant k_i et après p_i sont les mêmes en t et en $t+1$. Donc les seules valeurs clés impactant la mise à jour de es_i sont situées entre k_i et p_i.

Cas $i \leq m$ et $i = n+1$: la preuve pour ces deux cas découle directement du théorème 3 p. 121 de mise à jour du score d'érosion en incrémental.

Cas $m < i < n+1$: tout d'abord, montrons l'existence de p_i. Comme $\lambda_{ip_i}(t) = q$, cela signifie que la $p_i^{ème}$ érosion clé de $x_i(t)$ égale x_q, qui existe comme montré dans le théorème 3 p. 121.

De plus, le théorème 4 p. 122 montre que les dernières valeurs de $\lambda_{il}(t)$ pour l entre p_i et ω_i sont les $\lambda_{qr}(t)$ pour r entre 0 et ω_q d'où :

$$\sum_{l=p_i}^{\omega_{i-1}} \chi_{il}(t)\,(d_{i,l+1}(t) - d_{il}(t)) = \sum_{l=0}^{\omega_{q-1}} \chi_{ql}(t)\,(d_{q,l+1}(t) - d_{ql}(t)) = es_q(t)$$

Avec l'expression du théorème 1 p. 117, le score d'érosion en t s'exprime alors :

$$\begin{aligned} es_i(t) &= \sum_{l=0}^{k_i-2} \chi_{il}(t)\,(d_{i,l+1}(t) - d_{il}(t)) + \chi_{i,k_i-1}(t)\,(d_{ik_i}(t) - d_{i,k_i-1}(t)) \\ &\quad + \sum_{l=k_i}^{p_i-1} \chi_{il}(t)\,(d_{i,l+1}(t) - d_{il}(t)) + \underbrace{\sum_{l=p_i}^{\omega_{i-1}} \chi_{il}(t)\,(d_{i,l+1}(t) - d_{il}(t))}_{=es_q(t)} \\ &= \sum_{l=0}^{k_i-2} \chi_{il}(t)\,(d_{i,l+1}(t) - d_{il}(t)) + \chi_{i,k_i-1}(t)\,(d_{ik_i}(t) - d_{i,k_i-1}(t)) \\ &\quad + \sum_{l=k_i}^{p_i-1} \chi_{il}(t)\,(d_{i,l+1}(t) - d_{il}(t)) + es_q(t) \end{aligned} \tag{F.12}$$

De manière similaire en $t+1$, le score d'érosion se décompose :

$$\begin{aligned}
es_i(t+1) = &\sum_{l=0}^{k_i-2} \underbrace{\chi_{il}(t+1)\left(d_{i,l+1}(t+1) - d_{il}(t+1)\right)}_{=\chi_{il}(t)\left(d_{i,l+1}(t)-d_{il}(t)\right) \text{ car } l<k_i} \\
&+ \underbrace{\chi_{i,k_i-1}(t+1)}_{=\chi_{i,k_i-1}(t) \text{ car } <k_i} \Big(\underbrace{d_{i,k_i}(t+1)}_{=n+1-i} - \underbrace{d_{i,k_i-1}(t+1)}_{d_{i,k_i-1}(t) \text{ car } <k_i}\Big) \\
&+ \underbrace{\chi_{ik_i}(t+1)}_{=x_{n+1}} \underbrace{\Big(\underbrace{d_{i,k_i+1}(t+1)}_{=i-q} - \underbrace{d_{ik_i}(t+1)}_{=n+1-i}\Big)}_{2(i-m)} + \underbrace{es_q(t+1)}_{=es_q(t)} \\
= &\sum_{l=0}^{k_i-2} \chi_{il}(t)\left(d_{i,l+1}(t) - d_{il}(t)\right) + 2x_{n+1}(i-m) + es_q(t) \qquad \text{(F.13)} \\
&+ \chi_{i,k_i-1}(t+1)\left(n+1-i-d_{i,k_i-1}(t)\right)
\end{aligned}$$

La preuve se termine en calculant la différence entre les éq. (F.13) et (F.12) :

$$\begin{aligned}
es_i(t+1) - es_i(t) = &\sum_{l=0}^{k_i-2} \chi_{il}(t)\left(d_{i,l+1}(t) - d_{il}(t)\right) + 2x_{n+1}(i-m) + es_q(t) \\
&+ \chi_{i,k_i-1}(t+1)\left(n+1-i-d_{i,k_i-1}(t)\right) \\
&- \sum_{l=0}^{k_i-2} \chi_{il}(t)\left(d_{i,l+1}(t) - d_{il}(t)\right) + \chi_{i,k_i-1}(t)\left(d_{ik_i}(t) - d_{i,k_i-1}(t)\right) \\
&- \sum_{l=k_i}^{p_i-1} \chi_{il}(t)\left(d_{i,l+1}(t) - d_{il}(t)\right) - es_q(t) \\
= &\chi_{i,k_i-1}(t)\left(d_{ik_i}(t) - d_{i,k_i-1}(t)\right) \\
&- \sum_{l=k_i}^{p_i-1} \chi_{il}(t)\left(d_{i,l+1}(t) - d_{il}(t)\right) + 2x_{n+1}(i-m)
\end{aligned}$$

□

Annexe G

Expressions moyennes pour les calculs de complexité

Cette annexe est références dans la section 6.2.4 p. 126 concernant la complexité des méthodes incrémentales de calcul du score d'érosion par niveaux.

G.1 Distance moyenne au premier point inférieur à gauche

Note : la démonstration donnée ci-dessous est issue d'une discussion sur le forum Math Stack Exchange[1].

En notant d la distance du premier point inférieur à gauche du dernier point reçu, on a $d = n + 1 - q$ avec les notations du théorème 3 p. 121. En considérant d comme une variable aléatoire dans $\{1, ..., w\}$, nous proposons de déterminer $E[d]$, sa valeur moyenne.

Par définition :

$$\mathbb{E}[d] = \sum_{k=1}^{w} kP(d = k)$$

En supposant que les x_i reçus sont i.i.d. de distribution f_X et de fonction de répartition F_X, en notant x_n le dernier point reçu, la probabilité que $d = k$ s'écrit :

1. http://math.stackexchange.com/questions/900828/number-of-groups-containing-at-least-1-and-at-most-k-elements

$$
\begin{aligned}
P(d=k) &= P(x_n < x_{n-1} \cap ... \cap x_n < x_{n-k+1} \cap x_n \geq x_{n-k}) \\
&= \int_{x_n<x_{n-1},...,x_n<x_{n-k+1},x_n\geq x_{n-k}} f_X(x_{n-k}, x_{n-k+1}, ..., x_n)\, dx_{n-k}...dx_n \\
&= \underbrace{\int_{-\infty}^{\infty} F'_X(x_n)\, dx_n}_{P(x_n)} \underbrace{\int_{-\infty}^{x_n} f_X(x_{n-k})\, dx_{n-k}}_{P(x_n\geq x_{n-k})} \underbrace{\int_{x_n}^{+\infty} f_X(x_{n-1})\, dx_{n-1}}_{P(x_n<x_{n-1})} ... \\
&\quad ... \underbrace{\int_{x_n}^{+\infty} f_X(x_{n-k+1})\, dx_{n-k+1}}_{P(x_n<x_{n-k+1})}
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} F'_X(x_n) F_X(x_n) (1 - F_X(x_n))^{k-1} dx_n \\
&= \int_{-\infty}^{\infty} F'_X(x_n) (1 - (1 - F_X(x_n))) (1 - F_X(x_n))^{k-1} dx_n \\
&= \int_{-\infty}^{\infty} F'_X(x_n) (1 - F_X(x_n))^{k-1} - \int_{-\infty}^{\infty} F'_X(x_n) (1 - F_X(x_n))^{k} dx_n \\
&= \frac{1}{k} \int_{-\infty}^{\infty} \left(F_X^k(x_n)\right)' dx_n - \frac{1}{k+1} \int_{-\infty}^{\infty} \left(F_X^{k+1}(x_n)\right)' dx_n \\
&= \frac{1}{k} - \frac{1}{k+1} = \frac{1}{k(k+1)}
\end{aligned}
$$

La valeur moyenne de d s'écrit donc :

$$
\begin{aligned}
\mathbb{E}[d] &= \sum_{k=1}^{w} \frac{k}{k(k+1)} \\
&= H_{w+1}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[d] &= \sum_{k=1}^{w} k \frac{1}{k(k+1)} \\
&= H_{w+1} \\
&\sim \ln(w+1)
\end{aligned}
$$

où $H_n = \sum_{i=1}^{n} 1/n$.

G.2 Probabilité de l'apparition d'un nouvel extrema

Toujours en supposant que les données sont i.i.d., sans supposition sur leur distribution, la probabilité qu'une nouvelle valeur x_w soit supérieure à tous les points x_1 à x_{w-1} contenus

dans la fenêtre est :

$$\begin{aligned}
P\left(\bigcap_{i=1}^{w-1} x_w > x_{w-i}\right) &= \int_{-\infty}^{+\infty} F'_X\left(x_w\right) dx_w \times \prod_{i=1}^{w-1} P\left(x_w > x_{w-i}\right) \\
&= \int_{-\infty}^{+\infty} F'_X\left(x_w\right) dx_w \times \prod_{i=1}^{w-1} \int_{-\infty}^{x_w} f_X(x_{w-i}) dx_{w-i} \\
&= \int_{-\infty}^{+\infty} F'_X\left(x_w\right) F_X^{w-1}\left(x_w\right) dx_w \\
&= \frac{1}{w} \int_{-\infty}^{+\infty} \left(F^w\left(x_w\right)\right)' dx_w \\
&= \frac{1}{w}
\end{aligned}$$

Par symétrie, cette probabilité est également celle que x_w soit inférieure à toutes les valeurs précédentes.

Annexe H

Détermination de $P\left(d = \delta\right)$

Cette annexe présente la démonstration de la distribution de la statistique utilisée pour la méthode LDPE et présentée dans la section 8.1.2 p. 161.

En notant d la variable aléatoire qui mesure la déviation absolue moyenne des tailles de g groupes contenant n points, nous établissons dans cette annexe l'expression analytique de $P\left(d = \delta\right)$, la probabilité que d soit égale à δ étant donnés n et g. Cette probabilité est définie uniquement si $n \geq g$ et $g \geq 1$.

La preuve est réalisée en deux temps. D'abord, une expression générale de d est établie, permettant son calcul par décomposition des groupes en deux, avec d'une part ceux dont la taille est plus petite ou égale à la moyenne, et d'autre part ceux dont la taille est strictement supérieure à la moyenne.

Ensuite, à l'aide de cette décomposition, la probabilité est calculée comme le rapport entre le nombre de combinaisons des n points dans les g groupes de déviation δ divisé par le nombre total de combinaisons de n points dans g groupes.

H.1 Expression alternative de d

Dans cette sous-section, une expression alternative de d est donnée, basée sur la répartition des tailles de groupes autour de la moyenne.

Comme chaque point appartient à un groupe et un seul, la taille moyenne des groupes μ est donnée par :

$$\mu = \frac{1}{g}\sum_{j=1}^{g} s_j = \frac{n}{g} \tag{H.1}$$

Pour des raisons pratiques, les variables auxiliaires suivantes sont définies : $\mu^- = \lfloor \mu \rfloor$ et $\mu^+ = \lceil \mu \rceil$.

En notant L l'ensemble des indices de groupes dont la taille est inférieure ou égale à μ, U l'ensemble des indices de groupes dont la taille est strictement supérieure à μ, l la cardinalité de L et u la cardinalité de U, alors $L \cup U = \{1, ..., g\}$, $l + u = g$ et d

(cf. éq. (5.13) p. 106) s'écrit :

$$d = \frac{1}{g}\left(\sum_{j\in L}(\mu - s_j) + \sum_{j\in U}(s_j - \mu)\right) = \frac{1}{g}\left(l\mu - u\mu + \sum_{j\in U} s_j - \sum_{j\in L} s_j\right) \tag{H.2}$$

De plus, la somme des tailles de tous les groupes est égale à n donc $\sum_{j\in U} s_j = n - \sum_{j\in L} s_j$, et en notant $\theta = \sum_{j\in L} s_j$, l'éq. (H.2) devient :

$$d = \frac{1}{g}\Big(l\mu - (g-l)\mu + n - \theta - \theta\Big) = \frac{2}{g}(l\mu - \theta) \tag{H.3}$$

De cette expression découle directement $g^2 d = 2(nl - g\theta)$, donc $g^2 d \in \mathbb{N}$ car n, l et θ sont entiers. Donc $g^2\delta \notin \mathbb{N} \Rightarrow P(d = \delta | n, g) = 0$. Par la suite, nous supposons que $g^2\delta \in \mathbb{N}$.

H.2 Combinaisons telles que $d = \delta$

La deuxième partie de cette preuve consiste en l'étude d'un problème de combinatoire dont l'objectif est de déterminer le nombre de répartitions possibles de n points dans g groupes telles que la déviation absolue moyenne des tailles de groupes soit égale à δ.

L'éq. (H.3) implique que cette déviation peut être calculée uniquement avec la connaissance de la taille des groupes dont les indices sont dans L. Ainsi, nous donnons dans la section H.2.1 l'expression analytique des tailles possibles pour L, ou de manière équivalente des valeurs de l, pour un δ donné.

Ensuite, pour chaque l possible, nous donnons dans la section H.2.2 le nombre correspondant de répartitions des points dans les groupes. Ce problème est en fait celui du calcul du nombre de compositions de θ points dans l groupes sous certaines contraintes

Enfin, nous donnons dans la section H.2.3 p. 240 l'expression renvoyant le nombre de répartitions telles que la déviation absolue moyenne des groupes soit δ sachant les l possibles.

H.2.1 Valeurs possibles de l sachant δ

Les valeurs possibles de l sachant δ sont déterminées grâce aux contraintes sur l et θ.

D'abord, $1 \leq l < g$. Le cas particulier $l = g$ ne se produit que si tous les groupes ont μ^- éléments, donc lorsque $\mu^- = \mu$, i.e. $n \bmod g = 0$.

De plus, les groupes dont l'indice est dans L sont tes que $\forall j \in L$, $1 \leq s_j \leq \mu^-$, donc par définition de θ, $l \leq \theta \leq l\mu^-$. De manière symétrique, comme $u = g - l$ alors $\forall j \in U$, $\mu^+ \leq s_j$ et donc $\theta \leq n - u\mu^+$. Il en découle que $\theta \leq \min\left(l\mu^-, n - u\mu^+\right)$.

Par ailleurs, θ est entier car défini comme la somme du nombre entier de points dans les groupes de L. Ainsi, les valeurs possibles pour l sont telles que $\theta \in \mathbb{N}$.

A l'aide de ces contraintes, nous définissons Λ l'ensemble des valeurs l possibles étant

donné δ :

$$\Lambda = \left\{1 \leq l < g \text{ tel que } l \leq \theta \leq \min\left(l\mu^-, n - u\mu^+\right) \wedge \theta \in \mathbb{N}\right\} \cup \{g \text{ si } n \bmod g = 0\} \tag{H.4}$$

H.2.2 Nombre de compositions de n points dans g groupes de tailles comprises dans $[a, b]$

Comme les points d'une série temporelle sont ordonnés, le nombre de répartitions de n points dans g groupes de sorte que chaque groupe contienne au moins un point est égal au nombre de compositions de n dans g groupes. Ce nombre est égal au nombre de possibilités qu'il y a de placer $g - 1$ barres dans une liste de n points, i.e. de sélectionner $g - 1$ emplacements parmi $n - 1$ points (le dernier n'étant pas possible car il entraînerait la création d'un groupe vide) : c'est l'argument *stars and bars* proposé par Feller (1967, p.38).

Dans les paragraphes suivants, nous proposons un point de vue général sur la question du nombre de manière de placer n points dans g groupes de sorte que chaque groupe contienne au moins a et au plus b points, puis de manière plus spécifique au moins 1 et au plus b points, et enfin au moins 1 point uniquement sans contrainte sur le nombre maximal de points. Pour ce dernier cas, nous utilisons l'argument *stars and bars* présenté plus haut.

Cas général Le nombre de compositions de n dans g groupes de sorte que chaque groupe contienne au moins a et au plus b points est noté $N(n, g, a, b)$.

Si $bg < n < ag$ ou $b < a$ alors $N(n, g, a, b) = 0$. De plus, si $n = g$ et $a = 1 < b$, alors $N(n, g, a, b) = 1$.

Soustraire $a - 1$ aux g groupes ne change pas le nombre de compositions $N(n, g, a, b)$, et ce dernier peut donc être calculé comme le nombre de compositions de $n - g(a - 1)$ points en g groupes contenant au moins 1 et au plus $b - a + 1$ points, soit :

$$N(n, g, a, b) = N(n - g(a - 1), g, 1, b - a + 1) \tag{H.5}$$

Cas $a = 1$ Maintenant que $N(n, g, a, b)$ est exprimé en fonction de $N(n, g, 1, b)$, il convient de déterminer cette dernière expression. Comme détaillé dans notre discussion sur le site Math StackExchange [1], ce nombre peut être calculé récursivement en notant qu'il est égal à la somme des cas mutuellement exclusifs suivants :

— si aucun groupe ne contient plus de $b - 1$ éléments alors il est égal à $N(n, g, 1, b - 1)$

— si un groupe exactement contient b éléments alors il est égal à $\binom{g}{1} \times N(n - b, g - 1, 1, b - 1)$, soit le nombre de possibilité de choisir un groupe contenant b éléments parmi g groupes, multiplié par le nombre de compositions de $n - b$ points parmi les $g - 1$ groupes restant comportant moins de b points

1. `http://math.stackexchange.com/questions/900828/number-of-groups-containing-at-least-1-and-at-most-k-elements`

— si deux groupes exactement contiennent b éléments alors il est égal à $\binom{g}{2} \times N(n - 2b, g - 2, 1, b - 1)$ pour des raisons similaires

— plus généralement si k groupes exactement contiennent b éléments alors il est égal à $\binom{g}{k} \times N(n - kb, g - k, 1, b - 1)$

De plus, comme il ne peut y avoir plus de groupes « remplis » i.e. contenant b points qu'il n'y a de points à répartir, $k \leq \lfloor n/b \rfloor$, d'où la formule récursive suivante :

$$N(n, g, 1, b) = \sum_{k=0}^{\lfloor n/b \rfloor} \binom{g}{k} \times N(n - kb, g - k, 1, b - 1) \tag{H.6}$$

Notons que le dernier argument $b - 1$ garantit la convergence de la somme car $b < a \Rightarrow N(n, g, a, b) = 0$.

Cas $a = 1$ et $b = +\infty$ En ce cas, aucune contrainte sur le nombre maximal de points par groupe n'est donnée, d'où :

$$N(n, g, 1, +\infty) = \binom{n-1}{g-1} \tag{H.7}$$

Cette formule vient directement de l'argument *stars and bars* présenté plus haut.

H.2.3 Nombre de compositions sachant l et δ

Étant donnés δ, n, g et un l donné, le nombre $\widetilde{N}$ de compositions de n points en g groupes de taille telle que leur déviation absolue moyenne soit δ est égal au nombre de manière de créer les ensembles L et U avec leurs contraintes données dans la section H.2.1 p. 238. Sachant par l'éq. (H.3) p. 238 que θ est égal à $nl/g + g\delta/2$, $\widetilde{N}$ est le produit du nombre de compositions de θ points dans l groupes ayant au moins 1 et au plus μ^- points par le nombre de compositions de $n - \theta$ en u groupes ayant au moins μ^+et au plus $n - \theta - (u - 1)\mu^+$ points, le tout multiplié par le nombre de possibilité de choisir l éléments parmi g groupes, d'où, par les éq. (H.5) et éq. (H.6) :

$$\widetilde{N}(n, g, l, \delta) = \binom{g}{l} \times N(\theta, l, 1, \mu^-) \times N\left(n - \theta - \mu^-, u, 1, n - \theta - u\mu^+ + 1\right) \tag{H.8}$$

H.3 Probabilité de $d = \delta$

Étant donnés n et g, la probabilité $P(d = \delta)$ est finalement calculée comme le rapport entre le nombre de compositions de n en g groupes dont la déviation absolue moyenne en taille est égale à δ par le nombre total de compositions de n en g groupes, d'où, par

les éq. (H.8) et éq. (H.7) :

$$P\left(d=\delta\right)=\begin{cases}0 & \text{si } n<g \text{ ou } g<1 \text{ ou } g^2\delta\notin\mathbb{N}\\ \dfrac{\sum_{l\in\Lambda}\widetilde{N}\left(n,g,l,\delta\right)}{N\left(n,g,1,+\infty\right)} & \text{sinon}\end{cases} \tag{H.9}$$

Annexe I

Détail des résultats des expériences LDPE

Les tableaux présentés dans cette annexe contiennent les résultats détaillés obtenus dans les expériences décrites dans la section 8.5 p. 170.

Les tableaux I.1 et I.2 p.244 présentent les résultats associés au premier critère de qualité concernant le nombre de zones identifiées zE (cf. éq. (8.14) p. 171), tandis que les tableaux I.3 et I.4 p.245 détaillent les résultats obtenus sur le second critère de qualité lié aux nombre de groupes correctement identifiés comme appartenant à une zone périodique pC. Par ailleurs, le formalisme utilisé dans ces tableaux est identique à celui présenté dans le tableau 8.1 p. 172, à l'exception des méthodes utilisant une moyenne pondérée notées $wAvg$ ici.

L'interprétation de ces résultats est donnée dans la section 8.5.3 p. 174.

TABLEAU I.1 – Pourcentage d'apparition des valeurs de paramètres parmi les 30 meilleurs (i.e. plus faibles) résultats de zE sur l'ensemble des scénarios, configurations, répétitions, pour chaque méthode et pour chaque paramètre

zE	m1		m2		m3		fm1		fm2		fm3	
%	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg
Alpha												
1%	81%	97%	27%	7%	74%	95%	18%	33%	0%	0%	10%	8%
5%	19%	3%	54%	73%	26%	5%	19%	21%	13%	22%	20%	21%
10%	0%	0%	19%	18%	0%	0%	41%	28%	39%	35%	45%	45%
15%	0%	0%	0%	2%	0%	0%	22%	17%	48%	43%	25%	26%
Pi min												
0,2	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0,4	0%	0%	0%	0%	0%	0%	3%	5%	0%	13%	0%	0%
0,6	14%	31%	27%	10%	11%	18%	15%	26%	40%	58%	10%	8%
0,8	86%	69%	73%	90%	89%	82%	82%	69%	60%	29%	90%	92%
minSize												
2							21%	11%	16%	41%	30%	32%
4							31%	22%	16%	18%	30%	30%
6							25%	30%	39%	20%	23%	21%
8							22%	37%	29%	21%	17%	16%
minSep												
2							18%	37%	44%	28%	14%	15%
4							15%	37%	41%	12%	14%	11%
6							48%	26%	13%	23%	41%	42%
8							19%	0%	2%	37%	31%	32%

TABLEAU I.2 – Valeur moyenne de zE, pour chaque méthode et pour chaque scénario

zE	m1		m2		m3		fm1		fm2		fm3	
	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg
S1	354%	206%	206%	137%	406%	325%	46%	23%	33%	23%	66%	48%
S2	140%	105%	89%	83%	177%	155%	17%	13%	18%	18%	24%	22%
S3	97%	98%	71%	72%	107%	96%	20%	20%	25%	25%	21%	21%
S4	49%	46%	26%	25%	65%	62%	33%	33%	33%	33%	32%	33%
S5	69%	69%	54%	54%	69%	69%	49%	46%	28%	28%	49%	49%
S6	86%	151%	108%	0%	39%	0%	3%	16%	19%	0%	0%	0%
μ	133%	113%	92%	62%	144%	118%	28%	25%	26%	21%	32%	29%
σ	113%	58%	62%	48%	137%	113%	18%	13%	7%	11%	23%	19%
Rank	11	9	8	7	12	10	4	2	3	1	6	5

TABLEAU I.3 – Pourcentage d'apparition des valeurs de paramètres parmi les 30 meilleurs résultats de pC sur l'ensemble des scénarios, configurations, répétitions, pour chaque méthode et pour chaque paramètre

pC	m1		m2		m3		fm1		fm2		fm3	
%	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg
Alpha												
1%	0%	0%	0%	8%	0%	0%	0%	1%	0%	0%	0%	0%
5%	46%	55%	8%	22%	41%	69%	30%	35%	27%	14%	15%	3%
10%	53%	39%	56%	47%	56%	31%	38%	38%	39%	36%	45%	51%
15%	0%	6%	36%	23%	3%	0%	32%	25%	34%	51%	40%	46%
Pi min												
0.2	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0.4	0%	0%	0%	8%	0%	0%	0%	0%	14%	4%	0%	0%
0.6	0%	0%	100%	92%	0%	0%	8%	28%	62%	94%	0%	0%
0.8	100%	100%	0%	0%	100%	100%	92%	72%	25%	2%	100%	100%
minSize												
2							1%	0%	1%	0%	1%	21%
4							17%	11%	16%	17%	23%	31%
6							30%	32%	36%	47%	25%	24%
8							52%	57%	47%	36%	51%	24%
minSep												
2							48%	57%	87%	50%	36%	27%
4							38%	43%	13%	17%	29%	31%
6							15%	0%	0%	15%	28%	30%
8							0%	0%	0%	18%	7%	12%

TABLEAU I.4 – Valeur moyenne de pC, pour chaque méthode et pour chaque scénario

pC	m1		m2		m3		fm1		fm2		fm3	
	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg	Avg	wAvg
S1	89%	94%	92%	93%	82%	87%	92%	96%	94%	94%	86%	90%
S2	91%	93%	90%	90%	85%	86%	93%	94%	91%	90%	88%	88%
S3	86%	85%	83%	82%	81%	81%	88%	86%	84%	83%	83%	82%
S4	94%	94%	90%	90%	90%	90%	94%	94%	89%	89%	91%	91%
S5	80%	84%	91%	91%	76%	76%	86%	90%	95%	95%	81%	81%
S6	78%	63%	59%	90%	86%	96%	80%	55%	53%	97%	92%	100%
μ	86%	86%	84%	89%	83%	86%	89%	86%	84%	91%	87%	89%
σ	6%	12%	13%	4%	5%	7%	5%	16%	16%	5%	4%	7%
Rank	6	9	11	2	12	7	3	8	10	1	5	4