Analyse et visualisation pour l'étude de la qualité des séries temporelles de données imparfaites

Zied Ben Othmane

Thèse présentée pour obtenir le grade de docteur

Université de Reims Champagne-Ardenne

❧

CReSTIC – Centre de Recherches en STIC
École doctorale Sciences Numériques et de l'Ingénieur (ED SNI)

Discipline : Informatique

Analyse et visualisation pour l'étude de la qualité des séries temporelles de données imparfaites

---

Par : **Zied Ben Othmane**

Membres du jury:

Rapporteure : Lydia Boudjeloud-Assala, Maîtresse de conférences HDR, Université de Lorraine, Metz

Rapporteur : Arnaud Martin, Professeur des Universités, Université de Rennes 1, Lannion

Examinateur : Allel Hadjali, Professeur des Universités, ENSMA, Poitiers

Examinateur : Hacène Fouchal, Professeur des Universités, Université de Reims Champagne-Ardenne

Encadrant : Amine Aït Younes, Maître de conférences, Université de Reims Champagne-Ardenne

Directeur : Cyril de Runz, Maître de conférences HDR, Université de Tours

Date de soutenance : 23 janvier 2020

Remerciements

Je souhaite remercier en premier lieu mon directeur de thèse, M. Cyril de Runz, Maître de Conférences HDR au laboratoire CReSTIC dans l'équipe MODECO jusqu'en septembre 2019 et désormais membre du LIFAT de l'Université de Tours. Je lui suis reconnaissant pour le temps conséquent qu'il m'a accordé, pour ses qualités pédagogiques et scientifiques, pour sa franchise et pour sa sympathie. J'ai beaucoup appris à ses côtés et je lui adresse ma gratitude pour tout cela. Son énergie et sa confiance ont été des éléments moteurs pour moi. J'ai pris un grand plaisir à travailler avec lui.

J'adresse de chaleureux remerciements à mon co-encadrant de thèse, M. Amine Aït Younes, Maître de Conférences au CReSTIC de l'Université de Reims Champagne-Ardenne, pour son attention de tout instant sur mes travaux, pour ses conseils et réflexions avisés pour la bonne réussite de cette thèse.

Un grand merci à M. Damien Bodénès, chef de projets chez Kantar jusqu'en 2018, pour son implication dans ce projet de thèse. Il m'a beaucoup appris. J'ai apprécié son enthousiasme et sa sympathie. Je tiens à remercier Vincent Mercelot qui a pris la suite de M. Bodénès dans l'encadrement de mon travail au sein de Kantar.

Je remercie M. Arnaud Martin, Professeur des Universités à l'Université de Rennes 1, et Mme Lydia Boudjeloud-Assala, Maîtresse de Conférences à l'Université de Lorraine, d'avoir accepté d'être rapporteurs de ce mémoire. Ce mémoire a bénéficié de leur lecture très attentive et de leurs remarques précieuses.

J'associe à ces remerciements messieurs Allel Hadjali, Professeur des Universités à l'ISAE - ENSMA, et Hacène Fouchal, Professeur des Universités à l'Université de Reims Champagne-Ardenne, pour avoir accepté d'examiner mon travail et pour l'intérêt qu'ils y ont porté.

Je remercie la société Kantar pour m'avoir permis d'effectuer ce travail de thèse dans de très bonnes conditions. Je remercie aussi l'ANRT car sans son soutien financier, la thèse n'aurait pu se faire.

J'exprime toute ma reconnaissance à Zahia Guessoum et Herman Akdag pour les conseils et retours faits sur mon travail, notamment au sein du comité de suivi de thèse. Je souhaite aussi remercier Frédéric Blanchard pour la relecture de mon travail.

Je voudrais exprimer ma reconnaissance envers mes collègues de Kantar qui m'ont apporté leur soutien moral et intellectuel tout au long de ma démarche. Merci à Jean Marc Duhen pour ses conseils durant les 3 ans passés dans l'entreprise.

J'adresse mes plus sincères remerciements à ma famille : Mes parents Zine el Abidine Ben Othmane, Ibtissem Jemaiel, mes frères Iheb et Souheil et tous mes proches et amis, qui m'ont accompagné, aidé, soutenu et encouragé tout au long de ma thèse.

Finalement, je souhaite particulièrement remercier ma femme Mme Ben Othmane Mariem pour son précieux soutien et de m'avoir supporté durant ces trois années.

Table des matières

Table des figures

Liste des tableaux

Première partie

Introduction générale

Chapitre 1

Contexte et problématique

L'intelligence stratégique est un processus de gestion qui vise à identifier les informations pertinentes pour en construire une source d'inspiration permettant d'établir une stratégie spécifique. Elle a pour objectif la détermination des connaissances actionnables dans un environnement compétitif. Elle recouvre les différentes activités permettant à un décideur de lever les incertitudes auxquelles il est confronté et de donner divers éléments de décision. Cette démarche multidisciplinaire s'appuie essentiellement sur la veille informationnelle.

La veille informationnelle est, d'après la norme XP X 50-53[AFN98], « une activité continue en grande partie itérative visant à une surveillance active de l'environnement technologique, commercial, etc., pour en anticiper les évolutions ». Selon [Cac04], dans le Dictionnaire de l'information, elle est définie comme un processus continu et dynamique faisant l'objet d'une mise à disposition personnalisée et périodique de données ou d'informations, traitées selon une finalité propre au destinataire, faisant appel à une expertise en rapport avec le sujet ou la nature de l'information collectée. C'est donc un processus de détection des signaux, favorables ou défavorables pour un sujet particulier. Elle se fonde sur le renseignement et l'enquête pour construire de l'information pertinente qui a un sens par rapport à un projet. Ainsi, au sein d'une société, la veille est une discipline importante permettant de bien connaître son positionnement vis-à-vis diverses dimensions : veille technologique, veille économique, veille concurrentielle, etc.

La veille concurrentielle représente l'activité continue et itérative qui vise à une surveillance active des mouvements des entreprises concurrentes. Elle porte aussi sur les produits et services, les relations externes, les relations clients, etc. Elle consiste généralement à surveiller plusieurs canaux de données afin d'en tirer des informations utiles à l'entreprise qui les exploitera pour prévenir les menaces en provenance de la concurrence et de saisir des opportunités qui la rendront plus efficiente.

Dans le cadre de la communication médiatique, la veille concurrentielle vise à étudier le comportement d'un concurrent via l'analyse de ses activités sur les différents médias (TV, Radio, Presse, Internet, etc.). Pour une société, connaître les stratégies de communication misent en oeuvre par ses concurrents et le montant de leurs investissements est riche d'information. Ceci lui permet de définir ses stratégies de développement vis à vis de la concurrence.

La stratégie de communication médiatique d'une société dans sa volonté de promouvoir un produit (marque) consiste au minimum à diffuser sur des supports médias les publicités voulues. Dans les faits, une entreprise cherche à obtenir une communication médiatique efficace en terme de satisfaction finale de son objectif. C'est-à-dire, elle cherchera les différents supports de communication qui lui semblent les plus pertinents pour avoir le meilleur retour sur investissement.

La société Kantar, dans laquelle s'est effectuée cette thèse, a pour objectif de proposer des indicateurs de veille concurrentielle permettant de décrire l'activité de toutes les sociétés diffusant des publicités sur les différents médias. Dans le cadre de cette thèse, nous étudions des données des publicités diffusées sur les supports numériques et principalement sur certains sites web préalablement identifiés par la société.

1.1 Contexte industriel

Kantar media, filiale du groupe WPP [1], est un cabinet d'étude et un fournisseur de données sur les stratégies de communications des entreprises en France et dans le monde. Kantar media présente plusieurs produits et services à ces clients. Son activité est essentiellement dédiée au conseil, aux études de marché et au marketing.

En France, les activités de l'entreprise visent d'une part à la veille réputationnelle permettant d'apporter des informations sur ce que le monde peut dire autour d'une société et de son activité sur les différents médias (télévision, radio, presse, cinéma, affichage et internet). D'autre part, elle travaille sur la veille concurrentielle et la détection des messages passés via les publicités.

Au sein de l'entreprise, une activité importante est la détection de la diffusion des messages publicitaires sur les différents supports, leur catégorisation et leur valorisation afin de déterminer les stratégies des annonceurs et leurs comportements sur les différents médias. Il peut ainsi s'agir de la détection et de la catégorisation des stratégies de placement de publicités des différentes entreprises, de leur ciblage des clients, de la catégorisation des supports, etc.

Dans ce travail de thèse, nous nous sommes intéressés aux activités liées aux supports médias numériques, qui peuvent être, par exemple, les sites web, les réseaux sociaux, les moteurs de recherches, etc.

Au sein de ce contexte numérique, la distribution et l'affichage des publicités diffèrent d'un espace à un autre, d'un support média à un autre mais aussi d'un utilisateur à un autre.

1. WPP est le plus important réseau d'agences de publicité et de communication mondial : 179 000 employés à travers 111 pays

Afin d'étudier les placements publicitaires sur les supports numériques, Kantar déploie sur le web des robots récoltant des données de log sur les publicités affichées dans les bannières présentes dans les différents site utilisés. La société achète aussi des données issues de panel utilisateurs mais aussi des données d'audience. L'ensemble de ces données sur l'affichage des publicités sur des sites web sont exploitées dans cette thèse dans le cadre d'un projet interne nommé MMS (Mesures Multi-Sources). Les données récoltées par les robots sont des logs présentant à minima une étiquette temporelle, la page web sur laquelle la publicité a été affichée, et des informations sur la publicité affichées (nom du produit, de la marque). Par ailleurs, les données d'audience nous donnent, par période temporelle, l'audience des différentes pages web. Les données de panel utilisateurs permettent de compléter les données précédentes.

L'objectif de MMS est de modéliser le marché publicitaire sur internet, et ce, à partir de la récolte des données publicitaires faites par des sources de données internes et extrapolées avec des données de trafics fournies par des sources externes. A l'aide de ces données et après un redressement statistique, la société souhaite fournir des estimations, les plus fiables possibles, des investissements publicitaires faits par un ou plusieurs annonceurs sur le web. L'investissement publicitaire est le montant dépensé par un acteur (entreprise, annonceur, etc.) pour placer des bannières publicitaires sur le web afin de promouvoir ses produits ou services. Ceci permet aux clients de Kantar de prendre conscience de leur positionnement sur le marché.

Pour pouvoir fournir les bons chiffres d'investissements publicitaires, l'outil se fonde sur le nombre de vues et les volumes de publicités en cherchant à refléter la réalité du marché. Aussi, il est important d'être le plus exhaustif possible afin d'avoir le meilleur prototype pour que les résultats finaux soient proches de la réalité. Comme l'exhaustivité n'est pas possible, plusieurs sources (robots, panels, etc.) sont agrégées.

Cependant, et malgré cette agrégation, des tests comparatifs, entre les données estimées par le processus interne à la société et les données réelles provenant des régies, montrent des écarts importants aussi bien sur les volumes d'impressions que sur les investissements publicitaires. Les données acquises ne sont finalement que des indicateurs de tendances ou des estimations. Elles ne forment donc pas un reflet exhaustif de la réalité des placements publicitaires sur le web, contrairement à ce que l'on peut avoir sur d'autres types de supports médias (TV, radio, presse écrite).

L'idée de cette thèse est d'étudier la qualité des données récoltées et de son lien avec la véracité des résultats. Les attentes principales sont :

— de fournir des outils d'aide à l'analyse de la transformation de la donnée au niveau du processus du projet MMS ;
— de mettre en place des outils de visualisation et de fouille de données qui s'adaptent aux interactions avec les utilisateurs et permettent d'établir un modèle de la véracité des résultats produits.

1.2 Contexte scientifique

Cette thèse est le fruit de la collaboration entre la société Kantar et l'URCA (Université de Reims Champagne-Ardenne). À l'URCA, ce travail est effectué au sein de l'équipe MODECO du CReSTIC et s'inscrit dans l'axe fouille de données de l'équipe. Les travaux de l'équipe dans cet axe visent à une meilleure compréhension des données et étudient les questions liées à la véracité des traitements et des données.

Ce présent travail cherche à trouver des moyens et techniques d'estimation de la qualité des données récoltées tout en respectant leurs diverses spécificités. L'idée générale de cette thèse est de définir et de développer de nouvelles approches d'analyse de données et de visualisation qui devront permettre une meilleure compréhension de ces données volumineuses.

Du point de vue scientifique, les données étudiées sont assimilables à des flux de données imparfaites [2]. En effet, les données récoltées sont par nature incomplètes car nous ne pouvons pas obtenir des données exhaustives en terme d'affichage des publicités sur le web. Cette incomplétude de la récolte fait que les données acquises forment plus des estimations qu'une information précise sur la réalité. Elles sont donc également imprécises.

Les attentes scientifiques de cette thèse sont de proposer des solutions afin de mieux appréhender les données récoltées et d'en évaluer la qualité.

Pour cela, nous nous intéresserons particulièrement à :

— La représentation des données en considérant leurs imperfections et notamment leur imprécision. Ainsi, selon [Bou93 ; Sme98], nous modélisons nos données à l'aide d'ensembles ou d'ensembles flous.

— L'étude de la variabilité et de la stabilité des flux de données imparfaites récoltées.

Pour cela, il s'agira aussi de mettre en place des outils de visualisation et de fouille de données s'adaptant aux interactions avec les utilisateurs.

1.3 Problématique

Le projet de recherche, dans lequel s'inscrit cette thèse, a pour objectif de proposer des solutions afin, dans un premier temps, de mieux appréhender les données récoltées et leur qualité, et, dans un second temps, de travailler sur leur véracité en vue d'apporter des explications qui peuvent améliorer les résultats finaux.

Les deux questions industrielles principales auxquelles ce projet souhaite proposer des réponses sont :

Quels facteurs liés à la récolte des données peuvent influencer la précision des indicateurs construits pour l'investissement ? Et comment les superviser ?

2. L'imperfection de l'information, selon [Bou93 ; Bou95], peut venir de trois principales causes : l'incertitude, l'imprécision et l'incomplétude.

Cette problématique industrielle pose les deux questions scientifiques suivantes :

1. Comment étudier la qualité des données ?
2. Comment la qualité des données et de la récolte influence-t-elle la véracité des résultats en sortie ?

Dans ce mémoire, nous nous concentrons sur la première question.

Pour répondre à cette problématique, il s'agit dans un premier temps de travailler à la représentation des données en considérant leurs imperfections et à l'étude de la qualité en terme de volatilité des flux de données étudiés. Aussi les premières sous-questions posées sont :
Comment prendre en considération l'imperfection des données ?
Comment analyser la volatilité de la récolte des données ?

Afin d'y répondre, nous chercherons à représenter les données en considérant leur imprécision. Pour cela, nous exploiterons la notion de rang, d'une part via des représentations ensemblistes fondées sur la projection des données dans leur quantile, et, d'autre part, via des représentations floues. Nous chercherons ensuite à construire des indicateurs sur la variabilité et la stabilité des flux de données informant de fait sur leur volatilité et donc sur leur qualité.

Enfin, afin de permettre aux experts de Kantar de superviser la qualité de leurs récoltes, il s'agira de répondre à la question :
Quels outils de visualisation et d'interaction faut-il déployer pour permettre une exploration interactive de la qualité des flux de données ?

1.4 Résumé des contributions

Nos deux premières contributions visent à répondre à la question principale autour de l'étude de la qualité des données et de leurs récoltes. À l'aide de ces contributions, nous étudierons l'impact sur la véracité des données en sortie. La troisième contribution porte sur la proposition d'un outil de visualisation interactif qui repose sur des approches de visualisation développées et adaptées aux besoins des experts data pour l'exploration de la qualité (des récoltes) des données.

Dans ces contributions, nous considérons chaque donnée d'un flux de données particulier par rapport aux valeurs précédentes de ce dit flux. Cela correspond à la vision interne. Nous exploitons, aussi, une vision externe qui vise à positionner la donnée vis à vis de l'ensemble des données disponibles pour la même étiquette temporelle. Par ailleurs, nous excluons les valeurs manquantes de la construction de ces ensembles de références.

Dans notre première contribution, nous proposons de considérer l'imprécision des données d'entrée en les représentant par un ensemble de données. Nous déterminons cet ensemble à partir de leurs quantiles d'appartenance selon la vision interne et la vision

externe. Ainsi nous construisons pour chaque flux de données, deux nouveaux flux reposant sur les projections des données dans leur quantiles. Les données manquantes sont positionnées dans un quantile propre. Nous cherchons ensuite à répondre à la question de la volatilité par la construction d'un indicateur de variabilité de chaque flux de données construit. Cet indicateur est fondé sur les sauts dans les quantiles entre deux étiquettes temporelles successives. Nous agrégeons ensuite les valeurs des indicateurs interne et externe pour construire un indicateur d'instabilité du flux. Les communications suivantes sont liées à cette contribution [Ben+18b ; Ben+19b]

Dans notre deuxième contribution, présentée dans [Ben+19d ; Ben+19c], nous cherchons à généraliser l'approche précédente à un contexte où :

— Les données sont considérées dans un intervalle de confiance, que nous appelons imprecise-set, pour les représenter plus tôt dans le processus les imprécisions des données.
— La construction des ensembles de référence pour les données selon les deux visions (interne et externe) est fondée sur des regroupements, sans a priori sur les données, construits à partir de la densité. Le nombre de groupes est donc variable.

À l'aide de ces informations, nous définissons des ensembles flous sur les regroupements précédents, afin de proposer un nouvel indicateur de positionnement des données imprécises dans leur ensemble. Nous construisons ensuite la variabilité interne (resp. externe) en étudiant les sauts sur ces indicateurs de positions internes (resp. externes).

Notre troisième contribution est un outil de fouille visuelle interactive de la qualité des récoltes de données. Cet outil permet une exploration allant du plus général sur l'ensemble des données à du plus spécifique sur un sous ensemble de données voire sur un flux particulier pour une sous-période donnée. Les indicateurs fournis informent tant sur les absences/lacunes dans la récolte des données que sur la variabilité et l'instabilité. Il permet aussi d'exploiter la classification en méta-catégories et catégories des médias et publicités faites par la société. Cette contribution a fait l'objet des publications suivantes [Ben+18b ; Ben+18a ; Ben+19a].

1.5 Plan

Ce mémoire est organisé en quatre parties.

Le chapitre suivant (Chapitre 2) est dédié à la présentation des données exploitées dans la suite du manuscrit. Nous y introduisons aussi les premières notations liées à notre cadre applicatif.

La partie II présente l'état de l'art en trois chapitres. Le chapitre 3 introduit les notions de qualité et de véracité. Dans ce cadre, nous nous intéresserons plus particulièrement à la gestion des imperfections dans la partie 4, puis à la volatilité via les notions de variabilité et stabilité dans le chapitre 5.

La partie III introduit nos deux contributions pour l'analyse des flux de données imparfaites. Dans le chapitre 6, nous présentons notre première contribution proposant une approche par quantiles. Le chapitre suivant (chapitre 7) décrit notre deuxième contribution autour d'une approche floue généralisant notre première contribution.

La partie IV, constituée de deux chapitres, porte sur notre contribution concernant l'outil de visualisation interactive. Le chapitre 8 fournit une étude de la littérature dédiée à la visualisation de données et plus particulièrement de flux de données imparfaites. Le chapitre 9 présente notre outil et en illustre l'intérêt sur des scénarios d'utilisation.

Enfin, nous présentons, dans la dernière partie, une conclusion générale à ce travail et les perspectives que nous envisageons.

Chapitre 2

Données étudiées

2.1 Introduction

Dans ce chapitre, nous présentons les données que nous avons étudiées, manipulées et analysées pendant cette thèse. Nos données proviennent de différentes sources, sont agrégées et transformées afin de devenir nos données d'études. Ce chapitre présente les processus d'acquisition des données ainsi que les pré-traitements que nous avons effectués en amont des études présentées dans ce mémoire.

Nous commençons donc par présenter les données étudiées, leurs types, leurs natures et leurs caractéristiques. Ensuite, nous présentons le processus métier mis en œuvre dans l'entreprise afin de générer des informations exploitables pour notamment l'estimation des investissements faits sur les publicités. Ensuite, nous montrons que les données sont sujettes à différentes imperfections et que la considération de ces dernières peut induire des doutes sur la véracité des résultats fournis par l'entreprise. Ensuite, nous explicitons les processus de pré-traitement des données mis en place afin de pouvoir étudier la qualité des récoltes des données. Nous présentons les données qui en résultent et introduisons leur présentation plus formelle qui servira de support à nos contributions présentées dans la suite du manuscrit.

2.2 Présentation des données acquises par la société

La société Kantar, partenaire industriel finançant la présente thèse, est spécialisée dans l'étude des données publicitaires diffusées sur différents médias. Son objectif final est d'étudier les investissements publicitaires faits sur les publicités Web, i.e. le capital investi par les différents annonceurs pour un produit par exemple.

Pour cela, l'entreprise a mis en place un système de récolte de données web fondé sur des capteurs. Ces capteurs enregistrent les bannières vues sur différents sites web et toutes autres informations associées, comme par exemple le site où est parue la publicité, dans le but d'effectuer les analyses nécessaires pour pouvoir en déduire les investissements publicitaires de chaque annonceur.

Dans le présent travail, nous ne nous intéressons pas à l'étude de la déduction des estimations publicitaires mais plutôt à l'étude de la qualité des données récoltées par les capteurs afin de pouvoir agir sur la véracité des informations générées. En effet, la véracité des estimations financières dépend de la qualité des données brutes. C'est pourquoi, dans cette thèse, nous traitons la question de la qualité des flux de données issus de capteurs regroupés par période temporelle pour former des séries temporelles [1].

Les données de publicités peuvent être étudiées selon différents axes : axe orienté sur les publicités selon leurs supports de diffusion, axe produit où ce n'est plus la publicité en tant que telle mais le produit qu'elle représente, axe média où l'on observe les sites selon les publicités qu'ils diffusent, etc.

Pour chaque axe d'analyse, une vue particulière sur les données est nécessaire afin de produire les différentes variables d'étude pour le dit axe. Par exemple, pour l'axe média (ou sites web), nous étudions les données selon le nombre de vues (audience), le nombre de publicités affichées, le nombre d'URL visitées.

Afin de faciliter le travail sur les données de publicités – généralement des bannières (images), des vidéos, etc. – Kantar utilise un système de notation qui marque ces contenus par des identifiants et des tags (cf. tableau 2.1 à titre d'exemple) et les récoltes sous forme de logs. Les informations stockées sont composées d'un HASHURL représentant d'une manière cryptée l'URL sur laquelle la publicité a été vue, de l'identifiant IDMEDIA du média (e.g. un site web), du nombre de pages web vues lors de la visite (NBRPAGEVIEW), la date de l'enregistrement (DATE) de ces informations, de l'identifiant de la source (IDSOURCE), et l'identifiant de la visite (IDDATACRAWL).

Champs	HASHURL	IDMEDIA	NBPAGEVIEW	DATE	IDSOURCE	IDDATACRAWL
1	-7007601	10004	8474	12/01/2019	12	2226822
2	-700601	10025	6974	12/01/2019	12	2226784
3	-262925	10808	3074	12/01/2019	12	2226882

Table 2.1 – Exemple fictif de données brutes

Il est à noter que nos données sont principalement des informations sur la tendance car, comme indiqué précédemment dans ce manuscrit, nous ne pouvons prétendre à une récolte exhaustive.

De plus, les volumes de données sont amenés à augmenter naturellement au cours du temps car ce sont des flux de données. De plus, si le nombre de sites observés est actuellement limité, celui-ci peut être amené à augmenter.

1. Une série temporelle, ou série chronologique, est une suite de valeurs numériques représentant l'évolution d'une quantité spécifique.

Aussi, nous devons, dans nos traitements et analyses, tenir compte de cette possibilité d'évolution vers des Big Data [2]. L'idée dans ce travail est de structurer des millions de données dans des fichiers logs offrant une vision de traitement dite « à plat ». Les logs sont ensuite indexés, ce qui offre un requêtage rapide et puissant. Pour cela nous devons aussi opter pour des approches, outils et technologies scalables [3].

Pour le choix du système de gestion de nos données, nous nous sommes orientés vers des structures non relationnelles (NoSQL) orientées documents (un log est un document) qui ont l'avantage d'avoir une meilleure scalabilité que les systèmes relationnels tout en permettant un traitement efficace [Pok13 ; GRR14].

Par ailleurs, les prétraitements expliqués dans ce chapitre ont pour objectif de transformer, de manière scalable, les données acquises en des volumes de données restreints, compatibles avec leur analyse.

La section suivante a pour objectif d'expliquer le projet MMS, contexte applicatif de cette thèse, et, plus précisément, les procédés d'acquisition et de traitements mis en œuvre dans ce projet.

2.2.1 Le projet MMS

Nos travaux s'effectuent dans le cadre d'un projet qui se nomme MMS (Mesure Multi-Sources). Dans cette section, nous présentons son fonctionnement ainsi que toutes les étapes de génération des données.

Le projet a pour objectif d'estimer les investissements publicitaires sur le web en tenant compte du marché ciblé (France, Italie, etc.). Pour cela, il exploite un prototype fondé sur des étapes intermédiaires ayant pour finalité la production des chiffres qui serviront à l'étude du comportement du marché (voir figure 2.1).

L'information est récoltée à partir de trois sources différentes et complémentaires représentant les mêmes médias. Elles sont récoltées sous la forme de logs. Une quatrième source est utilisée pour les enrichir avec les informations d'audience. À partir de ces données, l'entreprise produit, à l'heure actuelle, des estimations des investissements publicitaires. Les estimations sont donc fortement dépendantes de la qualité des données sources. Par exemple, le nombre de vues d'une publicité et le volume des publicités détectées doivent refléter au maximum la réalité afin que le calcul des investissements soit le moins biaisé possible.

2. Les Big Data sont des ensembles de données devenus si volumineux qu'ils dépassent les capacités d'analyses des outils informatiques classiques de gestion de base de données ou de l'information

3. la capacité d'un produit à s'adapter à la montée en charge, en particulier sa capacité à maintenir ses fonctionnalités et ses performances en cas de forte demande et de volume de données très important

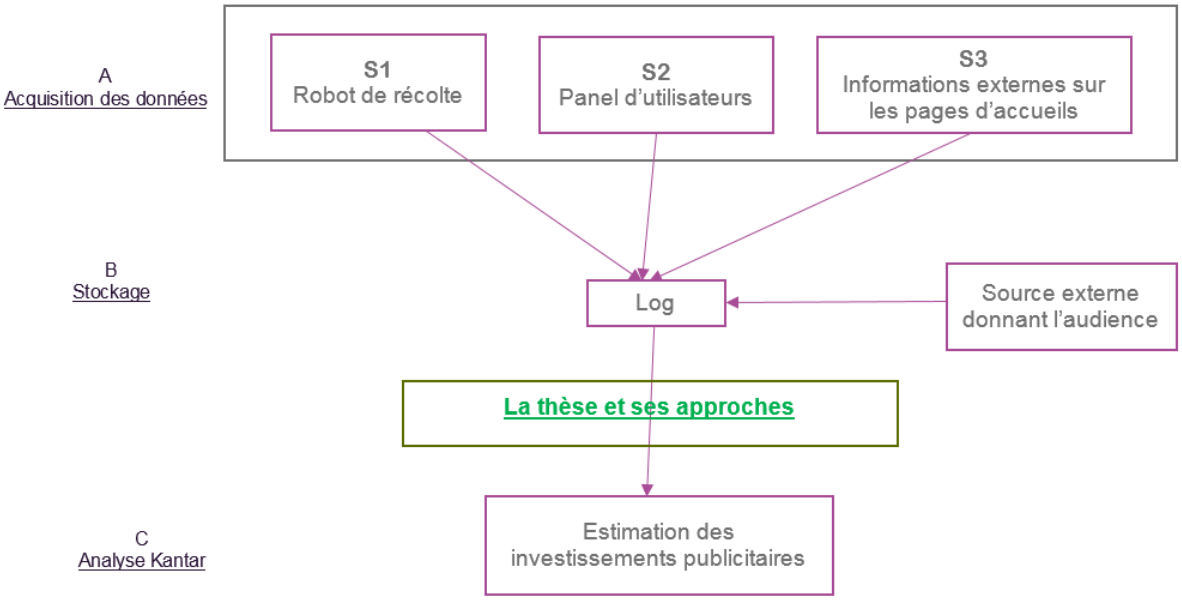


FIGURE 2.1 – Localisation du travail de la thèse par rapport au projet de l'entreprise

La figure 2.1 montre les différentes étapes du projet. Tout d'abord, l'information est récoltée par 3 sources différentes. Ensuite, elle est augmentée par extrapolation à partir des données d'audience venant d'une source extérieure. L'ensemble construit un log de données. À partir de ces logs, un ensemble de traitements est effectué pour obtenir les valeurs d'investissements estimées.

Le processus du projet MMS repose sur les trois étapes suivantes.

Étape A : Le projet met en œuvre une base de données de départ approvisionnée par 3 sources différentes.

- $\int_1$: cette source fournie les informations d'un ensemble de robots (scripts) qui captent les données des publicités, diffusées lors de leur passage sur 700 sites web, correspondant aux sites les plus visités de France.
- $\int_2$: cette source propose des données issues de la navigation d'un panel d'humains, effectuée dans le but d'enregistrer des données sur des publicités ciblées [4].
- $\int_3$: cette source procure des données sur les pages d'accueil de ces 700 sites web.

De manière pratique, l'ensemble des données est stocké dans une base de données non relationnelles de type document, où les logs, mais aussi les résultats des prétraitements sont des documents. Ce choix permet une certaine scalabilité du stockage des données et des traitements à effectuer.

4. Une publicité ciblée est une publicité destinée spécialement à un utilisateur au regard de sa navigation sur le web.

Etape B : Cette étape correspond à une phase d'enrichissement des données enregistrées dans la base précédente avec les données d'audience. Ces audiences sont données par page web et non par bannière. Une application, préalablement développée par la société, détermine à partir de ces audiences le nombre de vues associées à chaque bannière. Ceci donne, a priori, le nombre de personnes qui ont vu ou probablement cliqué sur les bannières récoltées dans une période de temps.

Etape C : Finalement, pour estimer les montants des investissements mis sur ces bannières, i.e déterminer l'investissement financier effectué par les annonceurs sur internet, il y a deux étapes majeures :

1. Chaque mois, Kantar se fonde sur une grille tarifaire qui détermine le coût estimé de chaque format de bannière sur le marché.
2. Kantar possède une information venant des régies publicitaires [5]. Ces régies fournissent les chiffres d'investissement d'une partie des annonceurs. Cette information est généralement trompeuse et ne reflète pas la réalité des chiffres réalisés vu les contraintes du métier, e.g. un annonceur peut interdire à une régie de publier ses chiffres d'achats de publicités.

En se basant sur ces données, Kantar livre des chiffres estimés mensuellement aux clients. Le projet MMS met en évidence différents problèmes.

Suite à de nombreux tests de comparaisons entre les données estimées par le processus du projet et les données réelles provenant des régies, il y a des écarts importants aussi bien sur les volumes d'impressions que sur les investissements publicitaires. Ces écarts sont assez variables et il est difficile d'en définir la ou les causes précises, compte tenu du grand nombre de paramètres qui rentrent dans les algorithmes de ce système.

Pour ce faire, le cadre principal de l'expérimentation visée à l'issu de l'analyse de ces données est d'isoler ces problématiques et d'essayer de les redresser. Ce besoin est né des deux constatations principales suivantes :

1. Les estimations sont le plus souvent sous-estimées. Ceci peut être dû au fait que certaines campagnes ne sont pas capturées par les mêmes sources, que les données sont redressées au niveau des URL, domaines et des sous-domaines.
2. Certaines typologies de campagnes ne sont pas correctement estimées. En effet, les volumes d'impressions sont souvent sous-estimés par les sources car elles ne capturent que certaines informations. Par exemple, la détection des publicités sur les pages web est en décalage avec la réalité du fait du dynamisme des publicités et du ciblage des consommateurs.

Nous présentons, maintenant, les problèmes sous-jacents à chacune des étapes du processus de traitement, mis en œuvre par la société.

5. Entreprises chargées de la vente d'espaces publicitaires aux annonceurs sur les différents médias, et support, comme par exemple sur le web.

Problèmes dans l'étape A :

- Les robots de la source $\int_1$ récoltent des données de publicités diffusées sur une page web. La majorité de ces données sont profilées par les navigateurs, i.e elles suivent la navigation et le comportement des utilisateurs grâce aux caches des navigateurs. Ainsi, au moment de la récolte, les bannières affichées ne sont pas les mêmes pour tous. Ici, il y a une perte remarquable dans la quantité et la variété des données stockées donnant lieu à un questionnement sur la qualité des informations générées à partir de ces données.
- Les robots utilisent des balises HTML particulières pour la récolte des bannières. Cependant, nous sommes conscients qu'il y a une perte dans la détection de bannières à ce niveau de par le dynamisme des codes des sites web. En effet, les sites web sont souvent mis à jour et les robots ont besoin de mises à jour pour qu'ils puissent suivre les nouvelles structures des pages. Malheureusement, cette dernière est rarement faite à cause des coûts humains et financiers.
- Les données de panel, issues des navigations humaines sur les sites web, ne présentent pas toutes les publicités affichées. Bien que ce panel puisse être représentatif des profils sociaux, il ne peut être exhaustif tant les profils des navigants sur le web sont variés et leurs historiques de navigation sont grands.

Problèmes dans l'étape B : Au niveau des estimations des audiences, le mécanisme du projet MMS enrichit les données, sur les bannières détectées dans l'étape A, par des sources externes donnant le niveau d'audience. Ces dernières fournissent l'information du trafic par page web et non par bannière. Ceci engendre les problèmes suivants :

- La fréquence temporelle de ces audiences, achetées de l'extérieur, n'est pas la même que celle des sources d'approvisionnement de l'étape A.
- Il peut y avoir des problèmes de correspondance et d'affectation entre nombre de vues et de bannières.

Problème Étape C : Cette étape cherche à explorer les montants investis à partir des données résultantes de l'étape précédente. Comme expliqué précédemment, les chiffres s'appuient aussi sur ceux fournis par les régies. Cependant, ces dernières peuvent ne pas être très fiables, de par de possibles contraintes du métier pouvant empêcher la fourniture de l'information exacte. Ainsi, dans cette étape, nous notons les problèmes, impactant la qualité des données et la véracité des résultats, suivants :

- La grille tarifaire, exploitée mensuellement par Kantar, n'est pas adaptée à chacun des sites web.
- Les chiffres résultants de l'étape précédente étant en cause, les résultats fournis par l'algorithme dans cette étape sont aussi questionnables.

D'un point de vu global, le processus de traitement est chaîné et linéaire. Chaque résultat d'une étape est l'entrée de la suivante. Aussi, comme les données initiales sont imparfaites, les données valorisées résultantes de chaque étape le sont, possiblement, aussi. Ainsi, les estimations finales des investissements publicitaires peuvent ne pas vraiment refléter la réalité du marché.

Nous constatons donc qu'en l'état, il est peu réaliste de prétendre atteindre les vrais chiffres concernant les investissements web mis en œuvre dans une campagne publicitaire. Aussi, cela fait apparaître clairement le besoin d'un système de supervision vérifiant la qualité des données valorisées et donnant des pistes d'amélioration et de détection des potentielles défaillances.

Pour conclure cette section, nous constatons que les données enregistrées sont susceptibles d'être imparfaites, à minima imprécises.

2.2.2 Caractéristiques

Dans cette partie, nous nous focalisons sur la nature de cette imperfection.

La figure 2.2 montre, d'une manière globale, le caractère temporellement très variant des données recueillies sur un échantillon de sites web, tirés aléatoirement, appartenant à une même catégorie d'usage cat_1. Une catégorie d'usage correspond à un label affecté par des experts aux sites web observés en fonction de leur objet (e.g., sites sportifs, sites d'annonce, etc.).

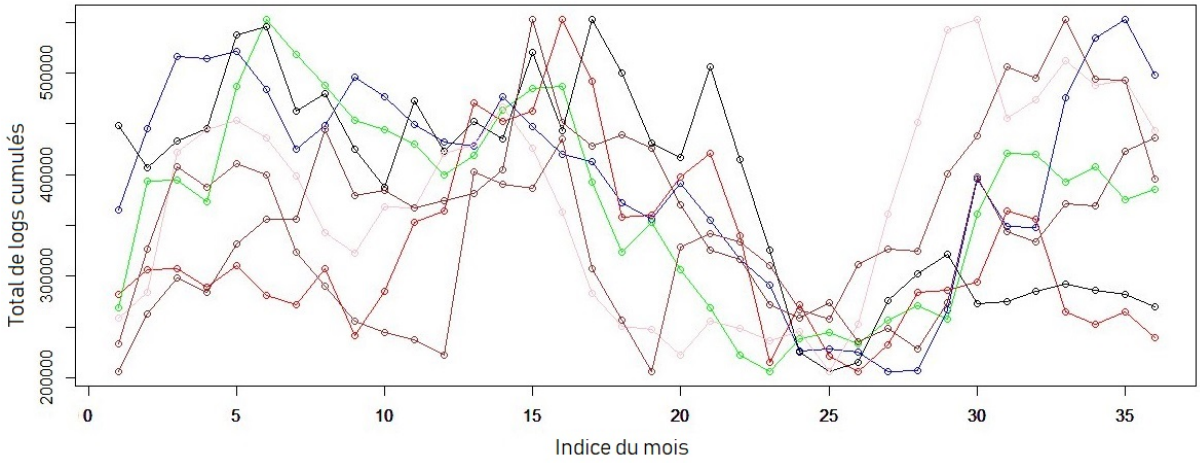


FIGURE 2.2 – Échantillon d'enregistrements temporels (en mois) par des capteurs en fonction du volume des logs pour la catégorie cat_1

Nous pouvons visuellement constater l'existence de beaucoup de chevauchements dans la figure 2.2. Les récoltes des publicités des sites de cette catégorie ont un caractère très volatile. Ces variations sont incompréhensibles sans connaissances préalables. Nous chercherons, dans les approches présentées dans ce manuscrit, à comprendre leur variabilité et le degré de leur qualité.

Que ce soit par une simple visualisation, comme expliqué précédemment, ou par des approches de fouilles de données, nous pouvons distinguer de possibles problèmes dans la récolte.

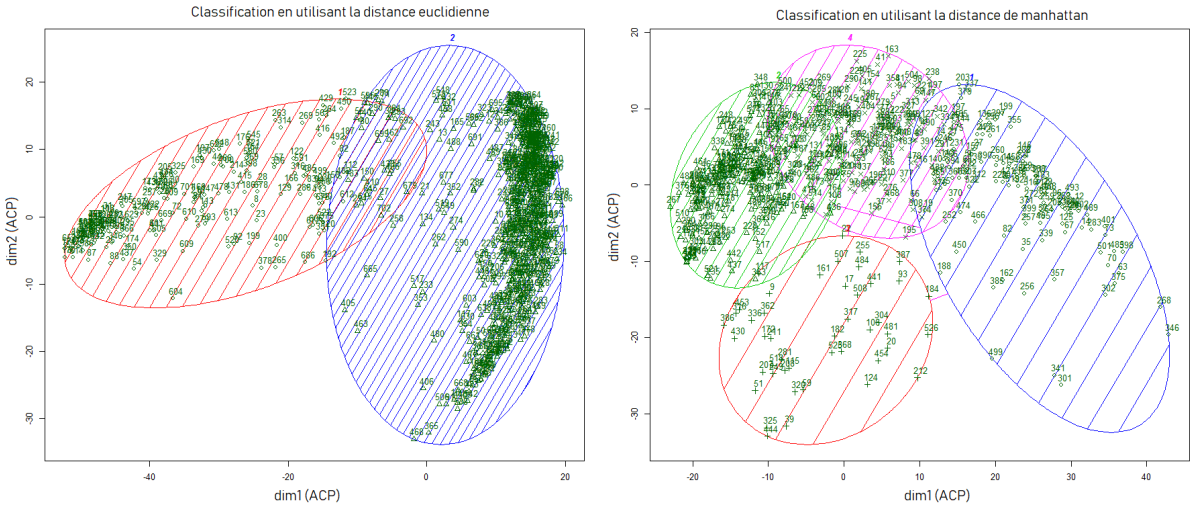


FIGURE 2.3 – Illustration des groupements, effectués à l'aide des k-moyennes, des données brutes, selon des k optimaux et des jeux de données différents

Prenons l'hypothèse que les stratégies de récoltes peuvent être le reflet des catégories d'usage définies par nos experts. La figure 2.3 montre une application d'un algorithme de clustering simple (k-moyennes [Mac67]) sur un jeu de données de 3 ans. Le facteur k est un facteur optimal [6]. La figure montre ainsi le nuage de points de diverses données groupées en 2 et/ou 4 groupes optimaux, mais en réalité nous possédons plus de 20 groupes de catégories de sites web. Ainsi, les stratégies de placement publicitaire sur les différents médias d'une catégorie ne sont pas forcément très différentes des celles des médias des autres catégories.

Ces données temporelles sont, aussi, le sujet d'une forte variation, soit au sein des catégories elle-mêmes, ou bien entre elles. Ainsi, les récoltes concernant deux sites web d'une catégorie, par exemple Sport, dans le marché des publicités, n'ont pas forcément un comportement similaire. La figure 2.2 est un exemple présentant des données récoltées sur des sites web d'une même catégorie. Nous pouvons nous apercevoir que certaines séries temporelles ont des comportements très différents des autres.

Une autre idée, sur les publicités, pourraient être que les récoltes de données aient un comportement saisonnier ou fréquentiel comme pourrait le laisser penser la figure 2.4. Dans cette figure, nous avons regardé le volume total de logs en agrégeant par mois pour les trois ans. Par exemple, le volume recueilli pour le mois 1 correspond aux volumes cumulés des mois de janvier 2015, janvier 2016 et janvier 2017. Ainsi, les récoltes effectuées durant les mois de mai sont inférieures à celles des autres mois. Cependant, la figure 2.5

6. le nombre de clusters donnant le maximum de l'indice de silhouette [Rou87]

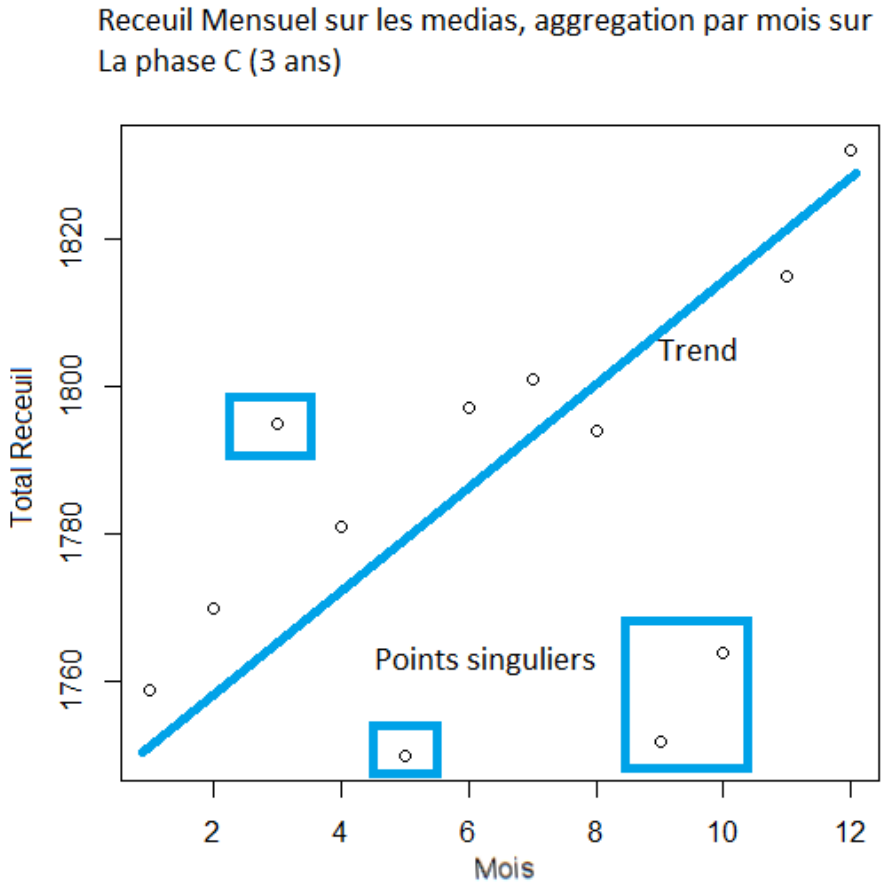


FIGURE 2.4 – Récolte de données tendancielles présentant des mois singuliers de récolte

permet de mettre en doute cette hypothèse, ce qui a été confirmé par d'autres informations indiquant que ces différences, dans les récoltes, sont plus dues à des conjonctions dans des changements de politiques de récoltes, qu'à des phénomènes saisonniers.

Nos données souffrent, aussi, de problèmes de discontinuités, comme l'illustre la figure 2.6. Dans plusieurs cas, la durée de l'absence de données pour un média peut être assez importante ou significative. Nos données sont, donc, aussi assujetties à de l'incomplétude.

En conclusion, les données de l'entreprise sont imparfaites. Cette imperfection se résume en deux facteurs essentiels : imprécision et incomplétude. Celles-ci donnent aussi lieu à de l'incertitude dans les résultats des traitements effectués sur nos données.

2.3 Préparation des données

Nous avons décrit, dans les sections précédentes, la nature des données de l'entreprise, la manière dont elles sont acquises, ainsi que les éventuelles imperfections qu'elles présentent.

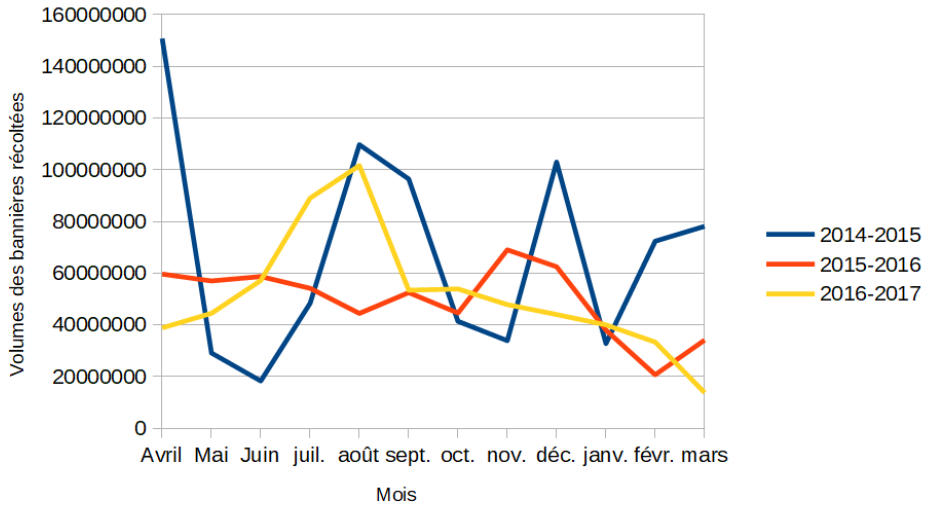


FIGURE 2.5 – Évolution des volumes de données recueillies par mois et par année de récoltes

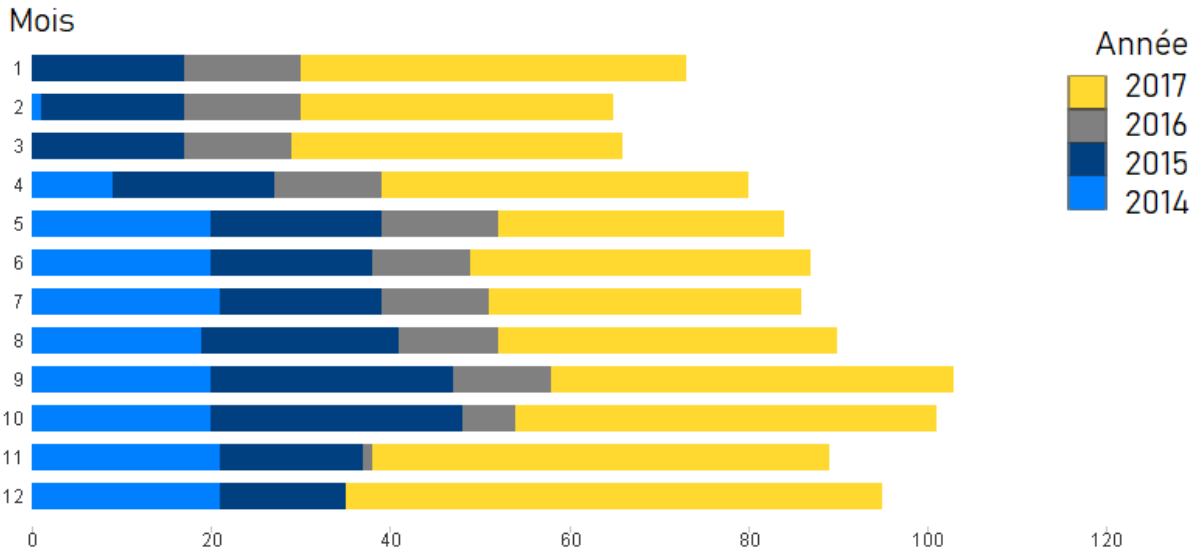


FIGURE 2.6 – Volume des absences dans les recueils journaliers par mois et par an

Nous présentons dans cette section les pré-traitements effectués, fournissant les données d'études exploitées dans ce manuscrit.

Pour pouvoir travailler sur la qualité des données, nous avons opté pour un des framework associé à ce contexte. La découverte de connaissance, en anglais Knowledge data discovery (KDD), est un ensemble de processus et modèles d'analyse à suivre pour la découverte de connaissances dans les bases de données. Les techniques d'exploration de données nécessitent plusieurs phases : la préparation, la sélection de données, le nettoyage de données, l'intégration des connaissances préalables sur les ensembles de données, l'ob-

tention de modèles sur les données, l'interprétation de solutions, etc. Un travail préliminaire pour la préparation des données est nécessaire en vue de leur qualification et de leur récolte.

Comme indiqué précédemment, nos données sont stockées dans une base centralisée. Vu que les données sont récoltées des sites Web par plusieurs capteurs formant l'ensemble des capteurs S, et plusieurs sources formant l'ensemble des sources $\mathcal{S}$, nous les agrégeons sur diverses variables permettant d'étudier leur qualité et leur véracité.

La figure 2.7 illustre ce processus de traitement où n capteurs captent p (dans la figure $p = 2$) types d'informations sur des sites web en fonction des variables (var1, var2 et la date). Ces données sont ensuite agrégées selon m instant (timestamps). Ainsi, les n robots/capteurs produisent des journaux contenant plusieurs informations ($var1$, $var2$, date). Ces informations sont agrégées en fonction des différents instants d'études afin de calculer les variables (v_1, v_2) qui sont analysées par nos outils.

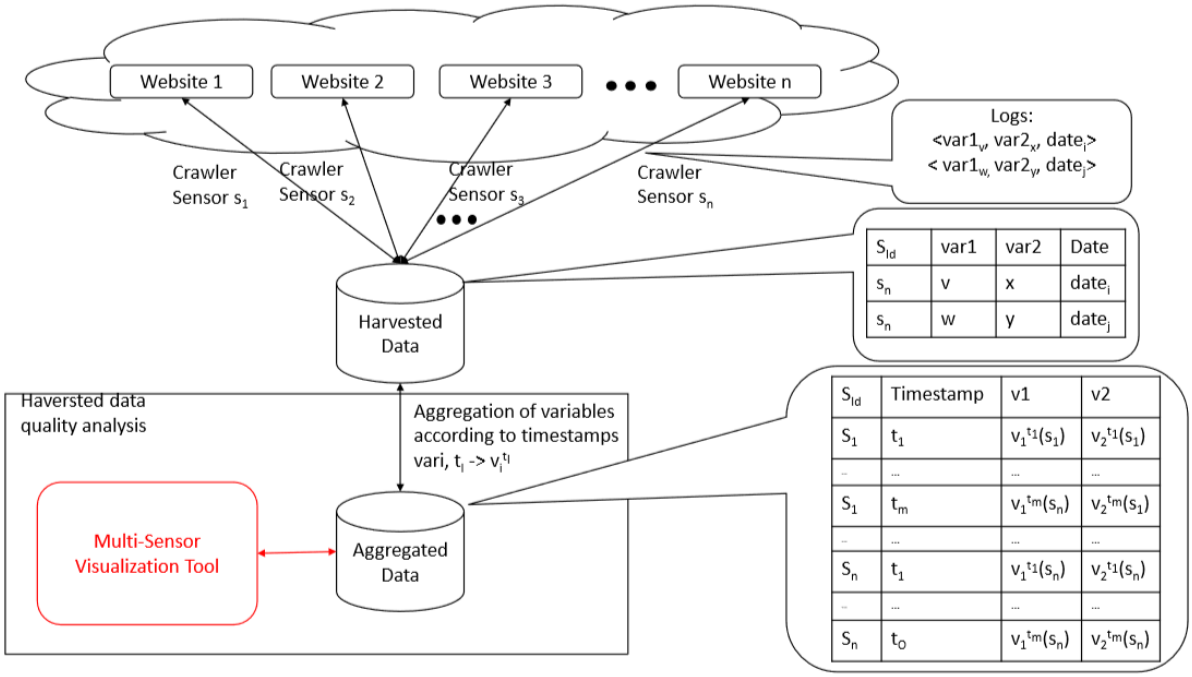


FIGURE 2.7 – Pipeline de la récolte des données à l'analyse de leur qualité.

Les données étudiées représentent les informations récoltées depuis 700 sites français. Les données sont enregistrées temporellement selon une fréquence moyenne de 2 minutes par page web, ce qui fait un important volume de données à traiter.

Nous avons stocké les données sur MongoDB [7], qui grâce à son principe d'indexation permet d'optimiser l'accès aux données.

7. Système de gestion de bases de données NoSQL : `https://www.mongodb.com/fr`

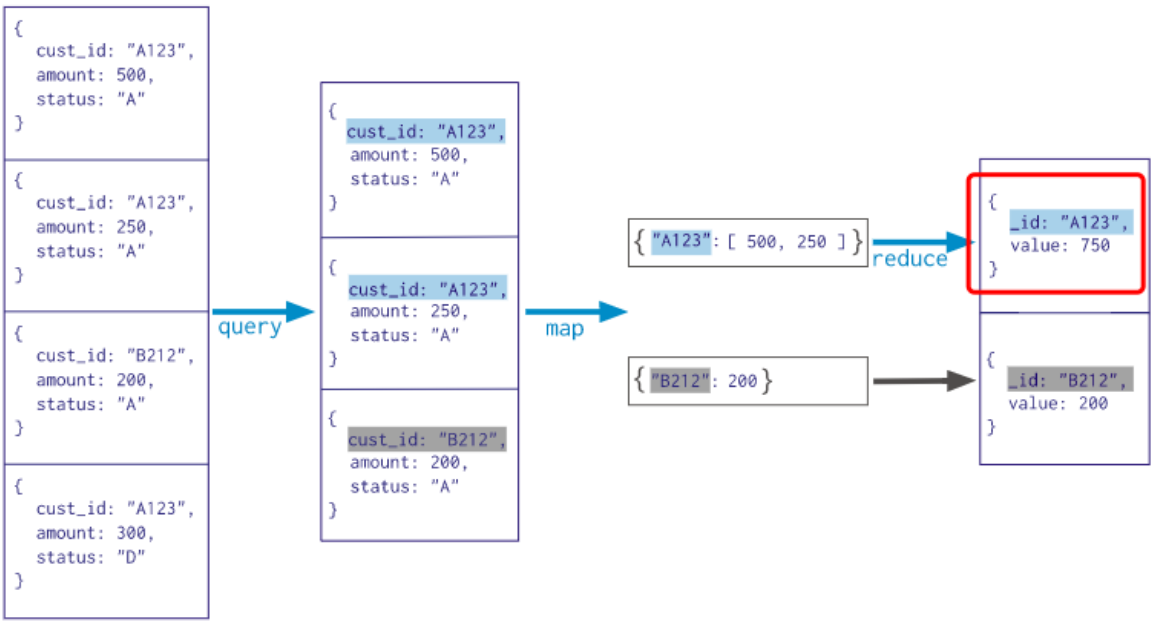


FIGURE 2.8 – Principe de MapReduce

Ensuite, nous cherchons à agréger les données par capteur, par variable et par mois (ou par jour), afin de produire des tableaux de données de taille compatibles avec l'utilisation d'outils d'analyse et de visualisation tel que R et QlickView. Afin de fournir ces agrégats, nous effectuons une transformation exploitant le modèle de MapReduce [DG08] (cf. figure 2.8). Ce modèle permet le traitement résilient et distribué de grandes quantités de données.

Les données sont finalement réduites comme le montre la figure 2.9. La figure 2.10 montre un échantillon de données transformées.


```
"ns" : "Avicenne.avicenne",
"size" : 50575913643.0,
"count" : 62223080,
"avgObjSize" : 812,
"storageSize" : 22407639040.0,
"indexSizes" : {
    "_id_" : 628776960,
    "jour_1_mois_1_année_1_media_text_categorie_text" : 2010193920.0,
    "jour_1" : 279990272,
    "mois_1" : 279986176,
    "année_1" : 280002560,
    "categorie_1" : 305233920,
    "media_1" : 314781696,
    "banniere_1" : 422825984,
    "urlDetails._id_1" : 2640875520.0,
    "media_1_année_1_mois_1" : 315080704,
    "banniere_1_media_1_année_1_mois_1" : 717393920,
    "année_1_mois_1" : 280109056,
    "année_1_banniere_1" : 425549824,
    "secteur_1" : 308371456,
    "mois_1_secteur_1" : 308465664
}
```


FIGURE 2.9 – Statistiques sur la réduction des données en base non relationnelle

```
{
    "_id" : ObjectId("5c62cf46518afe2cc8d8e2b9"),
    "année" : 2014,
    "mois" : 1,
    "jour»: 1,
    "media»: "@ L EXPRESS",
    "categorie»: "MEDIAS GENERALISTES PRESS",
    "banniere" : "84EF4F9B0523A95CA0021CAE01B7F866",
    "occurences" : 7,
    "urlDetails" : [
        {
            "_id" : "932669780709769107",
            "occurences" : 1
        },
        {
            "_id" : "7687902664862094441",
            "occurences" : 1
        },
        {
            "_id" : "7421014804224245802",
            "occurences" : 1
        },
        {
            "_id" : "-7273084931137422957",
            "occurences" : 1
        },
        {
            "_id" : "450412974998451329",
            "occurences" : 1
        },
        {
            "_id" : "1842807922796531463",
            "occurences" : 1
        },
        {
            "_id" : "66805031141819755585",
            "occurences" : 1
        }
    ],
    "produit" : "MAZDA 3",
    "segment" : "MODELES AUTO",
    "group" : "AUTOMOBILES",
    "mega_group" : "AUTOMOBILES POIDS LOURDS",
    "secteur" : "AUTOMOBILE TRANSPORT",
    "sous_secteur" : "AUTOMOBILES POIDS LOURDS",
    "source" : "Evaliant",
    "emplacement" : "Internet",
    "type_media" : "Site web"
}
```


FIGURE 2.10 – Échantillon de données en Json suite à une transformation dans MongoDB

2.4 Présentation des données pré-traitées

Cette section décrit les données pré-traitées qui sont en entrée de nos processus d'analyse. La période d'étude porte sur 3 ans (36 mois, 1095 jours). 700 sites sont observés par au moins une source parmi nos trois sources présentées précédemment. Le volume global de données à pré-traiter s'élève à 50Go et 62223080 logs au format journalier (logs re-

Variable	v_1	v_2	v_3
Moy	184	1369	104,1
Min	1	1	1
Q1	26	250	27
Mediane	36	718	70
Q3	53	1640	147
Max	12530	73690	6087

TABLE 2.2 – Statistiques résumant le jeu de données journalières à l'issue du prétraitement

Variable	v_1	v_2	v_3
Moy	1877,6	40170	1555
Min	1	1	1
Q1	255	10500	29
Mediane	706	23850	40
Q3	1251	48250	65
Max	59920	1444000	168400

TABLE 2.3 – Statistiques résumant le jeu de données mensuelles à l'issue du prétraitement

groupés par jour) et à 20Go au format mensuel (logs regroupés par mois). Le pourcentage global de valeurs manquantes dans la structure est de 6.9% pour les données mensuelles et de 24.53% pour les données journalières.

Nous étudions les données selon trois variables (v_1, v_2 et v_3). Les variables représentent : v_1 le nombre des bannières récoltées par les robots, v_2 le nombre des bannières unique récoltés et v_3 le volume des urls sur-lesquels apparaissaient les bannières. Les résumés statistiques de ces variables par jours et par mois sont donnés respectivement dans les tableaux 2.2 et 2.3. Nous pouvons voir qu'il y a une forte dispersion dans les valeurs.

Par ailleurs, selon une classification interne à la société, chaque site observé a été associé à une unique catégorie, elle même incluse dans une méta-catégorie. Il y 56 catégories définies dans la classification, et 12 méta-catégories.

2.5 Formalisation

Dans la suite de ce manuscrit, nous considérons les données pré-traitées. Nous nous positionnons principalement sur l'axe média – études des données recueillies pour chaque média – pour l'analyse de la récolte. Nous étudions l'ensemble des médias sans considération de leurs catégories (excepté pour l'outil de visualisation présenté dans la partie IV.

Nous considérons, à la vue de la multiplicité des mécanismes d'acquisition mis en œuvre pour un site, que les données acquises pour un site sont issues d'un capteur dont on ne connaît pas le fonctionnement interne (« boîte noire »). À chaque site est associé un unique capteur et chaque capteur correspond à un unique site.

Ainsi, nous considérerons par la suite les notations suivantes :

— Cat est l'ensemble des ca catégories : $Cat = \{cat_1, \ldots, cat_{ca}\}$

— $MCat$ est l'ensemble des m_{ca} méta-catégories : $MCat = \{mcat_1, \ldots, mcat_{m_{ca}}\}$

— $\mathcal{S}$ est l'ensemble des so sources : $\int = \{\int_1, \ldots, \int_{so}\}$

— S est l'ensemble des n capteurs : $S = \{s_1, \ldots, s_n\}$.

— T est l'ensemble des m *timestamps* : $T = \{t_1, \ldots, t_m\}$, un timestamp correspond à une agrégation à une échelle temporelle donnée.

— V est l'ensemble des p variables étudiées : $V = \{v_1, \ldots, v_p\}$.

— $v_i^{t_k}(s_j)$ est la valeur v_i au timestamp t_k pour s_j.

Pour nos données, on a :

— $ca = 56$.

— $m_{ca} = 12$.

— $n = 700$.

— $m = 36$ à l'échelle mensuelle et $m = 1095$ à l'échelle journalière.

— $p = 3$.

2.6 Conclusion

Dans ce chapitre, nous avons présenté le jeu de données à analyser qui présente des données issues de la diffusion de publicités sur Internet (secteur d'activité de Kantar). Elle sont principalement captées par différentes sources. Les données se caractérisent par leur volumétrie, la temporalité, les différents axes et variables d'études, etc. Elles présentent aussi certains problèmes comme par exemple : la forte variation des valeurs (données intrinsèquement peu comparables), des données singulières, des données manquantes, etc. Ces dernières sont donc imparfaites car elles sont imprécises et incomplètes.

La partie suivante présente les travaux connexes à l'étude de la qualité des séries temporelles de données et à la représentation des données imparfaites. Elle donne aussi notre positionnement par rapport aux approches et aux techniques de la littérature.

Deuxième partie

Etat de l'art

Chapitre 3

Qualité - Véracité

3.1 Introduction

De nos jours, les activités et la prise de décisions dans une organisation reposent sur des informations obtenues à partir des données brutes et des données analysées. Les informations générées sur cette base sont à l'origine de divers services, mis en place afin d'atteindre un objectif métier.

Comme les données constituent des ressources importantes dans toutes les organisations, la qualité des données est essentielle pour que les gestionnaires et les processus opérationnels identifient les problèmes de performance associés. De plus, les données de haute qualité peuvent augmenter les chances d'obtenir les meilleurs services dans une organisation. Par ailleurs, à l'égard des spécifications des systèmes d'information des entreprises et les nouveaux enjeux e.g la scalabilité, la volumétrie des données, la dépendance avec des facteurs extérieurs, etc., les techniques traditionnelles d'évaluation de la qualité des données ne permettent pas forcément de générer des informations en adéquation avec les processus métier [Ben+18b].

Le doute envers les informations générées mène à un questionnement important. Les informations générées sont-elles légitimes ? À quel degré pouvons-nous leur faire confiance ? Généralement, l'étude de la véracité des informations est un complément de l'étude qualité à l'égard de l'assurance de l'utilisabilité et de la fiabilité des données, dans un contexte surtout industriel.

Dans ce chapitre, nous commençons (section 3.2) par étudier les approches liées à l'étude de la qualité des données. Par la suite, dans la section 3.3, nous les complétons par l'étude des approches touchant à la véracité des informations.

3.2 Qualité des données

La qualité des données est un domaine de recherche actif en statistique depuis la fin des années 1960 [FS69]. À la fin des années 1990, les informaticiens ont commencé à travailler sur la mesure et à l'amélioration de la qualité des données [SLW97 ; JV97 ; Ber99].

Dans ce contexte, Wang [Wan98] suggère que la qualité d'une donnée doit être définie selon l'usage attendu par son utilisateur. Redman [Red97] considère le concept de qualité des données selon quatre dimensions : l'exactitude, la perfection, la fraîcheur et l'uniformité. Wang propose une méthode itérative d'amélioration de la qualité des données nommée Total Data Quality Management (TDQM) qui est structurée en quatre phases (définir, réaliser, contrôler, agir). Par ailleurs, Peralta propose dans sa thèse [Per06] de mesurer la qualité des données en fonction de leur but.

Au vu des diverses définitions possibles dans la littérature et qui dépendent du domaine d'application étudié, dans cette thèse, nous définissons la qualité de donnés comme suit (Définition 1) :

Definition 1. *La qualité des données est une perception ou une évaluation qui informe sur la cohérence, l'efficacité et la fiabilité des données. La qualité des données est déterminée par certains indicateurs tels que la continuité, l'exhaustivité, la variabilité, la stabilité, etc.*

3.2.1 Concepts et définitions

L'analyse de la qualité des données (QD) suit souvent le principe «adéquation à l'usage», c'est-à-dire la capacité d'une analyse à répondre aux besoins de jugement sur les données manipulées [WS96]. Ces données sont souvent évaluées sur la base de différentes dimensions.

Comme de nombreuses et divergentes définitions de la qualité ont été proposées dans la littérature, il existe plusieurs types de dimensions. Les dimensions les plus utilisées dans les travaux traitant de la qualité des données sont : l'exactitude, l'exhaustivité, la cohérence et la temporalité [Wan98 ; JV97].

À l'instar des divergences existantes sur la définition même du concept de qualité des données, il existe aussi des différences importantes sur l'ensemble de ses dimensions, ou sur la signification exacte de chacune, et cela bien que des normes (e.g [ISO11]) y fassent référence mais en se focalisant sur les principes de la qualité à prendre en considération.

Le tableau 3.2.1 illustre certaines dimensions de la qualité des données liées à notre travail. Du point de vu recherche, les dimensions de la qualité doivent s'adapter aux types de données, à leurs caractéristiques, et aussi au processus critique sur lesquels elles sont manipulées.

Dimension	Définition
La temporalité	La temporalité fait uniquement référence au délai qui s'écoule entre le changement d'état du monde réel et la modification de l'état du système d'information qui en résulte
La cohérence	C'est le degré de présentation des données dans un même format qui doit être compatibles avec les données précédentes
La précision	Les données sont précises lorsque les valeurs de données stockées dans la base de données correspondent aux valeurs réelles
La complétude	C'est le degré de présence des valeurs dans une collection de données
Accessibilité	Ce principe est étendu à la disponibilité des informations ou à la possibilité de les retrouver facilement et rapidement
Croyabilité	C'est la mesure dont laquelle l'information est considérée comme vraie et crédible
Fiabilité	C'est la capacité de la fonction à maintenir un niveau de performance spécifique lorsqu'elle est utilisée dans des conditions particulières

TABLE 3.1 – Exemple de dimensions de la qualité de données [Wan98 ; JV97]

3.2.2 Principales approches

Stratégies de surveillance des données manipulées

Il existe deux types de stratégies qui peuvent être adaptées pour améliorer la qualité des données, à savoir les stratégies dirigées par les données (Data-Driven) et les stratégies dirigées sur un processus (Process-Driven). Chaque stratégie utilise des techniques différentes [Sid+12]. L'amélioration de la qualité des données est l'objectif de chacune.

Stratégie dirigée par les données

La stratégie dirigée par les données est une stratégie visant à améliorer la qualité des données en modifiant directement leurs valeurs. Grâce à cette technique, les valeurs des données obsolètes sont mises à jour en actualisant la base de données. Parmi les techniques d'amélioration liées aux données, nous citons : la normalisation, la localisation et la correction des erreurs, etc. Dans ce cas, Held et al. [HL12] propose une approche fondée sur les commentaires de différents experts pour déterminer les valeurs qui doivent apparaître. Son approche a pour but d'identifier et valider une nouvelle dimension de la qualité des données.

Le couplage de données est également un autre moyen [Bat+09]. Il relie et combine les sources d'entrée dans une seule population en supprimant les valeurs en double/incorrectes. Dans ce cas, de nombreuses techniques y sont appliquées comme indiqué par batini et al. dans [Bat+09] :

— Techniques probabilistes et statistiques.

— Techniques empiriques utilisant des techniques algorithmiques telles que le tri, l'analyse d'arbres, la comparaison de voisins et l'élagage.
— Techniques fondées sur la connaissance, extrayant la connaissance de fichiers et appliquant des stratégies de raisonnement.

Selon Gupta et al. [GH11], il est recommandé de travailler sur **l'évaluation de la qualité des données avant d'appliquer une telle méthode**. Cependant, les données doivent être dans une structure bien définie et avec des relations cohérentes. Il est également recommandé de vérifier leurs significations pour éviter les informations erronées et les résultats confus.

Painter et al. [Pai+11] ont travaillé sur des données temporelles intégrées qui proviennent de divers capteurs. Les données sont considérées comme finales après plusieurs jours d'arrivée. L'approche utilise une régression par quantile pour évaluer la précision des enregistrements de chaque capteur, en calculant un rapport entre les données incomplètes et les données existantes avec un seuil de précision.

Plus globalement, la normalisation [HL12], la liaison d'enregistrements [Bat+09] et l'intégration de schémas [Pai+11] représentent les techniques les plus adoptées dans les méthodologies de la qualité dirigées par les données.

En effet, les stratégies dirigées par les données doivent prendre en considération la nature des données qu'elles manipulent. Dans le cas des données temporelles incomplètes, la normalisation ne peut pas être appliquée seule, car cela risque de provoquer une perte d'informations dans les traitements. En effet, moyenner des périodes de discontinuités différentes ne génère pas une information pertinente. Cela invite quiconque à prendre du recul sur les techniques à considérer dans ses analyses.

Stratégie dirigée par un processus

L'analyse de la qualité des données dirigée par un processus consiste à redéfinir le processus sur lequel les données sont examinées. Cette analyse consiste à produire ou à modifier les données afin d'améliorer leur qualité. C'est une stratégie qui améliore la qualité en mettant en œuvre un processus qui crée ou modifie des données.

Cette méthodologie adopte des techniques d'ingénierie qui permettent d'évaluer la qualité sur plusieurs étapes [MWC99], tout en reposant sur le processus de la production des données nommées dans [BT99] par «système de fabrication de l'information pour le produit d'information».

Dans [SWZ00], les auteurs exploitent cette méthodologie pour définir un modèle de production d'information fondé sur une carte (IP-MAP). Celle-ci modélise les sous-processus du traitement en les affectant à un modèle d'analyse graphique. Afin de juger de leur qualité, l'approche identifie en amont les phases critiques du processus, et donne les limites organisationnelles. Une fois les limites identifiées, des activités de contrôle qualité sont mises en place.

Des solutions complexes telles que IP-MAP ne peuvent pas toujours être adoptées en raison de leurs coûts élevés et, dans certains cas, l'impossibilité pratique d'une étape de modélisation approfondie du processus. Pour cette raison, d'autres méthodologies adoptent des solutions moins formelles, mais plus réalisables. Pour Scannapieco et Catarci [Bat18], par exemple, la définition repose sur un ensemble de matrices décrivant les principales relations entre les données, flux d'information, processus et unités organisationnelles.

Les techniques orientées processus sont plus performantes que les techniques dirigées par les données, sur une longue période, car elles éliminent complètement les causes profondes des problèmes de qualité. En revanche, l'utilisation de ces techniques est coûteuse par rapport à sa mise en place, mais elle est efficace en terme de résultats [Sid+12].

Dans le cas où les données se caractérisent par une temporalité et une imperfection, se diriger vers les stratégies orientées processus est une contrainte à garantir. Par exemple, pour les données Kantar qui s'accumulent par le biais des capteurs et sont synthétisées suite à l'application des différentes analyses, l'application des processus de traitement est le choix le plus adéquat à l'analyse de leur qualité. Cependant, pour avoir une vision claire sur la conception des processus et ses mises en œuvre, il va falloir adapter un modèle de qualité particulier.

Modèles de qualité de données

En pratique, l'évaluation de la qualité des données dans les systèmes de base de données a principalement été réalisée par des évaluateurs professionnels avec des audits de plus en plus compétitifs. Les approches de la gestion de la qualité industrielle et l'évaluation de la qualité logicielle proposent des extensions de gestion des métadonnées. Ainsi, l'utilisation de métadonnées pour l'évaluation et l'amélioration de la qualité des données a été préconisée par de nombreux auteurs [BP95 ; Ber07b ; Bea05]. Rothenberg [J96] a fait valoir que les producteurs d'information devraient effectuer la vérification, la validation et la certification (VV & C) de leurs données et qu'ils devraient fournir des métadonnées sur la qualité des données avec les jeux de données [J96]. Plusieurs propositions intègrent pleinement les métadonnées dans la conception de leurs processus qualité. Parmi ces approches, le programme TDQM (Total Data Quality Management) proposé par Wang et al. [WS96 ; Wan98] fournit une méthodologie incluant la modélisation de la qualité des données dans le modèle conceptuel Entité-Relation.

Il propose également des directives pour l'ajout de métadonnées étape par étape. D'autres travaux ont proposé des approches (similaires) pour modéliser et capturer la qualité de données relationnelles [PA03 ; ST03] et de données semi-structurées (par exemple, D2Q [SE04]).

La figure 3.1 présente un modèle générique de la qualité des données avec le formalisme UML qui synthétise les aspects communs des approches. Une ou plusieurs données des instances de qualité peuvent être associées à un élément de données (c'est-à-dire une valeur d'attribut, ensemble de valeurs, enregistrement, table, domaine, etc.). La qualité des données est composée d'une ou plusieurs dimensions avec des attributs publics représentant le type et la catégorie de la dimension de la qualité composée d'une ou de plusieurs mesures caractérisées par leur type, leur métrique et leur description. Chaque mesure a une ou plusieurs valeurs avec l'unité et la date de mesure correspondantes.

Nous remarquons que la plupart des modèles de qualité des données proposés reposent sur les métadonnées.

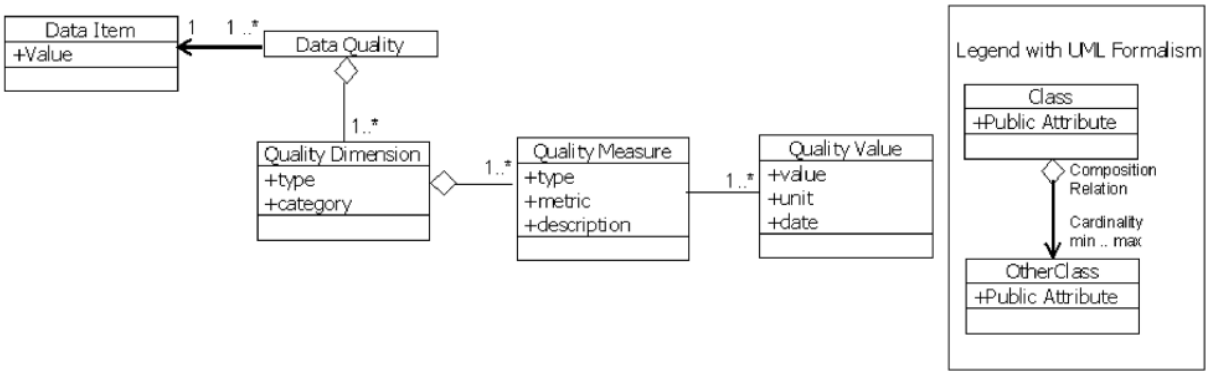


FIGURE 3.1 – Modèle de la qualité des données selon Berti-Equille et al. [Ber07a]

Ces approches reposent sur des méthodes de génération des métadonnées. Une métadonnée représente une caractéristique formelle normalisée et structurée utilisée pour la description et le traitement des contenus des ressources manipulés. Cependant, ces métadonnées ne sont pas toujours disponibles et leur génération ne suit souvent aucune norme qualité car elles sont spécifiques à la problématique traitée.

Bien que des efforts considérables aient été investis dans le développement des métadonnées standards pour l'unification de l'échange d'informations entre différentes ressources et de connaître les différents échanges entre les ressources d'un processus qualité, d'autres propositions ont donné une attention particulière aux journaux des données (fichiers log) i.e à l'historique de la transformation des données en informations [SVS05]. Contrairement aux métadonnées qui ont pour but d'agréger les constatations, les fichiers logs sont très utiles pour l'analyse et l'interprétation des données, par exemple, le débogage, la mise en œuvre de boucles de rétroaction de la qualité et l'analyse des causes d'erreurs de données, etc.

Il est donc souhaitable que les données puissent être contrôlées. Dans le cas où les données ne sont pas normalisées, certaines transformations sont nécessaires pour générer un processus de contrôle approprié. Les journaux de données construisent aussi un point fort considérable dans notre contexte du fait du suivi d'une donnée brute jusqu'à sa valorisation. En effet, pour produire un processus de contrôle des données il faut mettre en

place un processus de gestion particulier comme, par exemple, dans Painter et al. [Pai+11] qui proposent une approche fondée sur les quantiles pour étudier le comportement des données.

Processus de gestion de la qualité des données

La qualité des résultats de l'exploration finale ainsi l'interprétation des résultats reposent essentiellement sur le processus de préparation des données préétablit et sur la qualité des jeux de données analysés.

En effet, les processus et applications d'exploration de données nécessitent diverses formes de préparation, de correction et de consolidation de données, combinant des opérations de transformation de données complexes et des techniques de nettoyage.

Cela est dû au fait que les données d'entrées dans les algorithmes d'exploration de données sont supposées être conformes aux distributions de données et ne contenant aucune valeur manquante ou incohérente, incorrecte, etc.

La vue d'ensemble du cadre de ces manipulations est illustrée à la figure 3.2 et ce en suivant le framework KDD [GH07].

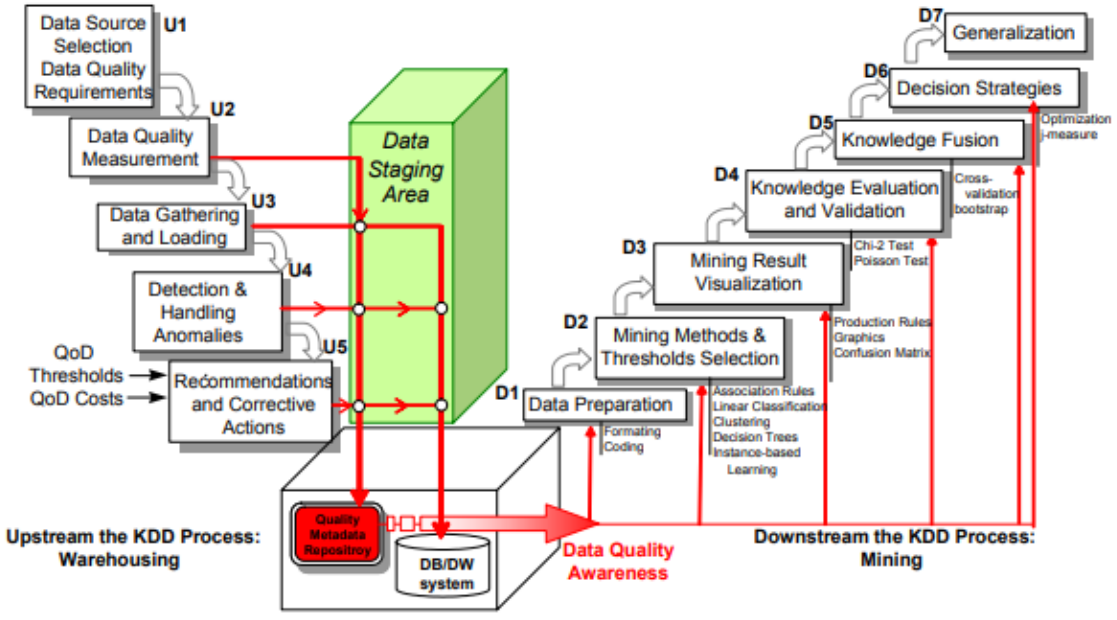


FIGURE 3.2 – Cadre général pour l'information sur la qualité des données dans un processus de découverte de connaissances (framework KDD) [GH07]

Analyse de la qualité des données pour des données fortement dynamiques : cas des données web

Après avoir vu les différentes stratégies d'analyse de données, les modèles et l'utilité d'adapter un processus de gestion, dans cette partie de l'état de l'art nous présentons des travaux connexes à notre domaine d'étude dans un processus qualité : l'analyse des données web.

Avec l'évolution des technologies, la nature des systèmes d'information évoluent et de nouveaux types de systèmes d'information font leur apparition. En ce qui concerne les problèmes liés au Web, les systèmes et les méthodologies associées répondent à plusieurs problèmes e.g., la qualité des données non structurées.

Pernici et Scannapiecov [SE04] proposent un modèle associant des informations de qualité avec les données Web récoltées, à savoir avec chaque élément d'une page Web, avec des pages et avec des groupes de pages. Plusieurs dimensions sont considérées, telles que la volatilité, la complétude et la précision sémantique et syntaxique.

La qualité des documents web revêt une pertinence croissante, car le nombre de documents gérés au format web augmente constamment. Rao et al. [RR40] ont montré qu'au début des années 2000, 40% de la matière sur le net disparaît dans un délai d'une année, alors que 40% supplémentaires sont modifiées, ne laissant que 20% dans leur forme originale. Toujours, au début des années 2000. Hey et al. [Hey04] indiquent que la durée de vie moyenne d'une page Web était de 44 jours et que le Web changeait complètement environ quatre fois par an.

En conséquence, la préservation des données Web devient de plus en plus cruciale. Le terme préservation du Web indique la capacité d'empêcher la perte d'informations sur le Web en stockant des versions significatives de documents Web. Ceci reflète le cas des données issues des capteurs et leur mise en étude. En effet, sur divers domaines comme le marché financier, les données changent quotidiennement et le faite d'étudier l'historique pour prévoir des comportements futurs reste un peu limité. Ceci n'est pas applicable forcément dans d'autres domaines comme les médias, car ce dernier est en parti cyclique/saisonnier. Par exemple, la rentrée scolaire, Noël, le Nouvel An, etc. sont des marqueurs systémiques dans le marché publicitaire.

Cappiello et al. [Cap15] proposent une méthodologie pour soutenir le processus de préservation tout au long du cycle de vie de l'information, de la création à l'acquisition, en passant par le catalogage, le stockage et l'accès. Les principales phases de la méthodologie sont comme suit :

1. Chaque fois qu'une nouvelle page est publiée, les données sont associées à des métadonnées, décrivant leur qualité, en termes d'exactitude, d'exhaustivité, de cohérence, d'actualité et de volatilité.

2. Lors de la phase d'acquisition, l'utilisateur spécifie des valeurs acceptables pour toutes les dimensions de la qualité. Si les nouvelles données répondent aux exigences de qualité, elles sont incorporées dans une archive sinon, elle seront rejetées.
3. Au stade de la publication, quand une nouvelle page remplace une ancienne page Web, la volatilité des anciennes données est évaluée. Si les anciennes données sont toujours valides, les données ne sont pas supprimées et sont associées à une nouvelle URL.

D'autres méthodologies d'évaluation pour évaluer les qualités spécifiques des sites Web sont proposées dans Atzeni et al. [AMS01], elle porte spécifiquement sur ses accessibilités et ses évaluations sur une base mixte quantitative / qualitative.

— L'évaluation quantitative de l'activité vérifie les directives fournies par le World Wide Web Consortium dans [Rag97].
— L'évaluation qualitative est fondée sur des expériences réalisées avec des experts.

Par ailleurs, Fraternali et al. [Fra+04] se concentrent sur l'accessibilité des données en proposant une approche fondée sur l'adoption de logs conceptuels, appelé des journaux d'utilisation du Web qui enrichissent les métadonnées déduites du schéma conceptuel prédéfini.

Les dimensions de la qualité liées aux systèmes exploitant les données web restent un problème ouvert, vue qu'ils ne sont pas encore définis dans une norme globale. Une possible solution consiste à s'appuyer sur la réputation (ou la fiabilité) de chaque source fournissant les données. En déterminant des valeurs de confiance envers les données et les sources fournissant l'information des questionnements recherche auront lieu. Nous en discutons certains dans la section suivante.

3.3 Véracité des données

Aujourd'hui le volume des données augmente constamment, par conséquent les problèmes de qualité de l'information sont devenus plus importants [LR14a].

Ainsi, les données peuvent avoir différentes caractéristiques qui affectent leurs exactitudes, par exemple en ce qui concerne les données massives récoltées du web par différents capteurs, ces dernières souffrent des problèmes d'imperfection notamment l'imprécision et l'incomplétude. Ces problèmes doivent être identifiés et pris en compte afin d'améliorer la compréhension des informations déduites. Généralement on parle de véracité des données.

La véracité des données complète les dimensions de la qualité des données. Lukoianova et al. dans [LR14b] fournit une feuille de route pour les définitions théoriques et empiriques de la véracité des données.

La véracité peut être définie comme suit (définition 2) :

Definition 2. *La véracité est liée à la présence de données incertaines ou imprécises qui sont dues à des erreurs, à des données manquantes ou invalides. Ces problèmes peuvent mettre en cause l'utilité des valeurs collectées [CSV18].*

Rieh et al. [Rie10] proposent une variété de méthodes d'évaluation de la crédibilité des systèmes de listes de contrôle. Leur méthodes vérifient la crédibilité et l'état des fournisseurs d'informations. Rubin et Liddy [RL06] définissent un cadre d'évaluation de cette crédibilité en concevant 25 indicateurs repris en 4 catégories : Identité de la source, fiabilité, qualité, déclencheurs. Weerkamp et de Rijke [WR12] ont étudié les indicateurs proposés par Rubin et Liddy et les ont intégrés à leurs approches de vérification de validité des informations. Ils ont montré que la combinaison d'indicateurs de véracité améliore considérablement l'efficacité de la compréhension d'un système d'information.

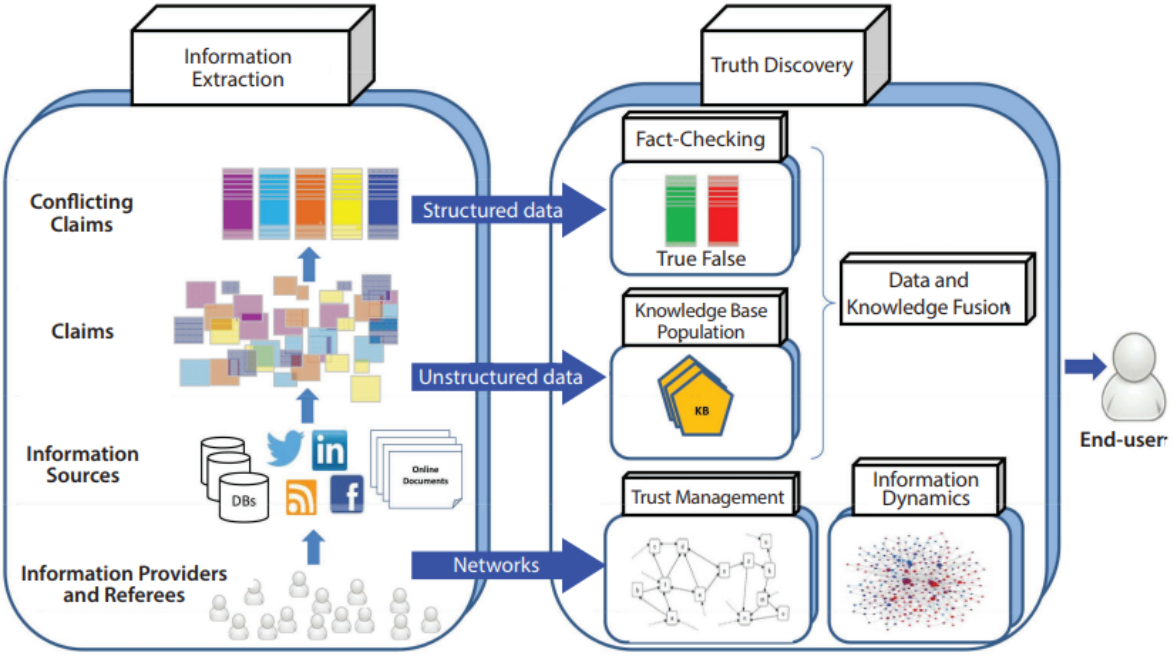


FIGURE 3.3 – Pipeline de découverte de la vérité. [BB15a]

Différents domaines de recherche ont contribué au domaine de la véracité des données : Extraction d'informations avec liaison d'entités, Vérification des faits et fusion de données structurées, Calcul de la confiance et gestion dans les systèmes de recommandation et de réputation, Étude du dynamisme des informations.

L'étude de la véracité des informations générées est liée à l'étude de la qualité. Nous la considérons comme une étape complémentaire mais aussi nécessaire pour valider les résultats à communiquer. En effet, selon la nature des données et ses caractéristiques, ces dernières doivent être traitées par des processus de traitement de qualité avant l'étude de leur véracité. Cependant, le recours à l'utilisation des formules et équations donnant des scores d'évaluation, comme utilisé dans [Rie10], n'est pas toujours possible car ceci

dépend du contexte de l'étude associée, e.g ce n'est pas possible de vérifier la crédibilité des sources extérieures alimentant une base de données interne. Pour cette raison et d'autres, Lukoianova et al. [LR14b] trouvent que le jugement sur les résultats en terme de croyabilité et d'authenticité doit suivre certains critères ou dimensions.

3.3.1 Concepts et définitions

Tout comme les dimensions définissant la qualité des données, la véracité admet aussi des dimensions particulières, Lukainova et al. [LR14b] proposent les dimensions suivantes : pour extraire une information sur les données manipulées :

- Objectivité / subjectivité : Cette dimension repose sur la vérification des informations venant des sources [MGD05], elle fait référence à la compréhension des informations sources et au jugement de leurs véracités. Par exemple, les nouvelles venant des sources officielles d'information pourraient avoir des préjugés explicites, alors que les nouvelles reprisent par des intermédiaires auront des jugements subjectifs.
- Vérité : La vérité réfère à l'écartement des informations intentionnelles, par exemple : une fausse conclusion sur un sujet [BB96]. Souvent cette dimension est liée à la capacité de pouvoir décortiquer les messages dans un système d'information afin d'aboutir à un niveau de confiance dans le contenu manipulé.
- Crédibilité : La crédibilité est la vérification de la véracité des informations issues d'un communicateur (un agent de communication) [FT99]. Elle vise par exemple à juger des informations prononcées par une entité dont les affirmations sont souvent mises en cause. En d'autres termes elle juge la qualité perçue de la fiabilité de l'expertise.

Ces fins doivent être vérifiées en gérant leur crédibilité et en garantissant la perfection des données. L'aspect conceptuel de ces trois dimensions, selon [LR14b], permet de déterminer un certain degré de véracité quand il s'agit des phénomènes touchant aux données volumineuses. Ces trois dimensions permettent la réduction des mauvaises interprétations.

Enfin, il est préférable avant tout contrôle d'information – e.g. contrôle d'authenticité, contrôle de validité, contrôle de croyabilité, etc. – de gérer l'incertitude des données vis à vis des utilisations dans un processus critique.

3.3.2 Principales causes affectant la véracité des données

Plusieurs problèmes dans les données peuvent affecter la véracité des informations. La découverte de la vérité dans un processus de traitement peut être affectée par diverses causes.

Voici quelques raisons limitant la pertinence des données :

— Des limites dans la collecte des données : les équipements techniques, par exemple les capteurs, peuvent avoir des mesures d'erreurs produisant des données incertaines et incomplètes [BB15b ; Ber15].
— Des limites lors de l'extraction de l'information : les techniques d'extraction de faits, d'événements, d'entités et de relations à partir des données non structurées ou peu structurées peuvent avoir une précision relativement faible. Ce qui rend la procédure d'établir un plan de qualité difficile au vue de la masse des paramètres à prendre en compte. C'est aussi le cas dans l'extraction des données publicitaires à partir du web : ces dernières sont fortement dynamiques et ne sont pas visualisées par tout le monde de la même façon, car elle dépendent généralement du comportement de navigation des utilisateurs.
— Des limites liées à l'intégration des données : l'intégration des données provenant des sources hétérogènes peut entraîner des incohérences au niveau de l'unification de l'enregistrement. Ceci peut être dû aux techniques d'agrégation, d'interpolation et d'extrapolation lors de la transformation des données brutes à des enregistrements unifiés.
— Ambiguïté et incertitude : la sémantique des données peut être différente et incohérente d'une source d'information à une autre, menant parfois à une interprétation erronée. Ceci reflète la mauvaise interprétation des résultats causée dans la plupart des cas par des problèmes de manque de connaissances envers les sources [DGH10].

Tout comme l'analyse de la qualité des données, dans la littérature on trouve aussi des méthodes et approches traitant de la véracité des données.

3.3.3 Principales approches

Les approches actuelles d'évaluation de la véracité peuvent être classées en 2 catégories : (1) les approches fondées sur le contenu et (2) les approches fondées sur des recommandations.

Approches fondées sur le contenu

Elles sont principalement représentées par des méthodes de vérification des faits et de fusion de données visant à calculer de manière itérative et/ou mettre à jour un score de fiabilité d'une source. Par exemple, Zhang et al. [ZHC07] utilisent une fonction de croyance sur les données afin de distinguer celles qui ont une forte ou faible croyabilité.

Par ailleurs, selon Berti-Equille et al. [BB15a], le principe est de concevoir des méthodes capables de juger les données sources. Deux manières de jugements complémentaires sont présentées pour juger, par exemple, la fiabilité d'un résultat venant d'une source.

La première est de déterminer un jugement issu du même environnement, i.e. un environnement mettant en œuvre toutes les variables et les attributs en commun entre les sources du même domaine. La deuxième manière est de trouver un jugement extérieure à l'environnement.

Prenons l'exemple suivant, pour une question donnée : Quel est le président de la république française en 2016 ? Cette question est posée à plusieurs entités (e.g la presse) construisant par la suite des sources de données d'un système d'information. Une valeur de fiabilité d'un résultat (Réponse) d'une source, dans ce cas, peut être tirée en le comparant par rapport au reste des résultats des autres sources. Plusieurs définitions d'un score de fiabilité pourront avoir lieu suite à la confiance envers les résultats reçus : e.g Calcul de probabilité, Etablissement des croyances envers les sources, etc.

La deuxième manière qui complète la précédente est d'avoir des réponses même incertaines sur les données venant de sources extérieures, e.g. non venant de la presse.

Il suffit ensuite de combiner les deux, par exemple, par des poids pour trouver un score unique caractérisant la croyance envers le résultat d'une source.

		S_1	S_2	S_3	S_4	Ground Truth	Conflicts
d_1	*USA*	Obama	-	Clinton	-	Obama	2
d_2	*Russia*	Putin	-	Medvedev	Yeltsin	Putin	3
d_3	*France*	Hollande	Sarkozy	-	Hollande	Hollande	2
d_4	*India*	Mukherjee	-	Patil	Patil	Mukherjee	2
Source Coverage		1	.25	.75	.75		

FIGURE 3.4 – Exemple d'informations issues de plusieurs sources concernant les responsables des différents pays en 2016 [BB15a].

La figure 3.4 met en évidence diverses constatations sur les données issues des différentes sources. Sur ces résultats on peut, par exemple, appliquer un vote majoritaire pour trouver les vraies valeurs. Cela peut toujours induire des conflits dans les conclusions. Par exemple, pour le président des USA, nous avons deux sources donnant chacune une réponse différente. Dans le cadre du vote majoritaire, aucune solution n'est meilleure que l'autre. Dans le cadre du vote majoritaire pondéré par la confiance dans les sources, la réponse sera Obama. Pour la France, que ce soit du vote majoritaire ou du vote majoritaire pondéré, c'est Hollande qui sera choisi. Cependant, pour l'Inde, avec ces deux méthodes, la réponse sera Patil ce qui est une réponse erronée pour l'année 2016. Ainsi, selon les opérateurs de fusion choisis, des réponses différentes et possiblement erronées peuvent être données. Cela montre l'importance des méthodes utilisées et aussi la dépendance à ces méthodes dans les approches de détermination de la véracité fondées sur le contenu.

Zhi et al. [Zhi+15] affirment que la compréhension de la véracité de certaines données dépendent de nos connaissances de base. Ainsi, la détermination des résultats aura une véracité élevée quand on ajoute des contraintes aux valeurs. Par exemple si on considère

que l'Espagne et le Royaume-Uni n'ont pas de président mais un premier-ministre ceci peut réduire le conflit probable dans le calcul des scores de fiabilité et par conséquent construire de vrais résultats.

Approches fondées sur des recommandations

Les approches fondées sur les recommandations reposent principalement sur la notion de réputation des sources. Généralement, elles jugent ces réputations par des méthodes de calcul qui reposent sur les valeurs des données générées. Elle peuvent également se reposer sur la popularité des sources et des mécanismes qui déterminent la consistance de l'enregistrement temporelle des valeurs. Dans ces approches, plusieurs modèles ont été proposés pour incorporer ces aspects [RHJ04].

Les méthodes les plus utilisées sont les approches bayésienne. Ces approches fournissent une base théorique pour mesurer l'incertitude afin de prendre une décision. Évidement, ces approches utilisent une base d'observations pour déterminer les jugements.

Pour évaluer la confiance d'une source par exemple, il suffit de connaître la distribution de données de chacun de ses capteurs et de déterminer un niveau de satisfaction envers les valeurs générées. Ces niveaux de confiance peuvent être binaires (satisfait, non satisfait), ou à plusieurs valeurs [GBS08].

3.4 Conclusion

La découverte de la vérité et ainsi l'évaluation de la fiabilité des faits d'un système se confrontent souvent à des défis techniques et logiques e.g adéquation des algorithmes, affectation des scores. La plupart des approches de mesures de la véracité s'alignent sur les caractéristiques des données manipulées en les affectant via une approche particulière, et ce, en prenant en compte des jugements externes. La quantité des jugements externes influe par conséquent sur la constatation finale. Toutefois, il est difficile dans certain cas comme "la récolte temporelle des données web publicitaires à fort potentiel dynamique" d'avoir des connaissances précises car cela dépend d'autres facteurs possiblement inconnus.

Chapitre 4

Imperfection : Incertitude, Imprécision, Incomplétude

4.1 Introduction

Comme nous l'avons vu précédemment, nos données sont imparfaites. Aussi, dans l'objectif d'étudier l'imperfection dans nos données, ce chapitre présente différents travaux de la littérature autour des différentes formes d'imperfection et des théories permettant leur représentation et leur gestion. Ce chapitre présente le support théorique dans lequel nos contributions s'inscrivent, plus particulièrement, celles présentées dans la partie suivante (Partie III). Nos choix de représentation et de modélisation des données imparfaites, effectués dans les parties suivantes, sont guidés par les concepts et théories introduits dans ce chapitre.

Ce chapitre s'articule de la manière suivante. Nous commençons par présenter la typologie de l'imperfection selon Bouchon-Meunier [Bou95]. À partir de cette typologie, nous indiquons en quoi nos données sont principalement imprécises et incomplètes. Aussi, la suite du chapitre introduit les théories et/ou approches permettant la gestion de données imprécises d'une part, et incomplètes d'autre part.

4.2 Typologie de l'imperfection

Les systèmes d'information fondés sur la récolte des données temporelles se confrontent souvent à la nécessité de la gestion de grands volumes d'information. De manière générale, le recueil peut être soumis à différentes formes d'imperfections pouvant en cas de non prise en considération de celles-ci jeter un doute sur la validité et la véracité des résultats. Aussi dans le cadre de l'étude de la qualité des données, il est important de considérer leurs imperfections car la validation des données impacte les résultats en sortie et donc les décisions finales.

Il est donc essentiel que la qualité des données temporelles soit étudiée et intégrée au processus d'analyse en fonction de la nature de l'information disponible.

Dans ce cadre, la compréhension des différentes formes d'imperfection est fondamentale [Run08 ; DDD19]. Ce chapitre examine la nature conceptuelle des différentes formes d'imperfection pouvant affectées les données acquises.

4.2.1 Concepts et définitions

À l'instar des concepts liés à la qualité des données, il n'existe pas de consensus dans l'ensemble de la littérature sur les notions et concepts liés à l'imperfection des données [DDD19]. Dans ce mémoire, nous nous fonderons principalement sur les travaux de Smets [Sme98], Bouchon-Meunier [Bou93 ; Bou95] en informatique, et ceux de Fisher [Fis05] et de De Runz [Run08] pour leur transcription dans les problèmes de gestion de données spatiales et temporelles, bien que leur cadre soit les données géographiques.

Selon [Fis05], d'un point de vue utilisateur de données, en étudiant celles-ci, nous pouvons observer si les classes des données ou les données manipulées sont bien ou mal définies. Dans le cas où la donnée et la classe sont bien définies, on sera soumis à de l'incertitude. Dans les autres cas où la donnée ou la classe sont mal définies l'imperfection des données sera due à l'imprécision, à l'ambiguïté et/ou à l'incomplétude [Run08].

D'un point de vue plus abstrait et informatique, selon Bouchon-Meunier dans [Bou93], « les connaissances disponibles sur une situation quelconque sont généralement imparfaites, soit parce qu'un doute peut être émis sur leur validité, elles sont alors **incertaines**, soit parce que nous éprouvons une difficulté à les exprimer clairement, elles sont alors **imprécises** [...] ainsi, le monde réel apparaît-il à la fois imprécis et incertain ». De plus, « les observations que nous recueillons peuvent être incertaines, mais également approximatives ou vagues ». Dans [Bou95], l'auteur distingue un troisième facteur d'imperfection des données : **l'incomplétude**.

La figure 4.1 illustre les facteurs de l'imperfection selon Bouchon Meunier.

Nous nous positionnons dans ce mémoire dans ce contexte sémantique. Aussi, nous suivons la terminologie de Bouchon-Meunier sur les trois formes d'imperfection :

— L'incertitude est liée à la validité d'une connaissance et le doute sur sa validité peut avoir plusieurs origines comme la fiabilité de la source, ou les erreurs. La gestion de l'incertitude tend à représenter la connaissance d'un agent sur la validité d'une donnée. Dans notre contexte, l'incertitude serait de savoir si nous récoltons pour un *timestamp* des données ou non. Gérer l'incertitude revient souvent à calculer des probabilités (chances d'apparition de l'événement). Ainsi, la théorie des probabilités est la théorie généralement utilisée pour représenter l'incertain [Sme98].
— L'imprécision est due au caractère vague ou approximatif de la sémantique utilisée. Le premier cas est dû à l'utilisation de connaissances subjectives. Le second cas est la conséquence de catégories aux limites mal définies. L'imprécision exprime

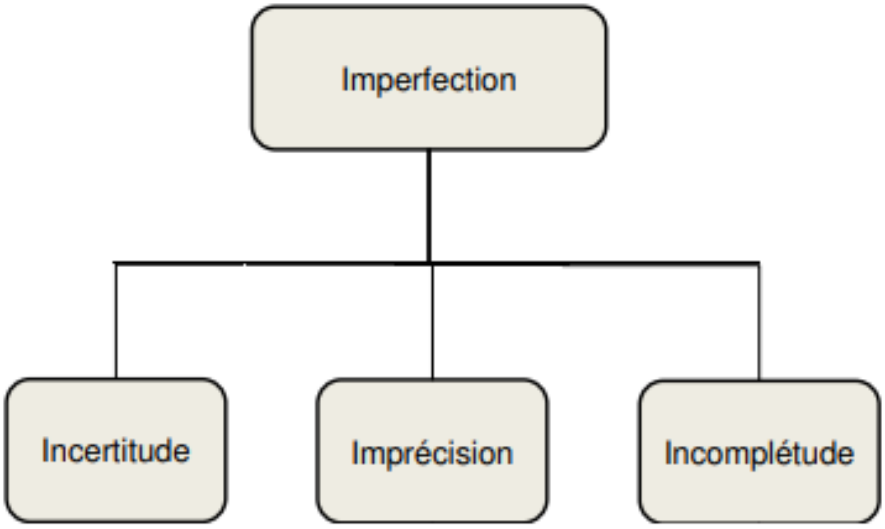


FIGURE 4.1 – Typologie de l'imperfection des données et connaissances selon [Bou95]

.

le manque de précision d'une mesure, d'un mot, d'une phrase ou d'un énoncé. Dans notre contexte, les données récoltées ne fournissent que des estimations des affichages réels des publicités sur les différents sites (cf. chapitre 2).

— L'incomplétude peut être issue d'une absence de connaissances ou d'une connaissance partielle sur certaines informations du système. Une donnée incomplète est une donnée pour laquelle la valeur de certains attributs est inconnue, ces valeurs sont dites manquantes. Ainsi dire que l'information générée ou acquise est incomplète représente le fait qu'une ou plusieurs valeurs d'attributs d'une donnée sont manquantes ou inconnues.
Dans notre contexte, nous avons à gérer des flux de données que nous traitons sous la forme de séries temporelles de données. Dans ce cadre, un attribut fait référence à un *timestamp*. Le fait de ne pas avoir de données récoltées pour un *timestamp* t, impliquera la présence d'une valeur manquante (valant NULL dans la base), et cela quel que soit le motif de la non récolte (bug, récolte impossible ou non présence de publicités). Comme indiqué dans le chapitre 2, nos données sont récoltées de manières irrégulières impliquant l'absence de données récoltées sur certains sites que l'intervalle de temps considéré soit journalier ou mensuel. Par ailleurs, nous ignorons aussi quelles sont les stratégies d'affichage utilisées par les diffuseurs. Aussi, les données récoltées ne forment que des informations partielles des véritables affichages. Les données sont donc intrinsèquement incomplètes.

La prochaine section introduit les concepts théoriques et formels pour la gestion et la représentation des informations imparfaites et plus particulèrement imprécises et incomplètes.

4.3 Théories de l'imperfection

La formalisation de l'imperfection a fait l'objet de différents travaux et propositions d'approches. L'ensemble des théories suivantes permettent de formaliser l'imperfection des données : la théorie des probabilités, la théorie des possibilités [Zad78], la théorie des ensembles flous [Zad65], la théorie des fonctions de croyance [Dem68 ; Sha76], etc. Elles se regroupent sous la terminologie théories de l'imparfait [Run08] [1]. Ces théories proposent des représentations formelles prenant en compte les imperfections dans les données. Chaque théorie permet de représenter un ou plusieurs types d'imperfection. Ainsi, par exemple la théorie des probabilités permet de gérer l'incertitude. Les probabilités imprécises incluant les fonctions de croyances et les possibilités/nécessités, permettent quant à elles de manipuler des informations à la fois imprécises et incertaines. Elles sont particulièrement utiles pour la fusion de données issues de sources contradictoires [ZMP18 ; Mar19].

Cependant, comme l'objectif de ce chapitre est de présenter le socle théorique de nos approches, nous nous concentrons dans les sections suivantes sur la représentation de l'imprécis à l'aide des ensembles et des ensembles flous, et sur la gestion de l'incomplétude.

4.3.1 Théories permettant la gestion de l'imprécis

La plupart des systèmes orientés données et traitant une problématique précise reposent généralement sur des hypothèses de précision et d'exhaustivité dans les données manipulées.

En réalité ces hypothèses peuvent ouvrir des questionnements sur la qualité des résultats des traitements. En effet, dans notre contexte les données viennent des différentes sources/capteurs et où les valeurs ne reflètent pas totalement la réalité. Par exemple, nos capteurs peuvent détecter 50 bannières affichées, mais cette valeur est plus une approximation : la valeur réelle est possiblement différente. Ceci peut remettre en question les informations générées suite à des traitements sur les données brutes.

Pour illustrer la notion d'imprécision, considérons la catégorie des signaux définie selon la relation "être stable". Sur une échelle de 0 à 1 (de la stabilité à l'instabilité absolue respectivement) mesurée par un indicateur de stabilité particulier, on peut dire qu'un signal est stable si la stabilité est égale à 0. Est-il instable si la stabilité est égal à 0.01 ou 0.02 ? Selon l'usage normal, la réponse sera non. Ainsi, le passage d'une mesure à une autre supérieure ne semble pas changer le jugement c'est à dire un signal stable à un signal instable (et réciproquement). Cependant, est-il pensable de dire qu'un signal de 0.80 est stable ? Évidemment non. Pour autant, peut-on définir un seuil différenciant les jugements pour lesquels un signal est dit stable ou instable ? Si oui laquelle ? Est-ce

1. En anglais le terme générique utilisé est Theory of Uncertainty

0.5 ? 0.4 ? Ou encore 0.3 ? Définir une valeur précise et admise par tous semble difficile. La catégorie regroupant les capteurs stable ne peut donc pas être définie précisément. L'information associée à cette catégorie est donc imprécise.

Dans les affirmations ci-dessus, la valeur de l'indicateur de stabilité, ne peut pas spécifier la catégorie d'un signal. Cependant, les systèmes fondés sur la manipulation des données et qui adoptent des modèles du monde réel se confrontent souvent à des problèmes majeurs de précision et d'exhaustivité. Tous modèles, toutes données ne forment par nature que des abstractions de la réalité. Pour qu'un système soit réaliste, il faut donc que les données récoltées soient « proches » de la réalité et donc de la vérité.

Plusieurs approches dans la littérature traitent l'imprécision dans les données. Dans cette section, nous présenterons la représentation de l'imprécis par la théorie des ensembles puis par les ensembles flous.

Théorie des ensembles

La théorie des ensembles commence par une relation binaire fondamentale entre un objet o et un ensemble A dans un espace Ω. Si o est un membre (ou un élément) de A, la notation $o \in A$ est utilisée.

Une relation binaire dérivée entre deux ensembles est la relation de sous-ensembles, également appelée inclusion. Si tous les membres de l'ensemble A sont également membres de l'ensemble B, alors A est un sous-ensemble de B, noté $A \subseteq B$. Par exemple, $\{1, 2\}$ est un sous-ensemble de $\{1, 2, 3\}$, de même que $\{2\}$ mais $\{1, 4\}$ ne l'est pas.

Un ensemble est un sous-ensemble de lui-même (pour tout ensemble A, $A \subseteq A$). Pour les cas où cette possibilité est inappropriée ou aurait du sens d'être rejetée, le terme sous-ensemble approprié est défini.

A est appelé un sous-ensemble approprié de B, noté $A \subset B$, si et seulement si A est un sous-ensemble de B et que $A \neq B$. Notez également que les éléments 1, 2 et 3 sont des membres de l'ensemble $\{1, 2, 3\}$ mais n'en sont pas des sous-ensembles. Par contre, les ensembles $\{1\}$, $\{2\}$ et $\{3\}$ en sont des sous-ensembles.

Tout comme la définition des relations arithmétiques dans d'autres théories, e.g. sur les opérations binaires entre les nombres, la théorie des ensembles utilise des relations particulières [Qui69]. Voici quelques propriétés de cette théorie :

- **Le complément** de A, noté $\bar{A}$, représente tous les éléments de Ω qui ne sont pas dans A
- **L'union** des ensembles A et B, notée $A \cup B$, est l'ensemble regroupant tous les objets membres de A, et ceux membres de B (en supprimant les doublons). L'union de $\{1, 2, 3\}$ et $\{2, 3, 4\}$ est l'ensemble $\{1, 2, 3, 4\}$. (voir figure 4.2 (b))
- **L'intersection** des ensembles A et B, notée $A \cap B$, est l'ensemble de tous les objets à la fois membres de A et membres de B. L'intersection de $\{1, 2, 3\}$ et $\{2, 3, 4\}$ est l'ensemble $\{2, 3\}$. (voir figure 4.2 (c))

— **La différence** de A et B, notée $A \backslash B$, est l'ensemble de tous les membres de A qui ne sont pas membres de B. $\{1,2,3\} \backslash \{2,3,4\}$ vaut $\{1\}$, tandis que, inversement, $\{2,3,4\} \backslash \{1,2,3\}$ vaut $\{4\}$. (voir figure 4.2 (d))

À ces propriétés, s'ajoutent d'autres définitions comme par exemple, la définition de l'ensemble vide qui est un ensemble qui ne contient aucun objet. Il est noté par une paire d'accolades avec rien à l'intérieur {} ou en utilisant le symbole $\emptyset$.

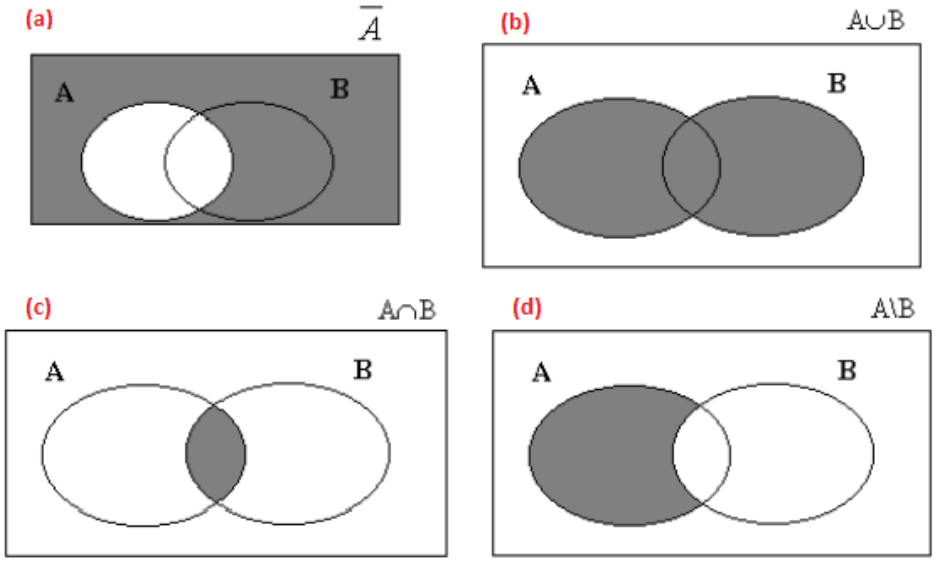


FIGURE 4.2 – (a) le complément de A ,(b) A union B ,(c) A intersection B ,(d) différence entre A et B

La théorie des ensembles élémentaires peut être étudiée d'une manière informelle et intuitive. L'approche intuitive suppose tacitement qu'un ensemble peut être formé à partir de la classe de tous les objets satisfaisant une condition ou définition particulière.

Un ensemble vu, comme une disjonction, peut être utilisé pour représenter une donnée avec son imprécision. En effet, l'imprécision peut avoir pour cause le fait que l'information disponible n'est pas assez spécifique. On peut donc supposer que la donnée imprécise peut être inscrite dans un ensemble ou intervalle l'encadrant. Soit ϕ_i une donnée imprécise définie sur $\mathcal{D}$. Il existe un sous ensemble $d \in \mathcal{D}$ borné par d_{min} et d_{max} incluant l'ensemble des valeurs possibles pour ϕ_i. Ainsi, représenter ϕ_i via d, permet de borner les possibles valeurs (disjonction) de ϕ_i et donc de représenter ϕ_i et son imprécision. La représentation via un ensemble, pouvant être un intervalle si une relation d'ordre est définissable sur le domaine, est donc une approche possible de gestion de l'imprécis.

La théorie des ensembles a donnée naissance à d'autres théories qui abordent divers limitations dans les objets manipulés et dépendent des différents contextes d'étude. La théorie des ensembles flous, objet de la section suivante, est une généralisation de la théorie des ensembles.

Traitement par la théorie des ensembles flous

La théorie des ensembles flous généralise la théories des ensembles en ajoutant une graduation de l'appartenance d'un élément à un ensemble.

Classiquement, les ensembles sont binaires dans le sens où un élément appartient à un ensemble ou est exclu de celui-ci.

Zadeh [Zad65] introduit cette idée d'ensemble flou dans la théorie des ensembles flous (appelée *fuzzy sets theory* en anglais). L'appartenance d'un élément à un ensemble flou n'est plus uniquement binaire (1 si l'élément appartient à l'ensemble ou 0 s'il est exclus) mais peut prendre une valeur quelconque dans $[0, 1]$. Cette valeur est appelée degré d'appartenance. La fonction qui associe un élément à un ensemble avec un degré d'appartenance est appelée fonction d'appartenance. L'idée est que nous ne sommes pas forcément en capacité de définir de manière précise si un élément appartient pleinement ou non à un ensemble. Par contre, on peut évaluer un certain degré d'appartenance.

La "fuzzification" consiste à transformer une valeur non floue en valeur floue à l'aide d'une fonction d'appartenance à l'ensemble visé. Soit un ensemble flou A, la fonction d'appartenance le caractérisant est usuellement notée μ_A, et le degré d'appartenance d'un élément x appartenant au domaine du discours sera noté $\mu_A(x)$. Cette fonction généralise la fonction classique de l'indicateur $I_A(x)$ de l'ensemble :

$$\begin{aligned} &I_A(x) = I - A(x) = 1 \text{ si } x \in A \\ &I_A(x) = 0 \text{ si } x \notin A \end{aligned} \tag{4.1}$$

Attention $\mu_A(x) = y$ ne signifie pas que l'on a $y\%$ de chance que x appartienne à A mais que x appartient à $y\%$ à A. Prenons l'exemple précédent de l'indicateur de stabilité. Si x a pour degré d'appartenance à l'ensemble Stable 0.8, il est globalement stable (membre de l'ensemble stable) mais pas tout à fait ce qui diffère du cadre probabiliste. En effet, si l'on se situait dans la théorie des probabilités, cela aurait voulu dire que pour la valeur de x, si nous faisons 10 tirages, alors x sera stable 8 fois sur 10, et instable 2 fois sur 10. Ainsi, dans la théorie des ensembles flous, les prédicats sont vagues, flous, imprécis car les mots utilisés pour les définir sont eux-mêmes mal définis, vagues, flous ou imprécis.

Zadeh [Zad65] remplace l'ensemble discret $0, 1$ par l'intervalle $[0, 1]$. De ce fait, plusieurs concepts comme les nombres flous, les quantificateurs etc. ont pu voir le jour [Bou93].

Les opérateurs de la théorie des ensembles classiques comme l'union, l'intersection et la négation ont été généralisés. La solution la plus classique est d'utiliser des opérateurs dits min-max [Zad65] :

$$\begin{aligned} &\mu_{\bar{A}}(x) = 1 - \mu_A(x) \\ &\mu_{A \cup B}(x) = max(\mu_A(x), \mu_B(x)) \\ &\mu_{A \cap B}(x) = min(\mu_A(x), \mu_B(x)) \end{aligned} \tag{4.2}$$

D'autres opérateurs appartenant à la famille des normes triangulaires et aux co-normes ont été proposés dans les travaux de [DOP00 ; Yag91]. La généralisation de l'implication se révèle moins évidente selon [Sme91] et plusieurs modèles ont été définis.

La loi du tiers exclu ne s'applique pas aux ensembles flous. En effet, la valeur $\mu_{A\cap\bar{A}}(x) = min(\mu_A(x), \mu_{\bar{A}}(x))$ peut être supérieur à 0. Cela se traduit de la façon suivante : « x appartient à la fois au concept "grand" et au concept "pas grand" ». Dans cette théorie, ceci est une propriété parfaitement valide. Cela ne représente cependant pas l'ignorance, mais l'appartenance partielle aux deux concepts.

La théorie des ensembles flous généralisant le concept des ensembles, le modèle peut être utilisé partout du moment qu'on travaille avec des ensembles, aussi, le concept n'est donc pas limité à une forme particulière des données. Un ensemble flou représente par nature un concept imprécis.

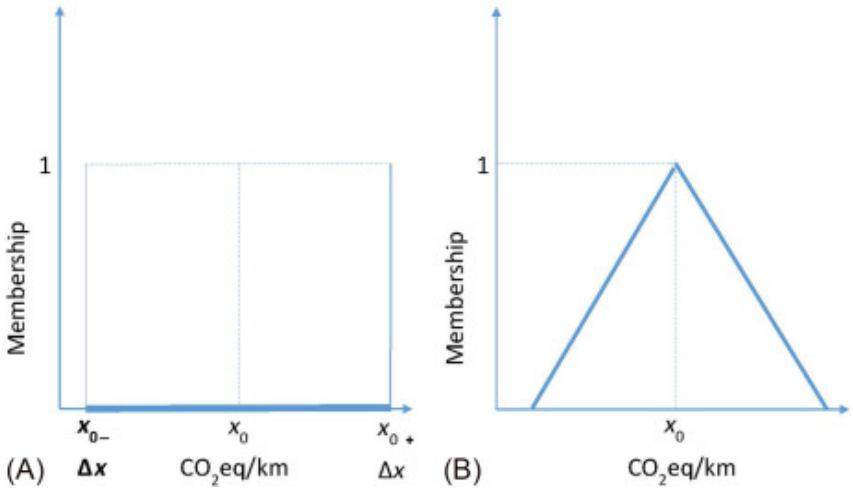


FIGURE 4.3 – Représentation de l'imprécision par un ensemble flou

Ces deux théories de représentation de l'imprécis servent de cadre de représentation de nos données imprécises et de socles de nos contributions présentées dans la partie III. Nos données étant également incomplètes, la section suivante présente des méthodes pour le traitement de ce type d'imperfection.

4.3.2 Méthodes de traitement de l'incomplet

L'intégration des connaissances sur l'aspect incomplétude, par exemple, lors de l'évaluation du comportement d'un recueil de données, est un facteur clé dans l'étude de l'information imparfaite et la gestion de la qualité des données.

Afin de traiter l'incomplètude dans les données, beaucoup de travaux évoquent l'importance des problèmes liés à la présence de valeurs manquantes [FPS96 ; HPK11 ; Pea06] et présentent différentes approches pour tenir compte de l'incomplétude des données lors d'un processus d'extraction de connaissances (cf. Framework KDD) et plus précisément lors du pré-traitement des données.

Pearson, dans [Pea06], souligne notamment les problèmes liés à la détection des valeurs manquantes qui ne doivent pas être traitées de la même façon que des attributs volontairement non renseignés. A l'inverse, dans certains cas, les valeurs inconnues, inapplicables ou encore non spécifiées sont encodées comme des valeurs valides.

Cette incomplétude peut fortement impacter la qualité des résultats obtenus à l'issue des processus de traitement quand ceux-ci dépendent en partie des données en entrée. L'étude et la gestion de l'incomplétude des données est donc importante [Fio07b]. En effet, dans les systèmes d'informations fondés sur les données, les valeurs manquantes sont souvent inévitables et peuvent avoir différentes significations.

Plusieurs propositions ont été faites pour déterminer l'incomplétude dans les données. Céline Fiot [Fio07a] cite les propositions suivantes :

1. Suppression des données comportant des valeurs manquantes ou des données incomplètes et/ou suppression d'un attribut du jeu de données si celui-ci est souvent non renseigné. Cette méthode n'est appropriée que si les valeurs manquantes sont rares, car si le pourcentage de valeurs manquantes est élevé, la perte d'information résultant de la suppression des données incomplètes n'est pas acceptable. De plus, la représentativité statistique de l'échantillon n'est pas toujours applicable.
2. Complétion des valeurs manquantes : Différentes approches sont alors possibles. Le remplissage par une valeur statistique (moyenne, médiane, etc.), difficilement applicable aux gros volumes de données, permet d'obtenir des résultats qui varient de manière importante selon l'estimation réalisée. Par ailleurs, la complétion doit être la plus proche possible de la réalité pour éviter d'introduire un biais trop important dans les données. Il s'agit, de fait, de traitement par imputation des données.

À ces approches, s'ajoutent aussi la gestion de l'ignorance. Comme la suppression des données incomplètes n'est pas envisageable dans notre problèmes, nous présentons dans la suite les approches reposant sur le traitement de l'incomplétude par imputation, puis les approches de traitement par gestion de l'ignorance.

Traitement par imputation

Plusieurs méthodes dans la littérature [SG02] reposent sur l'obtention d'une imputation convenable aux séries temporelles, e.g détermination d'une fonction spéciale, ou affectant un échantillonage sur le manque dans les données, etc. Voici quelques méthodes utilisées dans ces cas :

— Méthode de substitution : L'élément manquant est remplacé par l'utilisation d'une valeur initialement non présente dans l'échantillon, mais qui est estimée similaire à la donnée manquante.

— Méthode de la moyenne conditionnelle : La donnée manquante est remplacée par la moyenne des informations observées, i.e. la moyenne de l'échantillon. Nous obtenons donc une constante pour toutes les données manquantes. La technique est simple mais pas totalement satisfaisante selon Yoo et al. [YC07] car elle peut modifier la répartition des données réelles et donc impacter des indicateurs statistiques tels que la variance par exemple.
— Imputation multiple : La méthode d'imputation multiple arrive à résoudre la plus grande partie des problèmes des méthodes d'imputation. En fait, l'imputation avec une valeur constante ne peut jamais reproduire la variabilité typique d'un phénomène. Cette tendance conduit à une sous-estimation de la variabilité. Avec cette méthode, nous reconstruisons les données manquantes obtenues comme la synthèse de différentes valeurs échantillonnées à partir d'une distribution normale caractérisée par des indices de position et de dispersion typique aux phénomènes observés.

Traitement par représentation de l'ignorance

Dans cette démarche, nous considérons que nous n'avons pas (ignorance totale) ou peu (ignorance partielle) d'informations sur la valeur et que l'ensemble des valeurs sont possibles tout comme aucune valeur pourrait l'être. L'idée est donc de considérer l'incomplet comme une combinaison de l'incertain et de l'imprécis. À ce titre, la théorie des possibilités, celles des fonctions de croyances ou encore de manière plus générique les probabilités imprécises peuvent être utilisées en fonction des cas. Cependant, la construction des différentes fonctions nécessite des informations que potentiellement nous n'avons pas (par exemple sur la fiabilité des sources, et des capteurs). Aussi, nous ne nous positionnons pas dans ce cadre, dans ce manuscrit.

Une solution plus simple (mais plus limitée) de représentation de l'ignorance peut être d'attribuer aux cellules dont l'on ignore la valeur par une valeur en dehors du domaine du discours avec des calculs différenciés. C'est cette solution qui est en partie utilisée dans ce manuscrit.

4.3.3 Synthèse des traitements de l'incomplet

Selon le type de valeurs manquantes que l'on rencontre, leur traitement devra être adapté [Gro05 ; Pea06]. De plus, de nombreux algorithmes de fouille nécessitent ce prétraitement spécifique des données incomplètes.

Il apparaît donc nécessaire, dans un premier temps, de détecter s'il y a une raison pour que la valeur soit inconnue et si l'ignorer peut détruire une information potentiellement utile. Plusieurs approches contribuent à lever le conflit dans le cas de doute quand une discontinuité se présente. Nous citons deux principes majeurs : l'ignorance total et l'ignorance partielle.

4.4 Conclusion

Suite à notre revue de la littérature sur l'imperfection (Incertitude, Imprécision, Incomplétude), nous pouvons conclure que, suivant les données manipulées et les connaissances de leurs caractéristiques, différentes modélisations sont possibles. Aucun des modèles disponibles aujourd'hui ne peut gérer de manière satisfaisante toutes les contraintes des données imparfaites. Cependant, chaque problématique a des spécificités particulières quand il s'agit de l'imparfait, ainsi, sa résolution dépend du facteur de croyance envers les données manipulées.

Dans les contributions présentes dans cette thèse, nous avons traité l'aspect imperfection en faisant référence essentiellement à deux facteurs majeurs : L'imprécision et l'incomplétude. Nos contributions adoptent ainsi la théorie des ensembles dans une première contribution et la théorie des ensembles flous dans une seconde contribution, et ce pour gérer l'imprécision. Par ailleurs, au regard de la nature de nos flux de données, nous intégrerons dans nos approches des moyens spécifiques pour gérer l'incomplétude dans les flux de données temporelles imparfaites.

Chapitre 5

Variabilité, Stabilité

Lors de l'étude de la qualité de séries temporelles de données, la variabilité et la stabilité font partie, comme indiqué dans le chapitre 3, des possibles indicateurs à étudier. Ce chapitre est dédié à ces notions et à la manière dont elles sont généralement traitées.

De nombreuses approches d'analyse de données intègrent des aspects de vérification de l'évolution des informations captées. Parmi les aspects souvent utilisés, la variabilité et la stabilité ont pour objectif la quantification des changements dans les séries.

En statistique, étudier la variabilité revient à étudier la façon dont une série de données est dispersée ou étroitement groupée. Les indicateurs de dispersion construits pour la quantifier sont étroitement liés aux indicateurs de positions (e.g. la moyenne, la médiane) d'une série de données.

La stabilité peut caractériser la capacité d'un phénomène à être continu, à suivre une tendance, ou à avoir une faible variabilité. Contrairement au cadre classique de l'analyse des séries temporelles, nous ne cherchons pas dans ce travail à étudier la tendance ou à corriger les données, mais à en étudier la qualité. Aussi, la stabilité est considérée dans ce manuscrit selon ce prisme.

D'autres indicateurs que ceux présentés dans ce chapitre, ont aussi été proposés dans la littérature, notamment dans le cadre de données multidimensionnelles. Nous nous concentrerons dans ce mémoire sur les indicateurs associés aux séries de données à une dimension.

Nous commencerons (section 5.1) ce chapitre par un rappel sur les indicateurs de positions classiques. Ensuite, dans la section 5.2, nous présenterons certaines méthodes pour l'étude de la variabilité d'une série de données. Enfin, dans la section 5.3, nous décrirons une sélection d'approches mises en œuvre pour l'étude de la stabilité.

5.1 Rappel sur les indicateurs de positions

En statistique, un indicateur de position est un nombre réel permettant de situer l'ensemble des valeurs constituant une série statistique d'une variable quantitative. Il peut s'agir d'un indicateur de tendance centrale ou d'une valeur décentrée comme le maximum ou le minimum de la série.

Ils permettent de faire des comparaisons entre les séries de données. Ce sont des mesures liées à un problème ou à un phénomène clé dérivée d'une série de faits observés.

Notes	x1	x2	x3	x4	x5	x6	x7	x8	x9
Etudiant 1	4	5	3	6	5	7	5	4	6
Etudiant 2	9	10	2	0	1	10	2,5	9	1,5
Etudiant 3	5,5	6	5,5	6	6	0	5,5	5,5	5

TABLE 5.1 – Exemple de séries de données présentant les notes de trois étudiants notées sur 10

Indicateurs	Moyenne ($\mathcal{M}$)	Mode	min	max	Médiane (Q_2)	Q_1	Q_3
Etudiant 1	5	5	3	7	5	4	6
Etudiant 2	5	9	0	10	2,5	1,5	9
Etudiant 3	5	5,5	0	6	5,5	5,5	6

TABLE 5.2 – Valeurs des indicateurs de positions pour les données du tableau 5.1

Les indicateurs de position sont le plus souvent des moyennes (arithmétique, géométrique, quadratique...) ou des quantiles comme la médiane et les quartiles.

Nous considérons dans la suite, une série de données $X = \{x_i, i \in [1, n]\}$ de n observations. Cette série peut potentiellement être une séquence ordonnée chronologiquement (une série temporelle), i.e. x_i est apparu avant x_{i+1}.

Les différents indicateurs seront illustrés sur les trois séries de données présentés dans le tableau 5.1.

Le tableau 5.2 indiquent les valeurs des indicateurs de positions, présentées ci-après, et calculées sur les séries de notes du tableau 5.1.

La suite de la section présente certains des indicateurs de positions les plus fréquemment utilisés en commençant par les indicateurs fondés uniquement sur les valeurs et leurs occurrences, puis sur ceux fondés sur les rangs.

5.1.1 Indicateurs de positions fondés sur les valeurs

La moyenne

La moyenne est la somme des valeurs divisée par leur effectif total. En considérant la série de données X, sa définition est donnée dans l'équation suivante (éq. 5.1).

$$\mathcal{M}(X) = \frac{1}{n}(\sum_{i=1}^{n} x_i) = \frac{x_1 + x_2 + ... + x_n}{n} \tag{5.1}$$

L'utilisation d'une moyenne peut conduire à une mauvaise interprétation. Par exemple, la moyenne de chacun des étudiants dans l'exemple du tableau 5.1 est de 5 (cf. tableau 5.2). Cependant, les trois séries de notes sont très différentes les unes des autres avec notamment pour le troisième étudiant une perturbation de sa moyenne par son 0 qui est une valeur isolée très différentes des 8 autres. De plus, pour l'étudiant 2 qui a des notes qui suivent

une fonction bimodale, la moyenne de 5 ne correspond à aucune des deux modalités. En cas de distribution suivant une loi normale et avec un grand nombre de notes, la moyenne est toutefois un indicateur de position intéressant.

Afin d'avoir des indications plus fortes sur la série de données, nous pouvons aussi calculer le mode.

Le mode

Le mode est la valeur du caractère statistique dont la fréquence est la plus élevée.

Soit F une fonction de répartition, de X selon une relation F pour z, définie comme suit :

$$F(XRz) = \frac{|\{x_i Rz \| x_i \in X\}|}{|X|}. \tag{5.2}$$

Le mode est défini formellement comme suit :

$$Mode(X) = argmax_{x_i \in X}(F(X = x_i)). \tag{5.3}$$

Le mode a l'avantage de retourner une valeur présente dans les données de départ (et non une valeur construite comme la moyenne). Il a aussi l'avantage de représenter au moins une modalité. Cependant, il est très sensible à la discrétisation (pour le calcul des fréquences) et à la loi suivi par les données (en cas de loi uniforme, une valeur au hasard peut être prise). En cas de conjonction de deux phénomènes qui fournissent environ autant de valeurs l'un que l'autre, si l'amplitude des valeurs du second est plus large que celle du premier, le mode aura tendance à choisir la valeur la plus fréquente parmi celles fournies par le premier. Un exemple à cela est le cas de l'étudiant 2, où le premier phénomène fournit des valeurs dans l'ensemble $\{9, 10\}$ tandis que le second fournit des valeurs dans l'ensemble $\{0, 1, 1.5, 2, 2.5\}$. Le mode calculé est 9 (cf. tableau 5.2) avec une fréquence de 2. On peut noter que 10 aussi a une fréquence de 2. Le mode est donc dans le premier phénomène alors même que le second fournit plus de valeurs. Le mode est globalement plus robuste que la moyenne vis-à-vis des valeurs isolées.

5.1.2 Indicateurs de positions fondés sur les rangs

Pour les indicateurs de ce type, nous avons besoin d'ordonner nos données par ordre croissant. La série statistique $X = \{x_i, i \in [1, n]\}$, définie sur Ω donne lieu à la série statistique ordonnée $X = \{x_{(i)}, i \in [1, n]\}$.

La fonction *Rang* renvoie le rang d'une donnée dans son ensemble, c'est à dire l'indice de sa valeur dans la série ordonnée (cf. équation 5.4).

$$Rang : \Omega \times \Omega^n \to [1, n]$$
$$\forall x_i \in X, Rang(x_i, X) = j \iff x_i = x_{(j)} \quad (5.4)$$

À l'aide de cet ordonnancement, plusieurs indicateurs de positions peuvent être définis.

Minimum et maximum

Les minimum et maximum des valeurs d'une série temporelle sont des indicateurs décentrés de position définis comme suit :

$$min(X) = x_{(1)} \quad (5.5)$$

$$max(X) = x_{(n)} \quad (5.6)$$

Ces indicateurs peuvent être utiles pour caractériser l'étendue des valeurs prises (cf. indicateurs de dispersion - section 5.2.1), et possiblement des inclusions entre séries. Cependant, du bruit et des données isolés peuvent affecter la pertinence de ces valeurs (par exemple pour le minimum (0) pour l'étudiant 3 dans le tableau 5.2). Ces indicateurs ne donnent pas d'information sur la tendance de la série de données.

Médiane, quartiles et quantiles

L'idée générale de ces indicateurs est de déterminer les valeurs répartissant les données en un certain nombre de groupes de tailles homogènes. La médiane tend à partager la population en deux populations de taille égale : 50% des valeurs se situant en dessous d'elles, 50% au dessus. Les quartiles forment 3 valeurs tendant à partager la population en quatre groupes : 25% des valeurs de la population sont inférieures au premier quartile (appelé Q_1), 25% sont situées entre Q_1 et le deuxième quartile – la médiane – (Q_2), 25% sont situées entre Q_2 et le troisième quartile (Q_3), 25% des valeurs de la population sont supérieures à Q_3.

Nous notons r-quantiles l'ensemble des quantiles des multiples de la fraction $\frac{1}{r}$. Il y a (r-1) r-quantiles. Par exemple, il y a 3 quartiles (4-quantiles).

Le p-ième r-quantile de X est donc défini comme la valeur Q_p^r telle que :

$$F(X \leq Q_p^r) = \frac{p}{r} \quad (5.7)$$

ce qui est équivalent à :

$$F(X > Q_p^r) = 1 - \frac{p}{r} \quad (5.8)$$

C'est donc la valeur réciproque de la fraction p/r pour la fonction de répartition associée.

Les r-quantiles les plus utilisés sont la médiane (2-quantile), les quartiles (4-quantiles) et les déciles (10-quantiles).

Reprenons la série ordonnée $X = \{x_{(i)}, i \in [1, n]\}$. Soit p la proportion des éléments qui sont inférieurs au r-quantile recherché. Par exemple, $p = 1/4$ pour Q_1 et $p = 3/4$ pour Q_3 dans les quartiles.

Si $n \times p$ n'est pas un nombre entier, le r-quantile $Q^r_{r\times p}$ d'ordre $r \times p$ correspond à l'observation de rang $[n \times p] + 1$ (le premier entier supérieur à $n \times p$) :

$$Q^r_{r\times p}(X) = x_{([n\times p]+1)} \tag{5.9}$$

Si $n \times p$ est un nombre entier, deux conventions existent. Première convention :

$$Q^r_{r\times p}(X) = x_{(n\times p)} \tag{5.10}$$

Seconde convention :

$$Q^r_{r\times p}(X) = \frac{x_{(n\times p)} + x_{(n\times p)+1}}{2}. \tag{5.11}$$

Les valeurs des médianes, Q_1 et Q_3 pour les trois séries de notes d'étudiants sont données dans le tableau 5.2. Pour l'étudiant 2, on peut s'apercevoir que contrairement au mode, la médiane est dans la modalité de plus grande cardinalité. Les Q_1 et Q_3 pour ce même étudiant sont répartis sur les 2 modalités. Pour l'étudiant 3, la médiane et les Q_1 et Q_3 sont moins perturbés par la valeur isolée.

L'avantage important des quantiles est cette capacité à répartir les données en jeu de taille homogène de manière plutôt robuste. C'est pourquoi nous les utilisons, dans les parties suivantes, afin d'affecter chaque donnée à l'indice du quantile qui lui correspond. Le principe est le suivant. On range les données par ordre croissant. Le sous-ensemble des données immédiatement inférieures à chaque r-quantile ($r = 4$ pour le quartile, $r = 10$ pour les déciles) a une étendue de n/r (il contient environ $\frac{n}{r}$ données). Ces sous-ensembles, que nous assimilerons à leur r-quantiles, sont numérotés du premier au dernier (de 1 à r). Le numéro du r-quantiles auquel x appartient est défini selon l'équation 5.12.

$$\begin{aligned} Q^r &: valeur \times X \to [0, r]. \\ Q^r(x, X) &= \left\lfloor \frac{Rang(x, X)}{(n/r)} \right\rfloor + 1 \end{aligned} \tag{5.12}$$

Dispersions	Etendue	EIQ	$\mathcal{V}$	σ	$\mathcal{M}_{ecartsabsolus}$	$\mathcal{M}_{ecartscarrés}$
Etudiant 1	4	2	1,33	1,155	1,75	3,5
Etudiant 2	10	7,5	16,72	4,09	5,31	38,22
Etudiant 3	6	0,5	3,22	1,79	1,69	8,41

TABLE 5.3 – Valeurs des indicateurs de dispersion pour les séries de données du tableau 5.1

Les quantiles sont donc des indicateurs de position, et aussi de tendance, d'un échantillon donnant, en les combinant, une idée de l'étalement des valeurs tout en étant robuste au bruit, aux valeurs aberrantes ou isolées.

À partir de ces indicateurs de positions, différents indicateurs de dispersion peuvent être calculés pour quantifier la variabilité, qui est un indicateur de qualité dans notre contexte. C'est l'objet de la section suivante.

5.2 Variabilité

Plusieurs disciplines scientifiques se sont intéressées à l'identification et la quantification de la variabilité d'une série de données. En statistique, la variabilité fait référence à la nature de la dispersion des données dans le but de quantifier ses variations. Dans certaines approches de fouille de données, à l'instar de [Coe+08], la variabilité est définie comme la volatilité de la variation des mesures dans le temps. Elle est utilisée pour déterminer les événements extrêmes dans les séries chronologiques en adoptant des plages de quantiles. Par ailleurs, dans [Ded14], la variabilité est définie pour mesurer la robustesse d'un système d'information en terme de qualité des résultats. C'est pourquoi, nous nous intéressons particulièrement à sa caractérisation dans cette thèse.

Dans cette section, nous présentons, dans un premier temps, les approches statistiques traditionnelles de mesure de la variabilité par les indicateurs usuels de dispersion. Puis nous étudions d'autres approches issues de la littérature. Dans la continuité des observations faites sur nos données dans les précédents chapitres, nous introduisons ensuite des propositions de mesures introduites dans la littérature exploitant la théorie des ensembles flous. Nous expliquons ensuite notre positionnement au regard des limites de ces approches dans notre contexte.

5.2.1 Indicateurs de dispersion

En statistique, un indicateur de dispersion quantifie la variabilité des valeurs d'une série de données. Il est toujours positif et d'autant plus grand que les valeurs de la série sont étalées. Nous présentons dans la suite les plus courants. Le tableau 5.1 illustre leurs calculs sur les séries de données présentées dans le tableau 5.3.

Étendue ou amplitude

L'étendue, aussi appelée amplitude, est la mesure la plus évidente pour déterminer la dispersion. Elle représente la différence entre la valeur la plus haute et celle la plus basse dans un ensemble de données.

$$Etendue(X) = max(X) - min(X) \tag{5.13}$$

Cet indicateur de dispersion est une mesure simple à déterminer. L'amplitude est utile pour afficher la dispersion dans un jeu de données et pour comparer la dispersion entre des jeux de données similaires.

Suivant cet indicateur, il est assez évident que l'étudiant 2 a une forte variabilité car l'étendue de ses valeurs regroupe toutes les valeurs possibles. Cependant l'étudiant 3 aura une plus forte variabilité que l'étudiant 1 du fait uniquement de la valeur isolée. Dans notre contexte, au regard des grandes variations possibles dans nos récoltes, cet indicateur, sans autres prétraitements des données que ceux présentés dans le chapitre 2, n'est pas des plus utiles vu les forts écarts possibles dans les valeurs et les possibles facteurs de puissance ou de non comparabilité dans nos données. De plus, il ne prend pas en considération les imprécisions et les incomplétudes de nos données (ni leur chronologie).

Pour limiter l'impact des valeurs isolées, du bruit et des valeurs aberrantes, l'écart inter-quartile a été proposé.

Ecart inter-quantile

L'écart inter-quartile est une mesure qui indique comment les 50% des valeurs centrales sont dispersées. Il correspond à l'étendue de la série statistique après élimination de 50% des valeurs (25% les plus faibles et 25% les plus fortes). Cette mesure est plus robuste que l'étendue, qui est sensible aux valeurs extrêmes. Sa définition est fournie dans l'équation suivante (eq. 5.14).

$$EIQ(X) = Q_3(X) - Q_1(X) \tag{5.14}$$

L'écart inter-quartile fournit une image assez représentative de l'ensemble des données en ignorant les valeurs atypiques. Ainsi, si l'étendue pour l'étudiant 2 reste large du fait de la bimodalité de la série, celle de l'étudiant 3 devient la plus petite car la valeur isolée n'a pas été prise en considération.

Toutefois, comme pour la détermination des étendues, l'écart inter-quartile est une mesure de dispersion fondée uniquement sur deux valeurs. Afin d'essayer de considérer l'ensemble des valeurs, différentes mesures de dispersion ont été définies autour de la dispersion autour de la moyenne.

Dispersion autour de la moyenne

Tout d'abord, nous pouvons considérer la somme des valeurs absolues des écarts à la moyenne. Cela s'appelle l'écart moyen (eq. 5.15).

$$\mathcal{E}_{moyen} = \sum_{i=1}^{n} |x_i - \mathcal{M}(X)| \tag{5.15}$$

Cependant cet indicateur souffre de problème du fait de la non dérivabilité de la valeur absolue. Pour rendre positif les écarts, une solution peut être la mise au carré, principe utilisé par la variance.

Variance La variance mesure à quel niveau un ensemble de valeurs est écarté de sa valeur moyenne. La variance joue un rôle central en statistiques (statistiques descriptives, l'inférence statistique) et aussi dans d'autres domaines tels que la vérification d'hypothèses, la qualité de l'ajustement etc. La variance est une mesure importante en analyse des données. Sa définition est donnée dans l'équation 5.16.

$$\mathcal{V}(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mathcal{M}(X))^2 \tag{5.16}$$

Ce qui dans le cas de séries discrètes non triées, à l'instar de nos données, revient à :

$$\mathcal{V}(X) = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \mathcal{M}(X)^2 \tag{5.17}$$

L'observation des résultats du tableau 5.3 sur les exemples du tableau 5.1 montrent que contrairement à l'EIQ, la variance pour la série de l'étudiant 3 est plus forte que celle de l'étudiant 1 (de manière assez modérée), tandis que celle de l'étudiant 2 est beaucoup plus forte que pour les deux autres. La valeur isolée de l'étudiant 3 perturbe légèrement son calcul. Cependant, il ne faut pas oublier le rapport de puissance (le carré pour la variance) pour l'interprétation de ces résultats.

Écart type L'unité de la variance est le carré de celle de la variable étudiée (si cette dernière est en m, sa moyenne est en m mais sa variance est en m^2) d'où l'impossibilité d'additionner la moyenne et la variance. Pour résoudre ce problème et pour pouvoir combiner la dispersion avec la position (pour avoir une position contextualisée par exemple, ou des intervalles de confiance), l'écart type noté σ a été introduit. L'écart type est la racine carré de la variance (son unité est donc la même que celle de la moyenne).

$$\sigma(X) = \sqrt{\mathcal{V}(X)} \tag{5.18}$$

Lorsque les valeurs d'un jeu de données sont assez regroupées, l'écart-type est faible. Quand les valeurs sont écartées l'écart type sera relativement grand. L'écart type est généralement présenté en conjonction avec la moyenne.

Reprenons nos 3 étudiants. L'écart type des notes de l'étudiant 2 est plus de trois fois supérieur à celui de l'étudiant 1, et vaut plus du double de l'étudiant 3. Sa variabilité est donc clairement plus forte. Comme pour la variance, l'écart type des notes de l'étudiant 3 est supérieur à celui de l'étudiant 1 (environ 50% plus élevé). Cependant, les faibles valeurs de l'écart type pour ces deux étudiants amènent à considérer que leur variabilité est globalement faible.

Ces mesures sont sensibles aux valeurs aberrantes et aux valeurs isolées. De plus, l'imprécision des valeurs n'est pas considérée.

Prise en considération de la temporalité

Dans l'ensemble des approches précédentes, la variabilité est comprise sur l'ensemble des données sans considérer un possible ordonnancement temporel entre elles. Or c'est le cas dans les séries temporelles que forment nos données d'étude.

Dans ce contexte, nous pouvons considérer les écarts non plus à la moyenne mais deux à deux, en prenant des couples d'éléments qui se suivent. On peut dans ce cadre aussi calculer la moyenne des écarts absolus deux à deux et la moyenne des écarts au carré. Si on a n données, on a $n-1$ écarts.

$$\mathcal{M}_{ecartsabsolus} = \frac{1}{n-1}\sum_{i=1}^{n-1}|x_i - x_{i+1}| \tag{5.19}$$

$$\mathcal{M}_{ecartscarrés} = \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i - x_{i+1})^2 \tag{5.20}$$

Pour notre exemple des notes des 3 étudiants, on peut se rendre compte que ces deux mesures donnent des classements différents en terme de variabilité entre l'étudiant 1 et l'étudiant 3. Ces indicateurs sont naturellement sensibles à l'ordre des données et donc répondent à une problématique pour l'ordre chronologique. Ils sont aussi sensibles à la qualité des données, de petites variations dans les valeurs peuvent changer l'interprétation des comparaisons. Enfin, ces deux dernières approches ne permettent pas la gestion des données manquantes. Nous nous inspirons cependant de ces méthodes pour proposer dans nos contributions des approches tendant à lever leurs limites.

Plusieurs approches proposées dans la littérature visent à considérer la richesse de ces différents indicateurs, et les combinent pour mieux apprécier la variabilité des séries étudiées.

5.2.2 Autres approches

En biologie, Fishman et al. [Fis+12a] proposent d'étudier le comportement dynamique de l'ensemble des états de santé des patients pour déterminer la progression d'une maladie suite à un changement dans un état particulier, e.g. le rythme cardiaque. La variabilité est étudiée sur une période de temps déterminée. Le principe est de regrouper plusieurs jugements possibles, les jugements doivent être fortement corrélés pour avoir une idée sur l'état général. Techniquement, il faut quantifier les valeurs des données de chaque état temporel et déterminer les différences internes à l'état, puis entre les différents états. En outre, cette variabilité utilise diverses paramètres, comme, par exemple, le facteur temps permettant de réduire ou d'agrandir les périodes d'analyse, ou encore un paramètre de distance entre les données afin de déterminer différents groupements, etc.

Pour étudier le comportement des séries chronologiques avec un facteur temps, Chen et al. [Che+09] étudient la dispersion des points dans une même série temporelle sur une période donnée. La méthode proposée détermine une quantification de la variabilité à partir de l'éloignement entre les points en les associant à des groupes spécifiques (construits à partir d'autres relations) et en les comparant à la dispersion normale des valeurs.

Fishman et al. [Fis+12b] trouvent que la technique utilisée précédemment nécessite un ajustement des paramètres utilisés. Ils se fondent sur le fait que les valeurs varient temporellement et qu'il est donc difficile de déterminer des groupements réguliers de valeurs dans le temps. De plus, les valeurs aberrantes ne sont pas prises en compte dans cette méthode. Aussi, ils proposent que l'information sur les données doit être déduite à partir d'intervalles temporels et non à partir des groupements construits et ce, en découpant la période d'étude en des sous périodes pour mesurer la similarité entre elles. Ils effectuent ensuite une quantification sur les écarts déterminés.

Dans d'autres domaines comme par exemple, la surveillance des réseaux, Kang et Lee [Kan17] exploite l'imputation pour illustrer les informations temporelles. Ils comptent la variation entre les échantillons en utilisant des seuils de progression, ce qui fait l'objet du jugement sur la variabilité des flux.

Dans le domaine de l'étude du comportement climatique et de ses changements, Nishi Bhuvandas [Bhu+14] utilise une méthode statistique pour juger l'évolution de la température. Le principe est simple. Il part de la définition statistique de la variabilité qui peut être déterminée à partir des ensembles de mesures tels que l'écart type, la variance, l'aplatissement, etc. En prenant des ensembles de mesures enregistrées à chaque instant, un tableau de résumé global de ces valeurs, utilisant des indicateurs de dispersion, est ensuite généré. Un calcul de similarité entre ses lignes fait l'objet d'une étude de corrélation entre ces résumés, ce qui permet d'estimer un changement climatique futur.

Cependant, l'ensemble de ces mesures ne tiennent pas forcément compte de l'incomplétude des données ni de leur imprécision.

5.2.3 Mesures issues de la théorie des ensembles flous

Dans les approches floues, le calcul de la variabilité diffère d'une étude à l'autre. Dans [Hül14], elle est considérée comme une agrégation de la variance des mesures floues. Lubiano [Lub99] ainsi que par Couso et Dubois [CD09] la voient comme une variance scalaire de chaque élément à l'aide d'un modèle flou modélisant les connaissances imprécises.

Pour déterminer la variance scalaire, soit $\tilde{X}$ un intervalle de valeurs linguistiques. Soient λ_1, λ_2 et λ_3 des contraintes d'ajustement telles que $\lambda_1 + \lambda_2 + \lambda_3 = 1$, et X_1 et X_2 deux variables aléatoires définies sur Ω.

$\tilde{X}(x_i)$ représente le label assigné e.g $\{Grand, Petit, Moyen\}$. L'intervalle de mesure de la variabilité est $X_1(x_i) = inf(\tilde{X})$ et $X_2(x_i) = sup(\tilde{X})$, i.e. l'intervalle mettant en œuvre les variables linguistiques affectées, respectivement, aux valeurs inférieures et supérieures. la variabilité scalaire se représente comme suit :

$$SVar(\tilde{X}) = \lambda_1 \mathcal{V}(X_1) + \lambda_2 \mathcal{V}(\frac{X_1 + X_2}{2}) + \lambda_3 \mathcal{V}(X_2) \tag{5.21}$$

Cette méthode de calcul a été critiquée par Korner et al. [KN02] car elle présente divers cas particuliers. Il est compliqué de trouver une unique valeur satisfaisant $X_1(X_i) = inf\tilde{X}$ ou $X_2(\omega) = sup\tilde{X}$ et ce, au vu des relations présentes entre les ensembles flous (variables linguistiques). De ce fait, ils proposent de définir un représentant de chaque $X \in \tilde{X}$, i.e. chaque ensemble aura son représentant. La valeur du représentant est présentée par une pondération $\alpha \in [0, 1]$ donnant αX_1. Une fois que les représentants des valeurs des ensembles flous sont définis, un calcul de leur variance est donné. Cette variance représente la variabilité de l'ensemble $\tilde{X}$.

L'idée de calculer la variance scalaire à l'aide d'un point représentatif de chaque observation floue est utilisée par Baudrit et al. [Cou+07] dans le cadre des données imparfaites. Leur principe tient compte de l'incomplètude et l'aléa dans les données manipulées. Ils commencent par formaliser la propagation incertaine des données. Quant aux données manquantes, ils se fondent sur des hypothèses de dépendance entre les quantités inconnues. Ceci donne un résultat en fonction des croyances envers les données. Ce résultat est fourni par une fonction de croyance paramétrée. La fonction établit un modèle d'agrégation fondé sur les possibilités d'existence des valeurs dans des intervalles bien déterminés. La variabilité est ainsi déterminée ensuite par une évaluation des fréquences des mesures.

Cependant, Kruse et Meyer [KM87] considèrent que le degré d'appartenance de chaque nombre réel x appartenant à $\tilde{X}$ doit être considéré dans le calcul si, et seulement si, il a une signification précise et fait l'objet d'une assertion calculée.

Selon Couso et al. [CD09], les variabilités floues comme celle proposée par Baudrit et al. sont logiques si des valeurs finies caractérisant l'ensemble existent. Mais dans certains cas, le problème peut être NP-difficile. Ainsi l'agrégation, qui est à l'origine de cette variabilité, reflète uniquement la variation de l'imprécision.

D'autres propositions, venant de la logique floue, ont essayé de modéliser les variations quand il s'agit des ensemble flous. Paoli et al. [Pao+06] utilisent une fonction, appelée variogramme flou. Elle permet de pondérer les données disponibles et d'associer à ces valeurs une variance (appelée « variance de krigeage »). Ce modèle garantit selon eux une meilleure estimation des valeurs possibles en leur associant des attributs flous.

Bardossy et al. [BBK90] décrivent ces attributs à l'aide d'une variable aléatoire dont les valeurs sont des nombres flous triangulaires, ils affirment qu'une forme trapézoidale peut être aussi utilisée. La variable aléatoire est décrite par trois variogrammes (un variogramme modal, un variogramme supérieur, et un variogramme inférieur). Ces trois fonctions interviennent dans le calcul de la variance de Krigeage. La notion de variogramme flou [KM82] a été proposée pour prendre en compte les imprécisions dans la modélisation des ensembles flous.

Cependant, l'ensemble de ces mesures ne considèrent pas à minima l'une des caractéristiques de nos données : elles sont temporelles (séquentielles), incomplètes et imprécises. En nous inspirant de ces approches, nous proposons, dans les chapitres suivants, deux approches pour quantifier la variabilité des séries temporelles de données imprécises et incomplètes.

Si la variabilité quantifie les variations dans les mesures, la stabilité, potentiellement liée à la variabilité en fonction des approches, est aussi une mesure importante.

5.3 Stabilité

Felber [Fel15] considère que pour étudier la stabilité des données, il faut tout d'abord aborder des questionnements sur leur compréhension, i.e. savoir si les données ont un sens particulier, e.g. en essayant de déterminer l'orientation totale des valeurs enregistrées. Ainsi quand il s'agit des données temporelles, obtenir une idée générale sur leurs comportements est essentielle. Cette première vérification permet d'affecter les techniques adéquates à leurs analyses.

Plusieurs techniques de vérification de la stabilité peuvent être utilisées. Par exemple, quand il s'agit des données volumineuses, l'utilisation d'échantillons est une technique permettant de donner un jugement sur la stabilité en comparant les différents échantillons générés.

En statistique, la vérification de la continuation d'une distribution à suivre un certain modèle statistique ou probabiliste est aussi une méthode de vérification de la stabilité. Pour Rober et al. [AFG98], ce type de stabilité, qui calcule la différence entre deux distributions statistiques, doit tenir compte uniquement des régions d'intérêt i.e les régions centrales. Mikosh et al. [Mik+95] trouvent que pour tenir compte des extrémités i.e là où l'erreur de mesure est élevée, il faut centrer et réduire les valeurs. [FR99] discutent cette

méthode en affirmant qu'elle est souvent utilisée mais qu'elle a des conséquences quand il s'agit d'études sur des phénomènes précis. Il faut donc examiner la stabilité par situations particulières, en d'autres termes en étudiant les données qui se ressemblent le plus.

Battistelli et al. dans [BC14], décrivent la stabilité comme étant une forme de divergence dans les données. Soit $p(x)$ et $q(x)$ deux suites de distributions. Admettons en premier que $p(x)$ est une distribution réelle et $q(x)$ est une distribution théorique. Nous mesurons la perte relative, notée par D_{KL} et définie dans l'équation 5.22, entre les deux mesures.

$$D_{KL}(q(x)|p(x)) = E_p(ln(\frac{p(X)}{q(X)})) = \sum_{i=1}^{B} p(x_i) ln(\frac{p(x_i)}{q(x_i))} \tag{5.22}$$

Cette indice mesure la divergence de q(x) par rapport à p(x). Ce n'est pas une distance. Cette mesure a des limites quand les données convergent dans certaines expérimentations, e.g. à un instant t les distributions pourront avoir deux histogrammes très proches, dans ce cas D_{KL} tends vers 0.

Yurdakul dans [Yur18] propose un indice de stabilité d'une population nommé PSI. L'indice mesure la différence entre un échantillon présentant les valeurs observées (base) et un échantillon présentant les valeurs à étudier (cible).

L'indice de stabilité de la population examine la différence entre les proportions de base et cible et est calculé comme suit :

$$PSI(Base, Cible; C) = \sum_{i=1}^{n} (Base_i - Cible_i) * (ln(Base_i) - ln(Cible_i)) \tag{5.23}$$

En pratique, la valeur de l'indice peut être interprétée comme suit : si le PSI est inférieur à 10%, le modèle est approprié et si le PSI se situe entre 10% et 25%, il faut alors analyser l'échantillon actuel pour des raisons de PSI élevé. Si le PSI dépasse 25%, il est vivement conseillé de développer un nouveau modèle sur un échantillon plus récent.

La stabilité d'une série temporelle peut être aussi étudiée suite à l'application des tests de convergence. On parle de convergence lorsque la différence entre les données et/ou leur dispersion se réduit dans le temps. Il existe différentes méthodes pour étudier la convergence. Par exemple, nous pouvons étudier la corrélation entre deux séries sur un même niveau ou voir si deux séries convergent vers un point à partir d'un moment donné. Dans le cas des séries temporelles, on peut étudier la variance des points de X^T sur les périodes construisant T, *i.e.* dans des sous-périodes de T pour voir si l'écart de dispersion des points se réduit au cours du temps et jusqu'à quel point.

Si l'écart se réduit, on peut parler d'une stabilité dans la série temporelle (voir figure 5.1.a), si l'écart n'est pas réduit, on dit que X^T n'est pas stable (voir figure 5.1.b).

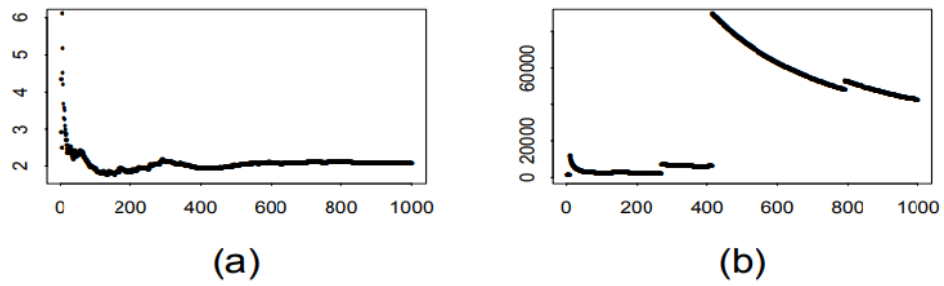


FIGURE 5.1 – Test de convergence pour l'analyse de la stabilité des données

D'autres approches existent dans la littérature. Elles reposent globalement sur les mêmes principes : soit elles comparent deux séries de données (passé/présent), soit elles utilisent des indicateurs de convergence ou de réduction de la dispersion en construisant des périodes temporelles.

Dans notre cas, l'idéal serait d'avoir des recueils homogènes au cours du temps. Nous considérons ainsi que la mesure de la stabilité quantifie une faible variabilité. Aussi, nos indicateurs de stabilité dépendent des indicateurs de variabilité.

5.4 Conclusion

Dans ce chapitre, nous avons introduit différentes approches pour quantifier la variabilité et la stabilité afin d'étudier la qualité de nos récoltes. De cette étude de la littérature, nous pouvons conclure qu'il n'y a pas, à notre connaissance, d'indicateurs de la variabilité et de la stabilité considérant l'ensemble des caractéristiques de nos données. Nous proposons dans la partie suivante, deux approches pour déterminer la variabilité de séries temporelles de données imprécises. Dans ces approches, nous proposons aussi un indicateur de stabilité construit par agrégation d'indicateurs de variabilités.

La partie suivante de ce manuscrit traite de la quantification de la variabilité et de la stabilité de la collecte des séries temporelles de données imparfaites. Dans ce but, nous proposons deux contributions. Ces contributions s'inscrivent dans un cadre qui tend à gérer toutes les contraintes de nos données.

Troisième partie

Contributions à l'analyse de la variabilité et de la stabilité des séries temporelles de données imparfaites

Chapitre 6

Contribution 1 : Approche fondée sur les quantiles

6.1 Introduction

Nous avons souligné précédemment, dans la partie II l'importance de la variabilité et de la stabilité pour l'évaluation de la qualité des données et de leur récolte. Nous avons aussi remarqué l'importance de bien gérer les caractéristiques et les imperfections des données afin de tendre à une meilleure véracité des résultats.

Les données dans ce travail se caractérisent par leur temporalité, leur facteur d'échelle et leurs imperfections (imprécisions et incomplétudes). Nous souhaitons donc être en capacité d'étudier leur variabilité et leur stabilité. Le chapitre 5 nous indique qu'il y a un manque dans la littérature pour traiter ce problème. Dans ce but, nous proposons dans ce chapitre une première approche fondée sur les quantiles, appelée QBA pour *Quantile Based Approach*, dans laquelle nous proposons un nouvel indice pour quantifier la variabilité et la stabilité de telles séries de données.

Ce chapitre est organisé comme suit. Nous commencons par présenter les objectifs (section 6.2) et les principes et hypothèses (section 6.3) guidant notre proposition. Ensuite, nous présentons dans la section 6.4, notre approche et nos nouveaux indices. Nous l'illustrons sur un exemple explicatif dans la section 6.5. Puis, dans la section 6.6, nous ferons une étude de la sensibilité [1]. Enfin, une analyse des résultats (section 6.7) et une conclusion (section 6.8) seront données.

6.2 Objectifs de l'approche

La présente approche (QBA) vise à juger la qualité de la récolte effectuée par nos capteurs, et ce, en évaluant la qualité des séries temporelles de données imparfaites manipulées. Pour cela, nous cherchons à étudier la variabilité et la stabilité qui sont deux

1. L'étude de la sensibilité consiste en l'analyse de l'influence des paramètres des indices sur les valeurs de ces indices.

aspects importants de la qualité (cf. Partie II). Nous entendons par variabilité, la volatilité d'une série temporelle au cours du temps, et, par stabilité, la faible variabilité d'une série au regard d'elle même et au regard des autres.

Dans cette contribution, nous cherchons à gérer les imperfections de nos données. Nous voulons ainsi, traiter leur imprécision et leur incomplétude.

Par ailleurs, nos données présentent des volumes très différents et sont donc difficilement comparables. Nous devons prendre en compte cet aspect dans l'analyse de nos séries temporelles.

De plus, compte tenu de la temporalité de nos données, nous cherchons à proposer une approche qui détermine les variations remarquables entre deux timestamps dans les séries temporelles, et ce, pour pouvoir détecter et analyser des phénomènes significatifs et/ou parfois cachés.

Notre contribution doit aussi tenir compte des tendances des valeurs récoltées. En effet, les données récoltées peuvent varier du fait de phénomènes internes au capteur, mais aussi du fait d'une tendance globale constatée sur l'ensemble des capteurs.

6.3 Principes et hypothèses

Afin de traiter l'imprécision de nos données, la présente approche utilise essentiellement la théorie des ensembles. L'idée est d'englober les valeurs des données dans des ensembles regroupant plusieurs données ayant un comportement similaire. Pour cela, on utilise les quantiles et les sous-ensembles qu'ils forment. Ces sous-ensembles forment des groupes de données de tailles homogènes dont les indices indiquent les positions globales. Nous considérons qu'affecter la donnée d'une série à son quantile dans la série permet donc de représenter la donnée avec son imprécision, car l'élément est considéré dans un ensemble de valeurs possibles ayant un comportement similaire vis-à-vis de l'ensemble.

De plus, les systèmes fondés sur les quantiles en tant que paramètres permettent le positionnement des données des capteurs dans le temps (voir figure 6.1). Lorsqu'on associe aux données les indices des sous-ensembles, formés par les quantiles, qui leur correspondent, on obtient une échelle de valeurs ordonnées allant de 1 à r, si r est le nombre de quantiles que l'on considère. Cette échelle permet de rendre comparable des valeurs qui l'étaient faiblement auparavant. Dans cette figure (6.1), le capteur fournit des données (entourées en rouge) qui sont successivement dans Q_4 puis dans Q_3. On s'aperçoit ainsi que malgré les variations importantes dans les valeurs des données, au regard des quartiles, il s'agit d'une perte d'un niveau. Ainsi, cette affectation des données à leur quantile permet à la fois de gérer à la fois l'imprécision et les problèmes d'échelles de nos données. Par simplification de langage, nous utiliserons les termes quartiles et quantiles à la fois pour leur valeur mais aussi pour les ensembles d'appartenance qu'ils forment.

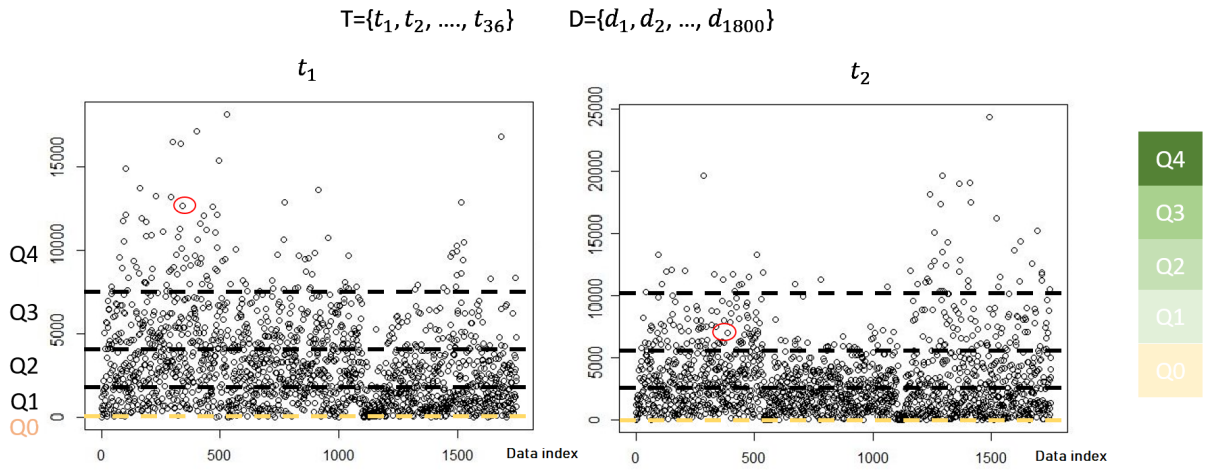


FIGURE 6.1 – Positionnement des données dans leurs intervalles, définis par les quartiles, respectifs à deux timestamps successifs t_1 et t_2.

Par ailleurs, les valeurs enregistrées par les différents capteurs souffrent d'incomplétude. Ainsi, puisque nos données sont lacunaires, nous adoptons les principes liés à la notion d'ignorance totale [DPS96] des valeurs nulles. Ces données avec des champs vides sont ensuite traitées séparément par notre approche. Pour considérer cet aspect, nous introduisons l'état Q_0 qui réfère à un état englobant des valeurs nulles. Cet état représente l'état d'ignorance sur nos données.

Dans l'approche QBA, la temporalité dans les données est traitée par la quantification des sauts significatifs sur les nouveaux positionnements des données. Pour cela, nous comparons les données successives deux à deux.

Par ailleurs, si l'ignorance dure dans le temps, cela peut vouloir dire que la récolte de ce capteur est arrêtée volontairement. En cas de reprise de la récolte après une longue durée d'interruption, il est sans doute préférable de considérer que c'est un nouveau début, et donc de ne pas considérer que c'est un saut significatif.

Enfin, nous observons l'évolution et la variabilité des données selon une vision interne et externe. La vision interne réfère à l'évolution interne de la récolte d'un capteur et la vision externe réfère à son évolution au regard des récoltes des autres capteurs. Ces deux visions permettent d'observer le comportement d'un capteur vis à vis de lui même (vision interne) et vis à vis des tendances globales (vision externe).

Sur ces principes et hypothèses, nous proposons dans les sections suivantes des indicateurs de variabilité selon la vision interne et externe. L'approche QBA propose d'unifier les deux en un indicateur de stabilité. Ce dernier donne une information synthétisant à la fois la cohérence locale et et la cohérence globale (vis-à-vis des autres) d'un ou plusieurs capteurs.

6.4 Présentation de l'approche QBA

Afin de répondre aux objectifs présentés dans la section 6.2, nous proposons une nouvelle approche, appelée QBA pour *Quantile Based Approach*, fondée sur les principes et hypothèses présentées dans la section précédente.

Notre approche est composée de trois principales étapes :

1. La représentation des données à l'aide des quantiles-set afin de pouvoir gérer les imperfections et les facteurs d'échelle. Cette représentation se fait selon une vision interne et selon une vision externe. Dans la vision interne, l'idée est de positionner la donnée récoltée par un capteur vis-à-vis des autres données récoltées par ce même capteur sur la période de temps étudiée. Dans la vision externe, l'idée est de positionner la donnée récoltée à un timestamp au regard des autres données récoltées par l'ensemble des capteurs à ce timestamp.
2. Le calcul de la variabilité, en considérant un indice spécifique que nous proposons, selon les deux différentes visions. La variabilité selon la vision interne donne une information sur la cohérence interne de la récolte du capteur. La variabilité selon la vision externe donne une information sur la cohérence du capteur vis-à-vis des tendances globales des récoltes faites par l'ensemble des capteurs.
3. Le calcul de la stabilité correspond à une agrégation des deux indices de variabilité afin de considérer la cohérence globale de la récolte d'un capteur (à la fois interne et externe).

6.4.1 Représentation à l'aide des quantiles-sets

Comme indiqué dans la section 6.3, nous prenons l'hypothèse que représenter les données à l'aide de leurs quantiles permet de gérer leurs imprécisions mais aussi de traiter les possibles facteurs d'échelle dans leur valeur.

En effet, pour la gestion des imprécisions, soit une donnée d appartenant à un quantile-set [2] Q_z d'indice de position z. Par construction, Q_z regroupe un ensemble de valeurs. Utiliser Q_z, via sa position z, au lieu de d permet de dire que les possibles valeurs de d sont dans Q_z et non uniquement d.

Pour les facteurs d'échelle, si on a r quantiles-sets, nos comparaisons se font entre des valeurs appartenant à $\{1, ..., r\}$ et non dans l'amplitude des données de la variable ou dans les rangs des données ($\{1, ..., n\}$). Bien entendu, r doit être beaucoup plus petit que le nombre de données ($r << n$).

Représenter la donnée par son quantile-set peut donc donner une information de son placement dans une série de données.

2. sous-ensemble défini par les quantiles

La série de données utilisée pour le calcul du quantile peut être l'ensemble des données récoltées par le capteur d'où elle est issue. Dans ce cas, la position obtenue correspond à une information sur la récolte du capteur sans considération des données issues des autres capteurs. Nous considérons donc que nous étudions par ce biais le comportement interne du capteur. Ce positionnement est donc issue d'une vision interne, et nous appelons les quantiles-sets obtenus les quantiles internes notés Q_{int}.

La série de données exploitées pour déterminer le quantile d'une donnée récoltée à un timestamp t peut aussi être l'ensemble des données récoltées à t. Autrement dit, la position obtenue pour notre donnée est relative aux positions des autres données récoltées en même temps. La relation au centre de ce positionnement n'est plus le capteur, mais le timestamp. Ce positionnement permet de comprendre la position d'une donnée vis-à-vis des données des autres capteurs. C'est donc une vision externe au capteur. Nous appelons les quantiles-sets obtenus par ce procédé, quantiles externes notés Q_{ext}.

La figure 6.2 donne une perception typique de notre nouveau placement d'une série de données issue d'un capteur, sans gestion de l'incomplétude. La figure 6.2(a) présente la chronologie des positions en quartiles-sets de la série de données calculées par rapport aux autres données récoltées par le dit capteur sur une période donnée. Cette figure donne une information sur la cohérence interne du capteur (ici beaucoup de variation interne). La figure 6.2(b) illustre la chronologie des positions en quartiles-sets de la série de données vis-à-vis des autres données récoltées aux mêmes instants. Cette représentation est complémentaire à la première en prenant en considération des mouvements globaux possibles dans l'ensemble du processus de récoltes.

Vu que l'absence de données collectées ne signifie généralement pas l'absence, par exemple, d'annonces sur un site Web, sa valeur quantitative dans l'étape d'analyse ne doit pas être considérée comme un 0 classique et ne doit pas être prise en compte dans le calcul du quantile. Pour cela, nous plaçons les données qui ont des indication "NULL" ou valant 0 dans un niveau spécifique appelé "quantile 0" ou Q_o. Ce niveau ne contient que les données manquantes ou valant 0.

Calcul des quantiles-sets

La représentation reposant sur les quantiles-sets permet donc d'informer sur les données et sur leur placement en fonction des autres. Dans notre contexte, le quantile-set d'appartenance de la valeur $v_i^{t_k}(s_j)$, obtenue pour la variable d'étude v_i par le capteur s_j au timestamp t_k, peut être calculé :

— d'un point de vue interne, à partir du même capteur s_j aux différents moments récoltés ;

— d'un point de vue externe, sur l'ensemble des valeurs des différents capteurs situés à l'instant t_k.

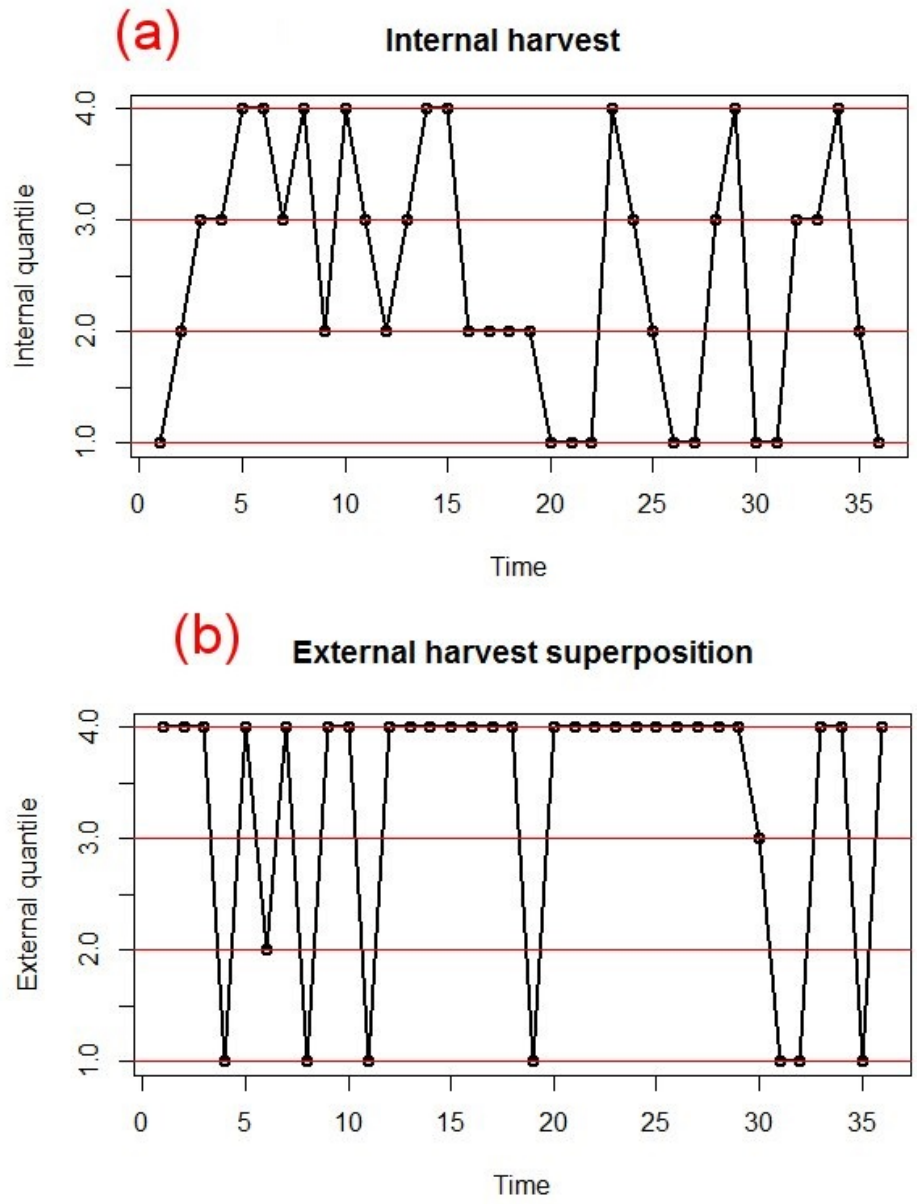


FIGURE 6.2 – Exemple de données d'un capteur représentées en quantiles-sets (ici des quartiles-sets)

Nous utilisons la fonction quantile présentée dans la section 5.1.2 et présentée dans l'équation 5.12.

Par construction, un quantile-set 0 représente un manque de données.

Q_{ext} - le quantile-set externe de $v_i^{t_k}(s_j)$ - est calculé sur l'ensemble des valeurs $\{v_i^{t_k}(s_z), z \in [1, n]\}$. Q_{ext} est exprimé comme suit (éq. 6.1) :

$$Q_{ext}(v_i^{t_k}(s_j), \{v_i^{t_k}(s_z), \forall z \in [1,n]\}) \rightarrow [0,r]$$
$$Q_{ext}(v_i^{t_k}(s_j), \{v_i^{t_k}(s_z), \forall z \in [1,n]\}) = \left\lfloor \frac{Rang(v_i^{t_k}(s_j), \{v_i^{t_k}(s_z), \forall z \in [1,n]\})}{(n/r)} \right\rfloor + 1$$
$$i \in [1,p], k \in [1,m], j \in [1,n]. \quad (6.1)$$

Pour rappel, r est le nombre de quantiles-sets (r - 1 valeurs), p est le nombre de variables, m est le nombre de timestamps.

Avec la même logique, Q_{int} - le quantile-set interne de $v_i^{t_k}(s_j)$ - est calculé sur l'ensemble des valeurs $\{v_i^{t_z}(s_j), z \in [1,m]\}$ est exprimé comme suit (éq. 6.2) :

$$Q_{int}(v_i^{t_k}(s_j), \{v_i^{t_z}(s_j), \forall z \in [1,m]\}) \rightarrow [0,r]$$
$$Q_{int}(v_i^{t_k}(s_j), \{v_i^{t_z}(s_j), \forall z \in [1,m]\}) = \left\lfloor \frac{Rang(v_i^{t_k}(s_j), \{v_i^{t_z}(s_j), \forall z \in [1,m]\})}{k/r} \right\rfloor + 1$$
$$i \in [1,p], k \in [1,m], j \in [1,n]. \quad (6.2)$$

Dans la suite, $Q_{ext}(v_i^{t_k}(s_j)$ (resp. $Q_{int}(v_i^{t_k}(s_j))$ sera noté $Q_{ext}(i,k,j)$ (resp. $(Q_{int}(i,k,j))$.

Cette représentation par quantile-set selon une vision interne et externe permet de gérer l'imprécision de données en affectant une donnée imprécise à un groupe spécifique, i.e. aux quantiles.

La mise en quantile-set permet de juger les données en volumes relatifs plutôt qu'en volumes bruts. Considérons par exemple $x_{v_1}^{t_1} = 850$ et $x_{v_1}^{t_2} = 40$. Ces deux valeurs peuvent appartenir à un même quantile-set : $Q_{v_1}^{t_1} = Q_{v_1}^{t_2}$. Dans ce cas, on dit que la variation x entre t_1 et t_2 n'est pas suffisante pour être considérée comme remarquable. En effet, x n'a pas changé de comportement au sens des quantiles bien que les valeurs réelles soient différentes.

L'ajout de Q_0 offre aussi la possibilité de repérer les instants où une valeur d'une donnée temporelle est absente. Ceci permet d'avoir des jugements sur l'absence des données.

La visualisation des distributions fondées sur les quantiles-sets permet d'obtenir des idées sur les informations relatives à la robustesse des mécanismes des robots, puisque nous passons à la présentation d'une valeur relative plutôt que d'une valeur brute. La récolte temporelle provenant de chaque capteur est maintenant présentée sur l'échelle définie par les quantiles. L'échelle fournit r niveaux permettant de repérer des données dans de nouveaux emplacements (voir la figure 6.3). Elle permet donc d'obtenir une vue claire et significative des enregistrements temporels.

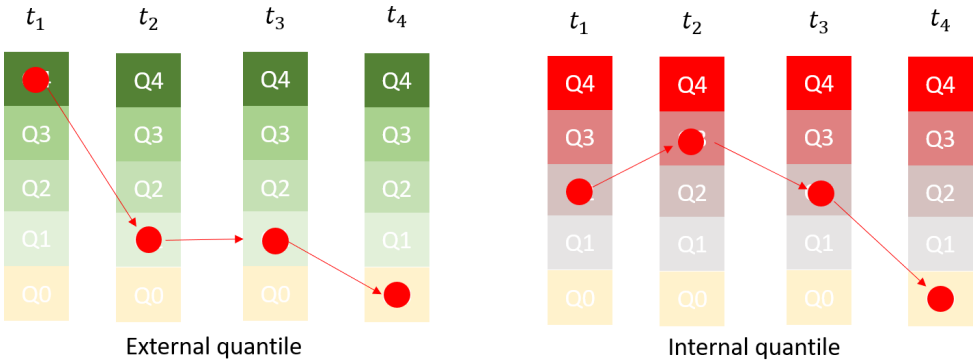


FIGURE 6.3 – Exemple d'une série temporelle représentée en quartile-set selon les visions interne et externe.

Plus particulièrement, dans la suite, nous étudions ces positions associées aux données et nous proposons des indicateurs spécifiques sur différents axes d'étude notamment la variabilité.

6.4.2 Indices de variabilité

L'objectif est de construire des indicateurs pertinents au regard du métier. Dans notre cas, il faut que les indicateurs, que nous définissons, fournissent des informations sur la qualité de la récolte des données web au travers notamment de la variabilité (dans cette section) et de la stabilité (section suivante).

Selon notre approche, le mot "variabilité" des données temporelles présente une spécificité particulière. En nous reposant sur la représentation par quantile interne et externe, nous proposons de calculer la variabilité interne des données récoltées par un capteur, mais aussi leur variabilité externe calculée au regard des données issues des autres capteurs.

La variabilité interne est la variabilité relative aux comportements de la récolte issue d'un capteur au regard de ce capteur, et donc relative au comportement du capteur. x^T est considéré comme très variable si, et seulement si, ses mouvements dans T sont très volatiles et loin d'être constants. Autrement, elle est moins variable si et seulement si, le positionnement des données a presque le même comportement dans T. Pour obtenir des informations sur cette particularité et sur le comportement d'un capteur s_i, nous souhaitons répondre à la question suivante :"Est-ce que le capteur s a le même comportement tout au long la période T ?"

La variabilité externe est liée au calcul du quantile-set externe. Sa détermination a pour but de répondre aux questions suivantes : "Est ce que le capteur s est volatile par rapport au reste des capteurs ? Et à quel point ?, Est-ce qu'il a le même comportement que les autres ?. Le calcul est donc lié au positionnement du reste des éléments. Si les valeurs d'une série temporelle reste au cours de T dans un quantile-set externe q, cette

série temporelle n'est pas tout à fait variable, car elle n'a pas varié dans T par rapport aux autres. C'est pourquoi l'approche par quantile quantifie les sauts pour calculer la variabilité.

Définitions formelles

Nous définissons Sc comme un score qui mesure la variation de la transition des positions entre deux timestamps successifs (t_{k-1},t_k) pour une variable donnée (v_i) dont les valeurs sont données en quantile.

Ce score est dépendant de deux seuils :

— Le premier, appelé jp, concerne la hauteur des sauts de quantiles entre deux timestamps successifs, i.e. c'est-à-dire la valeur absolue de l'écart des valeurs des deux positions successives exprimées en quantiles-set. Pour qu'un saut soit considéré comme représentant un passage remarquable, il faut qu'il fasse au moins la valeur de jp.

— Le second, noté b, concerne la durée de la période d'ignorance (sans données récoltées) consécutive (appelée *cip*) précédent une reprise d'activité, i.e. à t_{k-1} nous ne disposons pas de données et à t_k nous en avons. Si la durée d'inactivité n'est pas supérieure à b, nous considérons que la reprise d'activité est une transition remarquable. Si elle est supérieure, la période d'ignorance est telle, qu'il n'est pas envisageable de la considérer comme une transition significative.

Dans l'équation suivante (équation 6.3), le score Sc_{ext} est défini pour la vision externe (calculé à partir de Q_{ext}. Le même calcul peut être effectué pour Sc_{int} à partir de Q_{int}.

$$Sc_{ext}(i,k,j) = \begin{cases} 1, \text{si} & |Q_{ext}(i,k,j) - Q_{ext}(i,k-1,j)| \geq j_p \\ & \text{ou } Q_{ext}(i,k-1,j) > 0 \text{ et } Q_{ext}(i,k,j) = 0 \\ & \text{ou } Q_{ext}(i,k-1,j) = 0 \text{ et } Q_{ext}(i,k,j) > 0 \text{ et } cip \leq b \\ 0, & \text{sinon} \end{cases} \tag{6.3}$$

La variabilité de la série étudiée sur une période $T' = \{t_k, ..., t_{k+z}\}$ (eq. 6.4) :

$$Var_{ext}(i,k,k+z,j) = Var_{ext}(i,T',j) = \frac{\sum_{x=k}^{x=t+z} Sc_{ext}(i,x,j)}{z - IPMD(i,j)} \tag{6.4}$$

où $IPMD(i,j)$ est la somme des longueurs des séquences d'inter-périodes consécutives sans données de longueur strictement supérieure à b.

L'indice de variabilité d'un capteur s_j repose sur le décompte des sauts significatifs ($saut > jp$) entre les timestamps sur la période étudiée. Ce décompte est ensuite divisé par le nombre de transitions qui ne correspondent à une séquence consécutive sans données de durée supérieure à b.

Nous notons que nous sommes sur des échelles de quantiles puisque les données de base sont placées dans ces groupes particuliers. Nous considérons aussi, Q_0 le quantile-set adapté à l'absence de données détectées par un capteur s. Cet état Q_0 prend donc en considération l'ignorance que l'on a sur les valeurs d'une série temporelle.

La figure 6.4 présente un exemple avec un nombre de quantiles-sets égale à 4 dans une période de 5 inter-périodes actives (la période d'étude T présente 8 entre-périodes et 9 instants de temps).

Par conséquent, et suite à la formalisation présentée précédemment, nous en déduisons que plus la variabilité $Var(i, T', j)$ d'une série temporelle sur une période T' est haute (tend vers 1), plus la série temporelle varie sur T. Cela peut nous alerter sur le comportement du capteur en question.

Il faut aussi noter que nous considérons les sauts significatifs juste après le premier mouvement (ou juste après une période sans données d'une durée supérieure à b), c'est-à-dire que si par exemple nous étudions une période de 36 mois et que la première réaction détectée est dans le $5^{ème}$ mois, nous commençons le calcul à partir du mois suivant, i.e. le mois 6. C'est la raison pour laquelle, le calcul prend en considération le point de départ de son premier mouvement.

Dans ce qui suit, nous l'illustrons à partir de l'exemple de la figure 6.4. Dans cet exemple, nous cherchons à calculer l'indice de variabilité de la série temporelle présentée. Il s'agit d'un exemple fictif d'une série temporelle de données émises par un capteur et déjà représentées en quantiles. Nous montrons en rouge les mouvements significatifs, c'est-à-dire les sauts supérieurs à 2 et les sauts allant/venant à chacun de ces sauts est à considérer comme significatif. Dans les liaisons de données marquées en bleues, nous avons un mouvement ordinaire, par exemple entre t_3 et t_4, nous considérons qu'il s'agit d'une stabilité. Après avoir déterminé le nombre de mouvements significatifs, cette valeur sera divisée par la période effective de l'étude temporelle, c'est-à-dire la somme des périodes où nous avons un mouvement. Le résultat final de cet exemple est le suivant : $Var(i, k, j) = \frac{1+1+1+1}{8-3} = 0,80$. 8-3 car nous avons 3 inter-périodes d'inactivité de données au début ($IPMD = 3$) et que nous avons 8 inter-périodes. Comme cette variabilité est assez élevée (proche de 1), nous en déduisons la série varie beaucoup dans T, et que le comportement de ce capteur est potentiellement anormal.

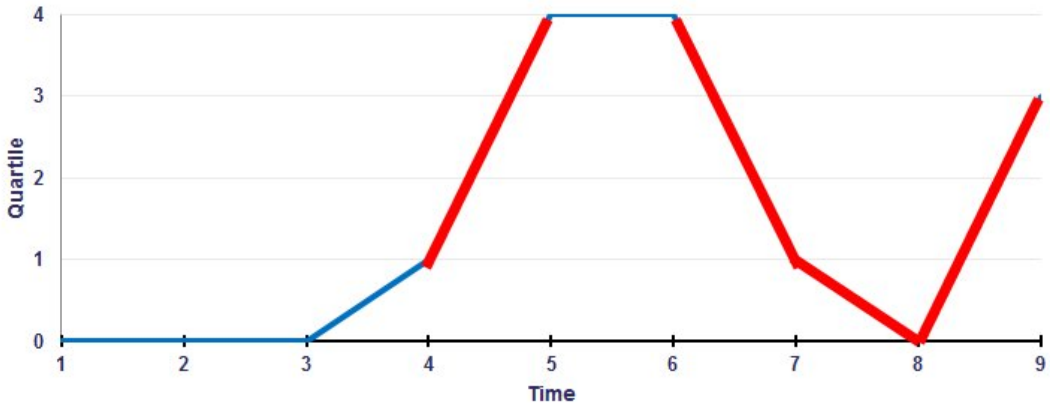


FIGURE 6.4 – Exemple d'évolution du flux de données du capteur avec $r = 4$, $jp = 2$, $b = 1$ et $var = \{v_i\}$.

6.4.3 Calcul des indices de stabilité des flux temporelles imparfaites

Afin d'obtenir un jugement unifié sur le comportement des capteurs, nous proposons une mesure qui regroupe la variabilité interne et externe d'un capteur ou d'un groupe. Cette mesure est un indice de stabilité.

On note cette mesure par $St(i, T, j)$. Cette mesure a pour objet de vérifier la cohérence du comportement d'une série temporelle d'un capteur pendant une période du temps T. Nous entendons par cohérence le faite que sa variabilité interne soit faible, et aussi que son comportement relatif aux autres séries temporelles ou groupes de séries temporelles ne soit pas très volatile (i.e. de variabilité externe faible). Dans ce cas on dit que le capteur est stable dans T. Plus le score est faible (tend vers 0), plus la série de données étudiée est stable, plus il est fort (plus il tend vers 1), plus elle sera instable.

Soit une variable v_i, une source s_j et une période de temps $T' = \{t_k, ..., t_{k+z}\}$, les scores de variabilité interne et externe sont respectivement notés Var_{int} et Var_{ext}. Nous définissons l'indice d'instabilité d'une série temporelle x^T comme suit (eq. 6.8) :

$$ISt(i, k, k+z, j) = ISt(i, T', j) \tag{6.5}$$

$$= \frac{\sqrt{Var_{ext}(i, k, k+z, j)^2 + Var_{int}(i, k, k+z, j)^2}}{\sqrt{2}} \tag{6.6}$$

$$= \frac{\sqrt{Var_{ext}(i, T', j)^2 + Var_{int}(i, k, T', j)^2}}{\sqrt{2}} \tag{6.7}$$

Considérons un plan orthonormé où la variabilité interne se situe sur l'axe des abscisses et l'externe sur l'axe des ordonnées. La valeur de l'instabilité d'une série X de variabilité interne $Var_{int}(X)$ et externe $Var_{ext}(X)$, l'instabilité est alors la distance entre le centre

du repère et le point de coordonnées ($Var_{int}(X)$,$Var_{ext}(X)$), normalisée en divisant par la distance maximale possible ($\sqrt{2}$). L'instabilité aura donc une valeur entre 0 et 1. Plus la valeur est grande, plus l'instabilité est forte. Plus elle est petite, plus elle est faible.

Nous aurions pu prendre d'autres fonctions de fusion. Nous aurions, par exemple, pu utiliser un indicateur de position ou encore des moyennes pondérées classiques ou ordonnées (OWA). La pondération est dans ce cas une question métier : il faut choisir les degrés d'influences à donner aux différentes valeurs.

A partir de l'indice d'instabilité, nous déterminons simplement l'indice de stabilité comme suit (eq. 6.8) :

$$St(i,k,k+z,j) = St(i,T',j) \qquad = 1 - ISt(i,k,k+z,j) = 1 - ISt(i,T',j) \qquad (6.8)$$

La visualisation des indices de variabilités et d'instabilités permet d'accroître la connaissance globale sur un ou plusieurs comportement de capteur comme l'illustre la figure 6.5. La découpe en zones de niveaux de stabilité, le regroupement de données par instabilité, la distinction de comportements singuliers (points isolés, bruits) sont des exemples d'analyse possible de ce type de visualisation.

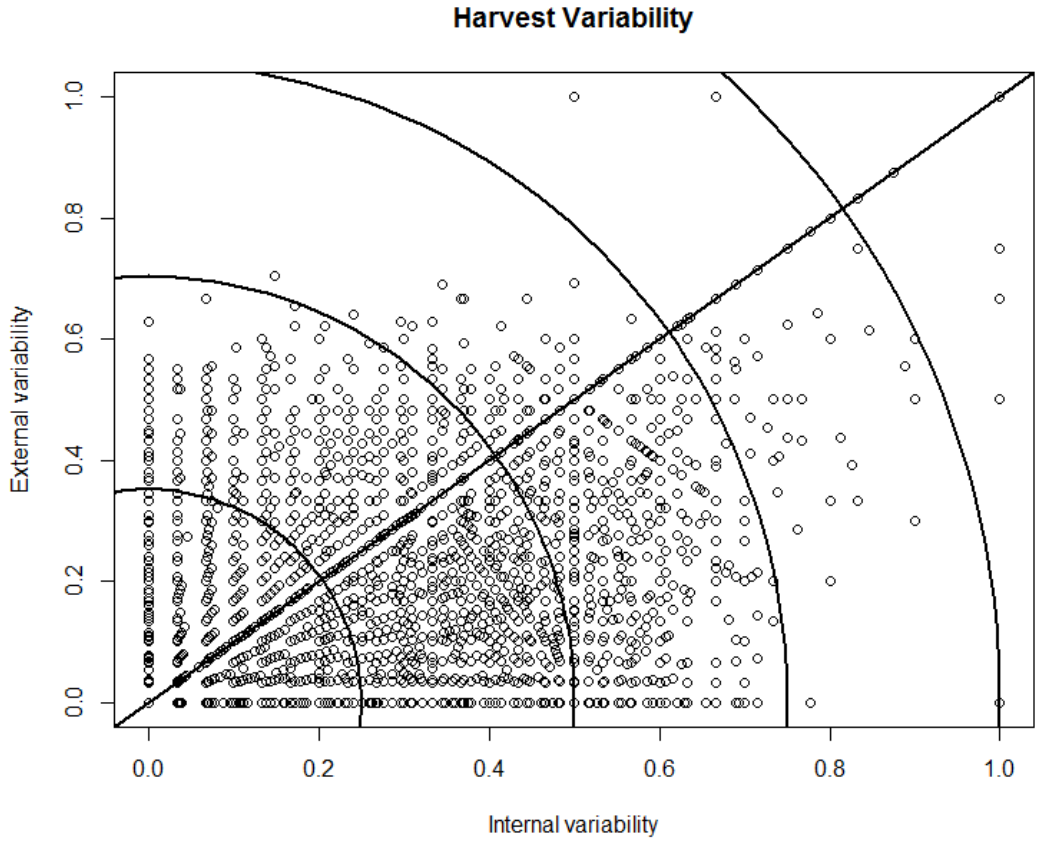


FIGURE 6.5 – Stabilité en fonction des indices de la variabilité interne et externe

Par ailleurs, les calculs précédents ont montrés un certain intérêt pour le jugement de la qualité des séries temporelles. Les définitions des indices permettent de gérer les différents facteurs de l'imperfection :

— imprecision→ Englobation en quantiles,

— incomplétude → Q_0 et le paramètre b.

L'ensemble nous permet d'extraire des connaissances préalables sur nos données, et de construire des indicateurs appropriés pour l'analyse de la qualité de nos données et de nos récoltes.

6.5 Exemple explicatif

Dans cette section, nous déroulons sur un exemple notre première approche pour le calcul de la variabilité et de la stabilité tenant compte des caractéristiques de nos données.

Supposons un ensemble S de 5 capteurs s_1, s_2, s_3, s_4, s_5 captant des données pour une variable donnée v – par exemple « le nombre d'affichages de bannières publicitaires sur une page web » – sur une période de temps $T = 10$ avec $t_i \in T\ \forall i \in [1, 10]$. Les valeurs brutes de la variable pour les 5 capteurs sur ces 10 timestamps sont données dans le tableau 6.1 et illustrées dans la figure 6.6.

Capteurs	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
s_1	50	70	80	0	0	120	320	70	50	65
s_2	80	0	150	55	95	50	70	0	100	35
s_3	0	320	0	50	65	70	80	100	0	120
s_4	70	80	100	90	120	75	0	160	130	20
s_5	75	0	160	130	20	75	150	0	130	20

TABLE 6.1 – Exemple fictif de données enregistrées pour v par les capteurs $s_{1\to5}$ sur une période T de 10 mois

Pour ce qui concerne nos paramètres, nous considérons dans cet exemple que $r = 4$ (représentation en quartile – il y a 3 valeurs de quartiles mais 4 intervalles définis par leur biais), $jp = 2$ et $b = 2$.

6.5.1 Affectation aux quantiles

Comme indiqué précédemment, l'approche QBA affecte à chaque données deux valeurs différentes de quantiles. La première correspond à la vision interne, la seconde à la vision externe.

L'idée est d'affecter la donnée $v_i^{t_k}(s_j)$ à un quantile-set qui lui correspond selon chacune des visions.

Les valeurs nulles ou valant 0 ne sont pas prises en compte dans la définition des quartiles car nous ignorons les valeurs que nous aurions dû avoir.

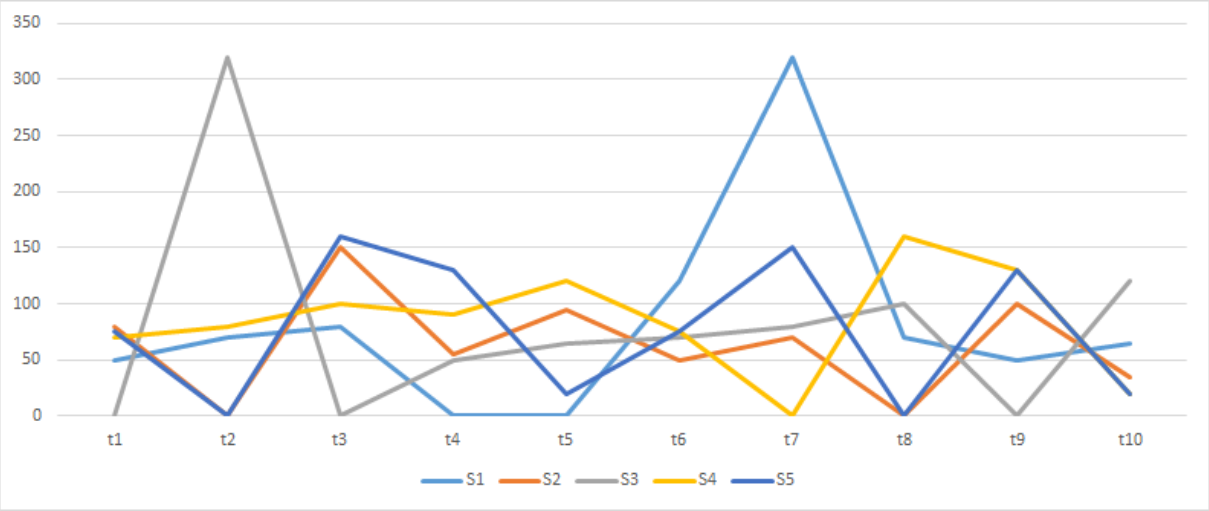


FIGURE 6.6 – Courbes des valeurs brutes de l'exemple du tableau 6.1

Quartiles/capteurs	s_1	s_2	s_3	s_4	s_5
Q_1	61,25	53,75	67,5	75	61,25
Q_2	70	75	80	90	102,5
Q_3	90	98,75	120	125	145

TABLE 6.2 – Valeurs des trois quartiles selon la vision interne pour chaque capteur

Vision interne

Dans la vision interne, il s'agit de calculer les quartiles-set par capteur : par ligne du précédent tableau. Le tableau 6.2 présentent les résultats pour l'ensemble des capteurs.

En utilisant la fonction d'affectation au quantile-set (eq. 5.12), nous pouvons obtenir les numéros de quantiles-set d'appartenance pour chacune des valeurs.

De plus, les valeurs nulles ou valant 0 sont affectées à un niveau spécial appelé Q_0. Ce niveau caractérise l'ignorance.

Cela nous permet d'obtenir les valeurs en quantiles-sets internes pour s_1 présentées dans le tableau 6.3. La courbe correspondante est donnée dans la figure 6.7.

t_i	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
s_1	50	70	80	0	0	120	320	70	50	65
$Q^T_{int}(s_1)$	Q_1	Q_2	Q_3	Q_0	Q_0	Q_4	Q_4	Q_2	Q_1	Q_2

TABLE 6.3 – Valeurs en quartiles internes pour s_1

Ce qui pour l'ensemble de nos capteurs donnent les résultats présentés dans le tableau 6.4 et la figure 6.8.

FIGURE 6.7 – Positionnement en quantiles internes des valeurs du capteur S_1

T	s_1	s_2	s_3	s_4	s_5
t_1	1	3	0	1	2
t_2	2	0	4	2	0
t_3	3	4	0	3	4
t_4	0	2	1	2	3
t_5	0	3	1	3	1
t_6	4	1	2	1	2
t_7	4	2	2	0	4
t_8	2	0	3	4	0
t_9	1	4	0	4	3
t_{10}	2	1	3	1	1

TABLE 6.4 – Quantiles d'appartenance des données du tableau 6.1 selon la vision interne

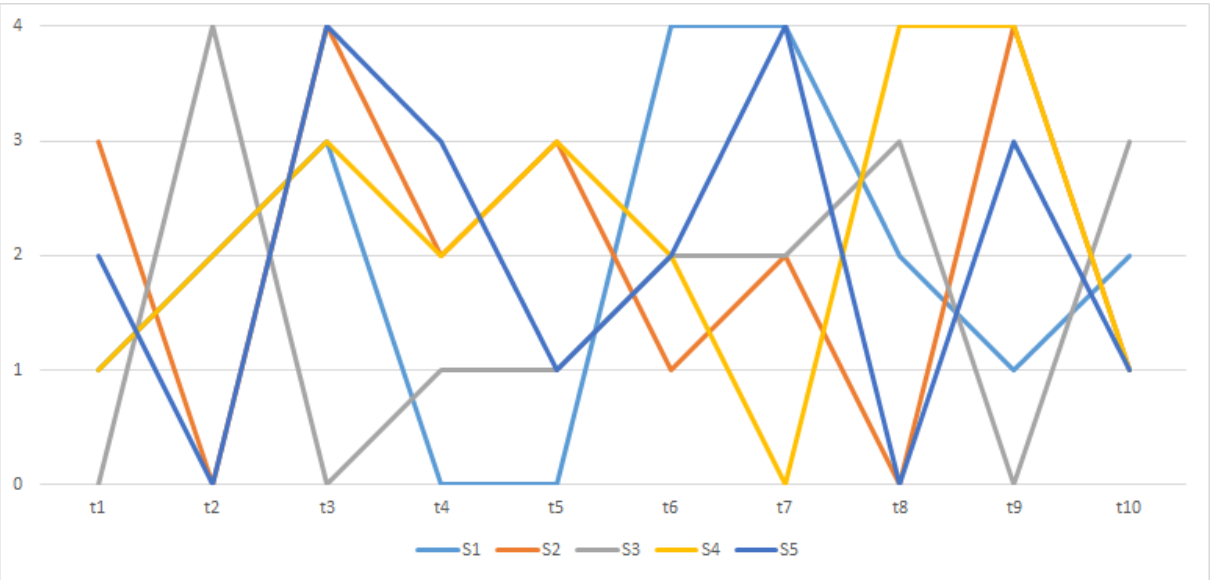


FIGURE 6.8 – Positionnement en quantiles internes des valeurs des capteurs

Vision externe

En suivant le même principe de calcul, nous allons utiliser la même logique pour calculer les valeurs en quartile-set externe relatif au capteur s_1. Pour ce faire, nous allons nous fonder sur les différents instants temporels. Pour chaque instant du temps t_i nous prenons toutes les valeurs enregistrées par tous les capteurs $s_{1\to5}$ et nous calculerons les intervalles de positionnement de la valeur de la donnée enregistrée par s_1 par rapport au reste des valeurs de données enregistrées par les autres capteurs.

Si par exemple nous cherchons à positionner la première valeur enregistrée par s_1 à l'instant t_1 par rapport à tout le reste des valeurs enregistrées, nous commençons par calculer les quartiles-sets comme suit :

1. Nous préparons les valeurs sur lesquelles seront calculés les différents quartiles externes à l'instant t_1 en enlevant les valeurs nulles (voir tableau 6.5)
2. Nous calculons, avec le reste de ces valeurs, les quartiles externes (cf. tableau 6.6)
3. Nous affectons à la valeur enregistrée par s_1 à l'instant t_1, qui vaut 50, à l'un des intervalles du tableau 6.6.
4. Nous itérons cet algorithme sur tous les timestamps $t_{1\to10}$.

capteur	s_1	s_2	s_4	s_5
v_i/t_1	50	80	70	75

TABLE 6.5 – Valeurs enregistrées par tous les capteurs ayant capté des données à l'instant t_1

min	Q_1	Q_2	Q_3	max
50.00	65.00	72.50	76.25	80.00

TABLE 6.6 – Quartiles externes à l'instant t_1 pour s_1

Après avoir effectué les calculs, la première valeur de s_1 enregistrée à l'instant t_1 est $50 \in [50, 80] \to Q1$. En suivant ce procédé, nous représentons son positionnement en quantiles-sets externes dans le tableau 6.7 et la figure 6.9.

t_i	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
s_1	50	70	80	0	0	120	320	70	50	65
$Q^T_{ext}(s_1)$	Q_1	Q_1	Q_1	Q_0	Q_0	Q_4	Q_4	Q_1	Q_1	Q_3

TABLE 6.7 – Valeurs en quantiles externes

En faisant de même pour l'ensemble des capteurs et des différents timestamps, nous obtenons le tableau 6.8 et la figure 6.10.

Ainsi, nous avons positionné chaque donnée selon ses volumes relatifs à la série temporelle de données du capteur d'où elle est issue (vision interne) et à la série de données obtenues pour l'ensemble des capteurs au timestamp où elle a été acquise.

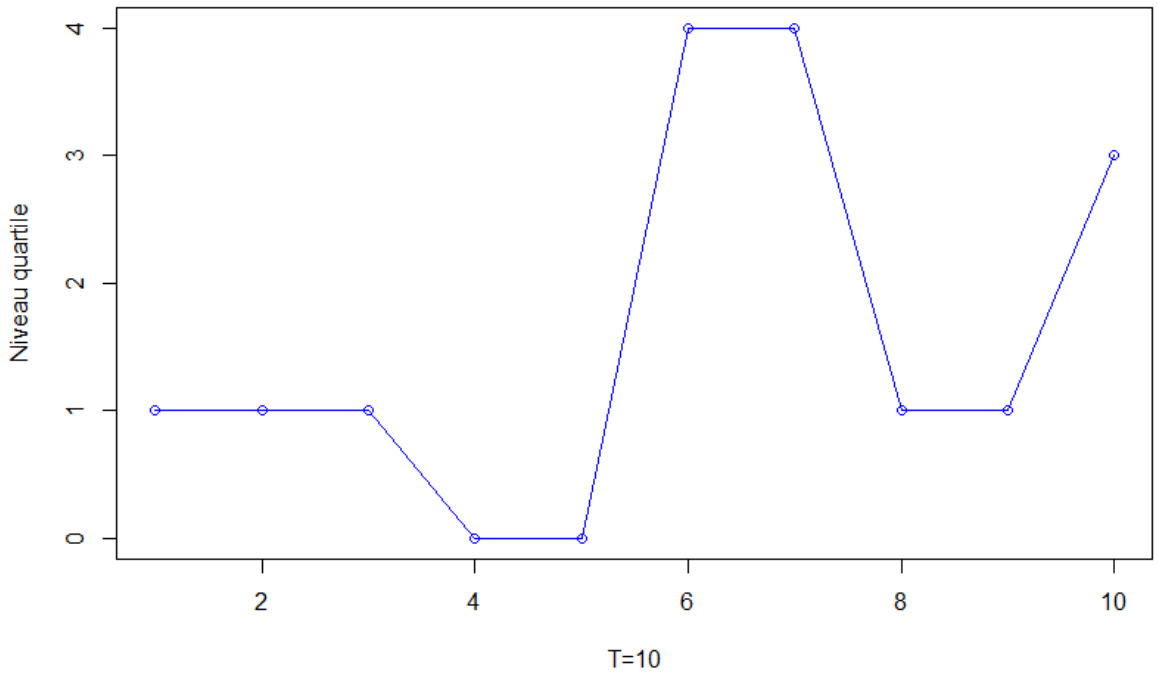


FIGURE 6.9 – Positionnement en quantiles externes des valeurs du capteur S_1

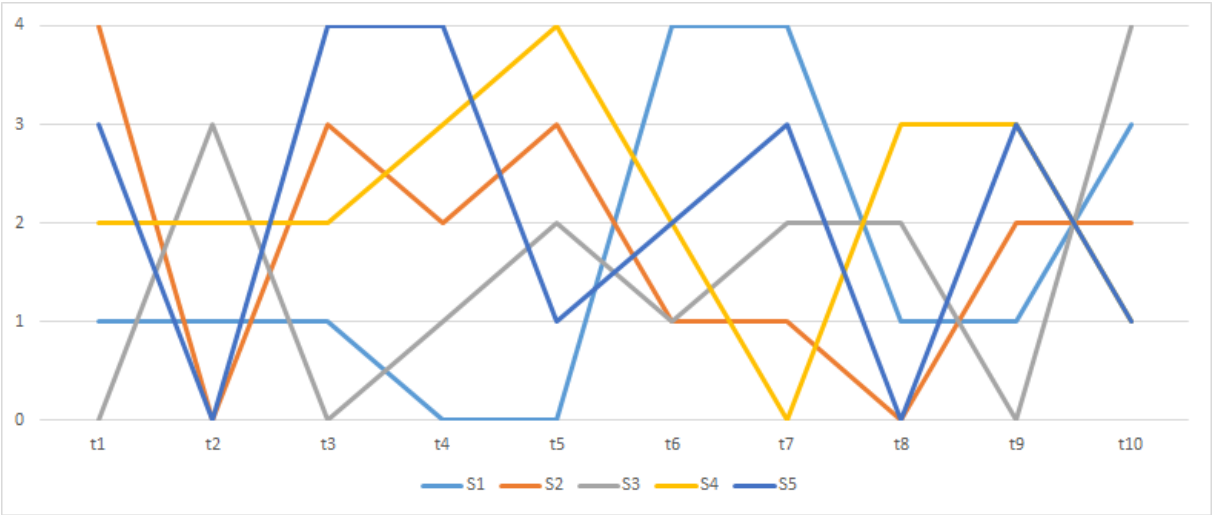


FIGURE 6.10 – Positionnement en quantiles externes des valeurs des capteurs

Nos valeurs sont désormais sur des rapports comparables (rapports entre quartiles), les imprécisions inhérentes aux données sont gérées par l'intermédiaire des indices des quantiles, et les incomplétudes sont gérées par l'introduction d'une valeur externe aux quantiles-sets.

T	s_1	s_2	s_3	s_4	s_5
t_1	1	4	0	2	3
t_2	1	0	4	2	0
t_3	1	3	0	2	4
t_4	0	2	1	3	4
t_5	0	3	2	4	1
t_6	4	1	1	2	2
t_7	4	1	2	0	3
t_8	1	0	2	4	0
t_9	1	2	0	3	3
t_{10}	3	2	4	1	1

TABLE 6.8 – Quantiles d'appartenance des données du tableau 6.1 selon la vision externe

6.5.2 Calcul des indices de variabilité

Nous calculons les indices de variabilité du capteur s_1 dans la période d'étude $T = 10$, en utilisant les valeurs précédemment indiquées de b et de jp.

$b = 2$ permet d'indiquer que si aucune donnée n'a été récoltée pour une période consécutive strictement supérieure à 2 timestamps, (1 inter-timestamp), nous considérerons que ce capteur n'est plus fonctionnel et que s'il est réactivé après ce délai, la phase d'ignorance est trop longue pour présupposer de la valeur qu'aurait dû prendre ce capteur. Les périodes d'inactivité et de reprise ne sont pas considérées dans le calcul de la variabilité. Par contre, si la période d'inactivité est inférieure ou égale au seuil b (ici inférieure ou égale à 2 timestamps, i.e. couvrant au plus une inter-période), les changements d'état vers l'état d'ignorance doivent être pris en compte dans le calcul. Nous pouvons constater que s_1 est toujours en état de fonctionnement normal, à l'instar des autres capteurs de l'exemple. Les périodes de pause supérieures à b sont exclues du calcul de la variabilité. Le choix du paramètre b influe sur le calcul de la variabilité et l'indice final.

jp est le seuil caractérisant la hauteur de saut minimale pour qu'un changement soit considérer comme remarquable. Dans cet exemple, $jp = 2$ indique que tout saut (écart entre deux positions successives) d'une hauteur plus grande ou égale à 2 constitue un passage remarquable. Par exemple, pour s_1 et le passage de t_1 à t_2, on a un saut de 1 dans l'échelle des quantiles, aussi ce saut ne sera pas remarquable et donc ne sera pas comptabilisé dans le score de variabilité. Par contre, entre l'instant t_7 et t_8, on a un saut de 2, aussi ce saut sera compté dans le calcul de la variabilité. Les passages à l'état actif - état d'ignorance, et état d'ignorance - état actif, sont tous ici comptabilisés car les durées de pertes d'information sont inférieures ou égales à b. Soit Scc, le Sc (6.3) cumulé au fur et à mesure du temps.

Calcul de la variabilité interne

Jusqu'à présent, nous avons fixé les paramètres $r = 4$, $b = 2$, et $jp = 2$.

Passage – Hauteur du saut	s_1	s_2	s_3	s_4	s_5
$t_1 - t_2$	1	3	4	1	2
$t_2 - t_3$	1	4	4	1	4
$t_3 - t_4$	3	2	1	1	1
$t_4 - t_5$	0	1	0	1	2
$t_5 - t_6$	4	2	1	1	1
$t_6 - t_7$	0	1	0	2	2
$t_7 - t_8$	2	2	1	4	4
$t_8 - t_9$	1	4	3	0	3
$t_9 - t_{10}$	1	3	3	3	2

TABLE 6.9 – Hauteurs des sauts entre positions selon la vision interne

Le tableau 6.9 présente les hauteurs de sauts selon la vision interne pour notre exemple.

Le tableau 6.10 illustre pas à pas le calcul du score cumulé permettant d'obtenir la variabilité interne pour le capteur s_1.

Itération	Transition temporelle	Hauteur du saut	Sc	Valeur du cumul de Sc (Scc)	Remarques
1	$t_1 \rightarrow t_2$	1	0	0	$saut < jp$
2	$t_2 \rightarrow t_3$	1	0	0	$saut < jp$
3	$t_3 \rightarrow t_4$	3	1	1	passage remarquable à un état d'ignorance
4	$t_4 \rightarrow t_5$	0	0	1	pas de données
5	$t_5 \rightarrow t_6$	4	1	2	et $cip <= b$ et $saut > jp \rightarrow$ passage remarquable vers un état actif
6	$t_6 \rightarrow t_7$	0	0	2	$saut < jp$
7	$t_7 \rightarrow t_8$	2	1	3	$saut \geq jp \rightarrow$ passage remarquable
8	$t_8 \rightarrow t_9$	1	0	3	$saut < jp$
9	$t_9 \rightarrow t_{10}$	1	0	3	$saut < jp$

TABLE 6.10 – Exécution pas à pas du calcul du nombre de sauts remarquables selon les quantiles-set internes et $jp = 2$

Après cette exécution, notre variable d'itération Scc est égale à 3, i.e nous avons trouvé trois passages remarquables selon les paramètres du début.

En appliquant la formule de la variabilité (eq. 6.4), on obtient pour s_1 :

$$Var_{int}(v, T, 1) = \frac{3}{9} = 0.333 \tag{6.9}$$

Si nous regardons maintenant l'ensemble des capteurs, nous obtenons les scores de variations pour les différents passages présentés dans le tableau 6.11.

Passages	s_1	s_2	s_3	s_4	s_5
$t_1 - t_2$	0	1	1	0	1
$t_2 - t_3$	0	1	1	0	1
$t_3 - t_4$	1	1	1	0	0
$t_4 - t_5$	0	0	0	0	1
$t_5 - t_6$	1	1	0	0	0
$t_6 - t_7$	0	0	0	1	1
$t_7 - t_8$	1	1	0	1	1
$t_8 - t_9$	0	1	1	0	1
$t_9 - t_{10}$	0	1	1	1	1

TABLE 6.11 – Scores de variations remarquables entre les différents timestamps

	s_1	s_2	s_3	s_4	s_5
Variabilité interne (Var_{int})	0,33	0,78	0,56	0,33	0,78

TABLE 6.12 – Valeurs de l'indice de variabilité interne des capteurs sur T

Au final, nous obtenons les variabilités internes présentées dans le tableau 6.12. Nous pouvons remarquer que s_2 et s_5 semblent avoir la plus grande variabilité.

Calcul de la variabilité externe

En suivant, le même principe, nous calculons la variabilité externe de s_1. Le tableau 6.13 montre une exécution pas à pas faite sur les valeurs du tableau 6.7 pour s_1. À partir du score cumulé, la variabilité externe pour ce capteur est calculée dans l'équation 6.10

Itération	Transition temporelle	Hauteur du saut	Valeur du *Scc* en cumul	Remarques
1	$t_1 \rightarrow t_2$	0	0	$saut < jp$
2	$t_2 \rightarrow t_3$	0	0	$saut < jp$
3	$t_3 \rightarrow t_4$	(1)	1	passage remarquable à un état d'ignorance
4	$t_4 \rightarrow t_5$	0	1	$saut < jp$
5	$t_5 \rightarrow t_6$	4	2	$cip <= b$ et $saut > jp \rightarrow$ passage remarquable (sortie d'un état d'ignorance)
6	$t_6 \rightarrow t_7$	0	2	$saut < jp$
7	$t_7 \rightarrow t_8$	3	3	$saut > jp \rightarrow$ passage remarquable
8	$t_8 \rightarrow t_9$	0	3	$saut < jp$
9	$t_9 \rightarrow t_{10}$	2	4	$saut \geq jp \rightarrow$ passage remarquable

TABLE 6.13 – Exécution pas à pas du calcul du nombre de sauts remarquables selon les quantiles-set externes et $jp = 2$

	s_1	s_2	s_3	s_4	s_5
Variabilité externe (Var_{ext})	0,44	0,56	0,56	0,44	0,67

TABLE 6.14 – Valeurs de l'indice de variabilité externe des capteurs sur T

$$Var_{ext}(v, T, 1) = \frac{4}{9} = 0.44 \tag{6.10}$$

Le fait que $Var_{ext}(v, T, 1) > Var_{int}(v, T, 1)$ (0.44>0.33) montre que les valeurs données par ce capteur varient plus par rapport aux autres que par rapport à lui même. En effet, on note qu'il y a 4 passages remarquables en externe contre 3 en interne. On peut donc aussi dire que le comportement du capteur est plus cohérent avec lui même qu'avec les autres.

Pour l'ensemble des capteurs, les scores de variabilité externe obtenus sont donnés dans le tableau 6.14.

Nous pouvons nous rendre compte que la variabilité interne et externe de s_3 sont identiques.

6.5.3 Calcul de la stabilité

L'indice d'instabilité est une mesure proposée pour combiner les deux indices de variabilité. C'est une mesure qui cherche à unifier le jugement sur une série de données temporelle tout en se fondant sur le mécanisme de l'approche QBA. Si les deux indices de variabilité sont mesurés sur une échelle de 0 à 1, une valeur unifiée peut donc avoir un maximum de 1.

Selon l'indice proposé dans notre approche, la valeur de l'indice d'instabilité pour s_1 est donnée dans l'équation 6.11.

$$ISt(v, T, 1) = \frac{\sqrt{Var_{ext}(v, T, 1)^2 + Var_{int}(v, T, 1)^2}}{\sqrt{2}} = \frac{\sqrt{0.44^2 + 0.33^2}}{\sqrt{2}} = 0.3889 \tag{6.11}$$

Cependant, comme indiqué dans la section 6.4.3, d'autres approches sont possibles. Toute méthode d'agrégation est en soi possible.

Prenons, par exemple les opérateurs d'agrégation suivant (qui sont aussi des indicateurs de positions) :

- Moyenne
- Min
- Max
- Moyenne pondérée avec trois configurations différentes :
  - MoyPondérée 1 : la variabilité interne compte le double de la variabilité externe

	s_1	s_2	s_3	s_4	s_5
Var_{int}	0,33	0,78	0,56	0,33	0,78
Var_{ext}	0,44	0,56	0,56	0,44	0,67
ISt	0,39	0,68	0,56	0,39	0,72
$\mathcal{M}$	0,39	0,67	0,56	0,39	0,72
min	0,33	0,56	0,56	0,33	0,67
max	0,44	0,78	0,56	0,44	0,78
MoyPondérée 1	0,37	0,70	0,56	0,37	0,74
MoyPondérée 2	0,41	0,63	0,56	0,41	0,70
OWA	0,41	0,70	0,56	0,41	0,740
Moyenne des différentes variabilités	0,39	0,67	0,56	0,39	0,73
Medianne des différentes variabilités	0,39	0,68	0,56	0,39	0,72

TABLE 6.15 – Scores d'instabilité selon différents agrégateurs

Opérateurs	Ecarts absolus aux moyennes	Ecarts absolus aux médianes
ISt	0,005	0,000
$\mathcal{M}$	0,016	0,019
min	0,294	0,297
max	0,262	0,259
MoyPondérée 1	0,089	0,089
MoyPondérée 2	0,096	0,096
OWA	0,077	0,073

TABLE 6.16 – Somme des écarts absolus aux moyennes et aux médianes selon les différents agrégateurs

— MoyPondérée 2 : la variabilité externe compte le double de la variabilité interne
— OWA (Ordered Weighted Average) : la variabilité la plus élevée compte le double de la plus faible.

Les résultats de ces différentes approches sont donnés dans le tableau 6.15. Leur moyenne et leur médiane par capteur sont aussi données.

Si on somme les écarts absolus aux moyennes et aux médianes pour l'ensemble des capteurs, on obtient le tableau 6.16. On peut remarquer que les valeurs, issues de l'opérateur que l'on a choisi, se situent pour nos données au centre des valeurs issues des différents opérateurs testés. Aussi, notre mesure d'instabilité semble intéressante.

Les valeurs de stabilité qui en découle selon l'équation 6.8 sont données dans le tableau 6.17. Pour rappel, $St = 1 - ISt$.

Sur notre exemple, le capteur fournissant la récolte la plus stable est donc le capteur s_5.

	s_1	s_2	s_3	s_4	s_5
St	0,61	0,32	0,44	0,61	0,28

TABLE 6.17 – Valeurs de l'indice de stabilité

6.6 Étude de sensibilité

L'objectif de cette section est d'examiner le choix des différents paramètres en suivant une méthodologie de vérification de l'impact des paramètres de nos définitions sur la variabilité et sur la stabilité de la récolte.

Cette méthodologie cherche à étudier les différents paramètres de base de l'approche QBA en faisant une étude de sensibilité. L'impact sera étudié selon des critères que nous préciserons en fonction des expériences.

Dans ce qui suit, nous présentons les étapes de cette méthodologie en utilisant les données de Kantar.

Cette étude de sensibilité s'intéresse aux paramètres de nos indices qui sont :

— r : nombre de quantiles ;
— jp : hauteur minimale de saut entre les quantiles successifs pour considérer une transition comme un passage significatif ;
— b : durée maximale d'une période d'ignorance pour considérer la réactivation comme un passage significatif.

Influence du nombre de quantiles-set

Nous commençons par examiner le choix du premier paramètre r. Nous évaluons la cohérence des ensembles de données obtenus par l'utilisation des quantiles-sets à l'aide des indices Silhouette [Rou87] et Dunn [PBM04]. Ces indices quantifient la qualité (pertinence) de l'appartenance des données à leur quantile.

Les résultats statistiques présentés dans le tableau 6.18 sont calculés sur la base de l'ensemble de nos données (cf. section 2.4) à l'échelle mensuelle à la fois selon la vision interne (quantile interne) et la vision externe (quantile externe). Pour rappel, nous avons 700 capteurs, donnant des données sur 3 variables pour une période d'étude de 36 mois consécutifs.

Afin de choisir un nombre optimal (r^*) de quantiles-sets, nous pouvons par exemple nous fonder sur un ou plusieurs résumés statistiques. Si par exemple, nous considérons la moyenne comme mesure de validation, $r = 4$ semble la valeur la plus appropriée.

Si nous choisissons une autre méthode de choix multicritères, comme par exemple un vote majoritaire, un vote pondéré, etc., nous pouvons obtenir un autre r^*.

Nous choisissons dans la suite de cette étude de sensibilité, $r^* = 9$ afin d'avoir un nombre de quantiles suffisamment important pour examiner le choix de jp^*, i.e le seuil optimal pour la hauteur minimale de saut entre les quantiles.

Silhouette Index

r	Min	Q_1	Mean	Max
$r = 4$	-0.2864	0.4566	0.4593	0.7540
$r = 5$	-0.2974	0.4356	0.4506	0.6278
$r = 6$	-0.3816	0.4165	0.4452	0.7061
$r = 9$	-0.2790	0.3794	0.4322	0.5532
$r = 10$	-0.2740	0.380	0.4307	0.586

Dunn Index

$r = 4$	0.0004234	0.4603	0.4462	1.6703
$r = 5$	0.28640	0.4603	0.4645	1.6700
$r = 6$	0.00079	0.4603	0.4276	1.6990
$r = 9$	0.00031	0.3298	0.3801	2.7190
$r = 10$	0.0007039	0.285	0.3710	1.108

TABLE 6.18 – Influence du nombre de quantiles-sets sur la qualité de la distribution de données selon les mesures Silhouette et Dunn.

Influence des seuils

Nous cherchons à observer l'impact de la valeur du seuil sur la hauteur de saut entre positions successives exprimées en quantiles.

Dans cet objectif, nous calculons les valeurs médianes et moyennes de variabilité en fonction du seuil sur la hauteur de saut. Pour cela, nous fixons les autres paramètres avec les valeurs suivantes : $r^* = 9$ et $b = 2$. Les résultats sont donnés dans le tableau 6.19.

Nous pouvons constater que la variabilité dépend du choix du jp. Si le jp augmente, la variabilité diminue car il y a de fait moins de sauts à considérer.

$r^*_{b=2} = 9$	Median(V_{ext})	Mean(V_{ext})	Median(V_{int})	Mean(V_{int})
$jp = 2$	0.7996737	0.7732519	0.7444	0.5358
$jp = 3$	0.7996240	0.7732425	0.7444	0.5255
$jp = 4$	0.7995924	0.7732330	0.7412	0.5179
$jp = 5$	0.7995292	0.7732235	0.7355	0.5113

TABLE 6.19 – Exemple d'indices de variabilité interne et externe calculés pour plusieurs sauts compris entre 2 et 5 pour $r^* = 9$, $b = 2$ et $T = 36$.

Nous avons aussi diversifié le choix du paramètre b en fonction de jp. Les résultats sont présentés dans le tableau 6.20. De ce tableau, nous pouvons déduire que la variabilité dépend aussi du paramètre b. Nous pouvons voir que si b augmente la variabilité augmente. Cela est dû au fait que le nombre de passages remarquables augmente car le nombre de réactivations (passage de l'ignorance à des données récupérées) non considérées augmente. Par conséquent, il faut trouver un ajustement entre ces deux paramètres pour en déduire les mesures optimales.

Pour le choix de b, nous pouvons par exemple nous fonder sur des mesures de calcul de similarité entre les lignes du tableau 6.20, par exemple en appliquant la distance de Jaccard, la distance de Hamming, etc. ou choisir d'autres stratégies d'études.

Mesures moyennes de la variabilité interne

mean	jp=2	jp=3	jp=4	jp=5
b=2	0.5358	0.5255	0.5179	0.5113
b=3	0.7685	0.7685	0.7685	0.7685
b=4	0.7625	0.7623	0.7622	0.7621

Mesures moyennes de la variabilité externe

	jp=2	jp=3	jp=4	jp=5
b=2	0.7732	0.7732	0.7732	0.7732
b=3	0.77110	0.7710	0.7710	0.7710
b=4	0.7684	0.7684	0.7684	0.7684

TABLE 6.20 – Variabilité interne et externe en moyenne pour $r^* = 9$ et $T = 36$. Les indices sont calculés en faisant varier les seuils b sur les pauses et jp sur les sauts.

Admettons maintenant que $b^* = 2$. Selon nos données, cette valeur représente la plus faible mesure obtenue sur la première ligne du tableau de la variabilité externe et des mesures acceptables (si l'objectif est d'avoir globalement une variabilité faible) sur la variabilité interne. A noter aussi que ce paramètre peut être défini par les experts, et ce choix n'est qu'une recommandation.

Ainsi, après avoir fixé r^* et b^*, le choix de jp n'est pas encore déterminé. Pour cela nous allons étudier comment la variation de jp peut affecter le jugement sur la stabilité de la récolte. La figure 6.11 montre l'impact de jp sur la stabilité. Nous avons fait varier jp et utilisé la distribution de densité sur les valeurs de la stabilité. Nous pouvons remarquer que si jp augmente, la distribution se rapproche de 1, c'est-à-dire que la récolte est considérée comme stable. Ce qui est tout à fait logique.

En effet, plus la valeur du seuil sur la hauteur du saut est grande, plus le nombre de sauts remarquables est petit. Aussi, comme indiqué précédemment, plus la valeur de jp est grande, plus les valeurs de la variabilité sont petites. Par conséquence, plus jp est grand plus l'instabilité est petite. Or, la stabilité est le complément de l'instabilité, du fait que nous considérons que St(v,T,1) = 1 - ISt(v,T,1). En conséquence, la valeur de jp impacte donc fortement les indices de stabilité.

Pour éviter les erreurs de jugement sur le choix de jp, ce dernier doit être à considérer selon les objectifs de l'étude. Si l'objectif est de ne considérer que les variabilités très prononcées, une valeur élevée, au regard du nombre de quantile, pour jp est à privilégier. Si l'objectif est d'obtenir une variabilité fine, on aura plutôt tendance à choisir un jp petit.

Si nous n'avons pas de contraintes sur l'interprétation des variabilités, le choix peut se porter vers une valeur intermédiaire ou raisonnable. Dans ce cas, un choix pour la valeur du seuil du saut peut être par exemple $jp^* = \frac{r^*}{2}$. Nous pouvons aussi nous fonder sur la distribution des valeurs de la stabilité en fonction de jp pour en déduire un choix. Par exemple, au regard de la figure 6.11 et des formes des distributions, un choix pourrait être entre $jp^* = 1$ et $jp^* = 2$ car leurs distributions sont plus étalées et moins centrées sur 1.

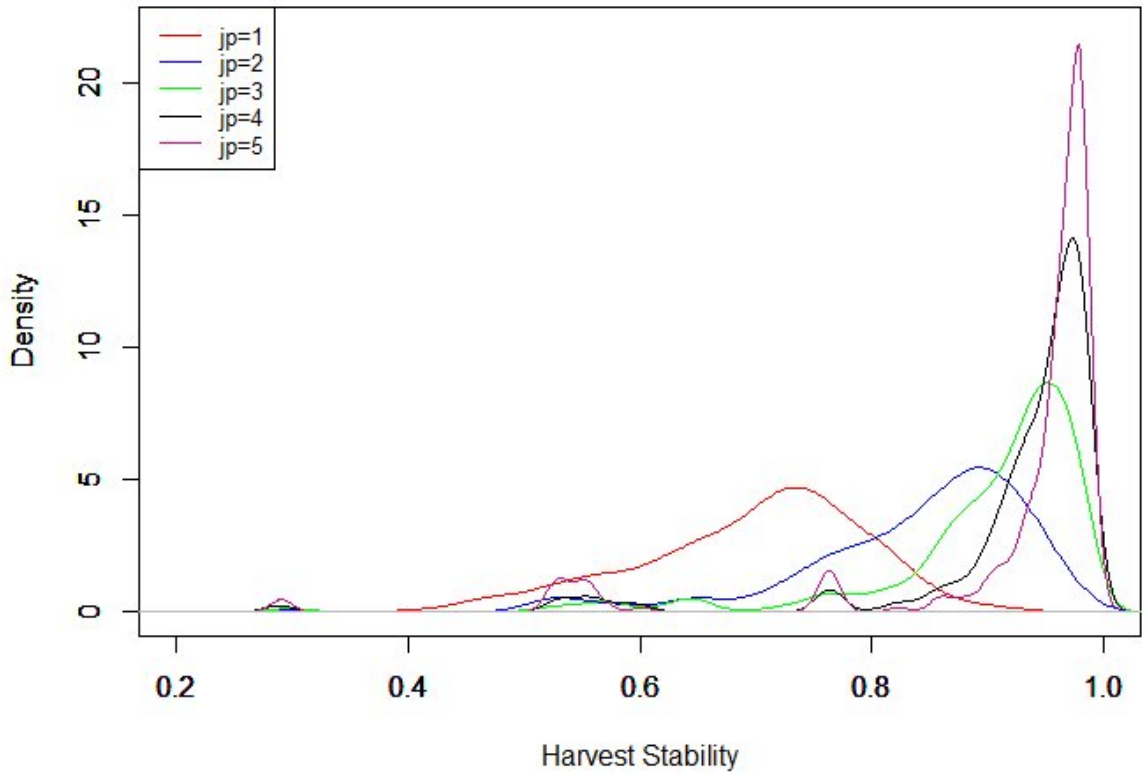


FIGURE 6.11 – L'impact de la variabilité sur la distribution de stabilité concernant plusieurs jp. Les calculs sont effectués en raison de $r^* = 9$, $b^* = 2$ et $T = 36$.

Dans cette étude de sensibilité, nous observons que les paramètres influent grandement sur la mesure de la variabilité et que leur estimation requière une attention particulière. Nous recommandons donc cette méthodologie afin de déterminer les potentielles anomalies dans les séries temporelles de données imprécises comme c'est le cas des données de Kantar. Dans ce qui suit, nous présentons une étude sur nos données reposant sur ces indicateurs.

6.7 Résultats et discussion

Dans cette section, nous utilisons l'approche QBA pour déduire des indications pertinentes sur la qualité de la récolte.

Les résultats indiqués dans cette section sont obtenus sur nos données en faisant varier le choix des variables d'études, la période et le nombre des instants temporels.

Comme les indicateurs de variabilité peuvent donner une idée préalable de la mise en place significative dans les séries chronologiques, nous avons appliqué une Classification Hiérarchique Ascendante [3] (CAH) avec le critère de Ward [4] aux indices de variabilité sur cette période selon nos trois variables d'études. L'objectif est de voir s'il est possible de distinguer certains liens comportementaux entre les capteurs et possiblement la manière dont les annonces sont placées dans les sites Web, les catégories de sites Web, les méta-catégories, etc. La figure 6.12 montre que nos robots d'exploration suivent quatre stratégies différentes. Chaque cluster représente un comportement similaire des capteurs regroupés et met en évidence leur variabilité en fonction de toutes les variables. Ce type d'intégration visuelle a été réalisé afin d'exploiter la véracité des informations et de différencier les comportements des capteurs.

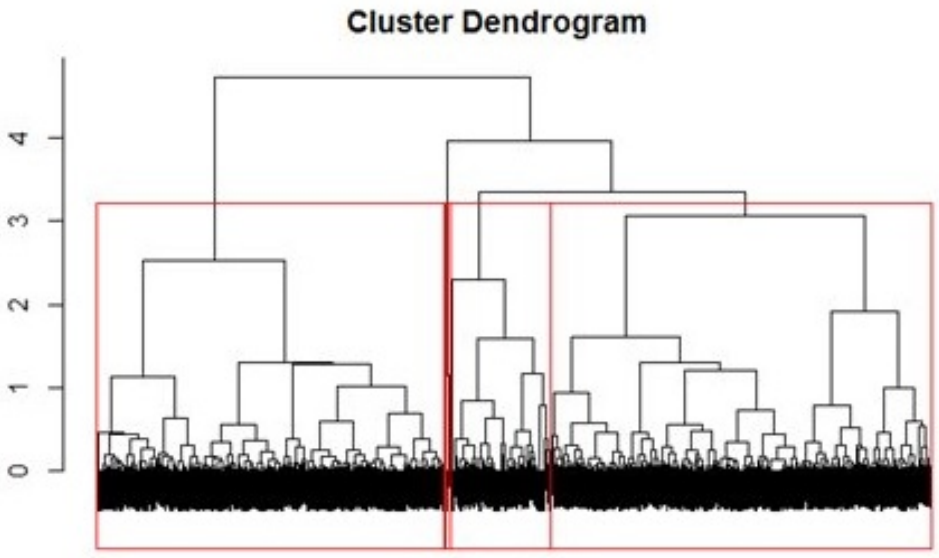


FIGURE 6.12 – Classification de la variabilité externe par CAH et critères de Ward sur 36 mois. Les indices de calcul mensuels sur la e de $r = 4$, $jp = 2$, $b = 1$ et $var = \{v_1, v_2, v_3\}$ fournissent 4 stratégies d'analyse distinctes.

Regrouper les comportements des capteurs en fonction de leurs scores de variabilité, nous permet de détecter des phénomènes significatifs dans nos données. Par exemple, nous avons clairement un ensemble réduit de capteurs pour lesquels le comportement de leur récolte est très éloigné des autres. Leur variabilité interne comme externe sur les trois variables est beaucoup plus importante.

Ce type de groupement permet aussi l'identification de certains comportements atypiques i.e soit un capteur a récolté énormément de données sur un site ou bien l'inverse ce qui donne aussi une nouvelle information sur la qualité de la récolte.

3. Méthode d'analyse par grappes qui cherche à établir une hiérarchie de grappes
4. En statistique, la méthode de Ward est un critère appliqué dans l'analyse hiérarchique par grappes

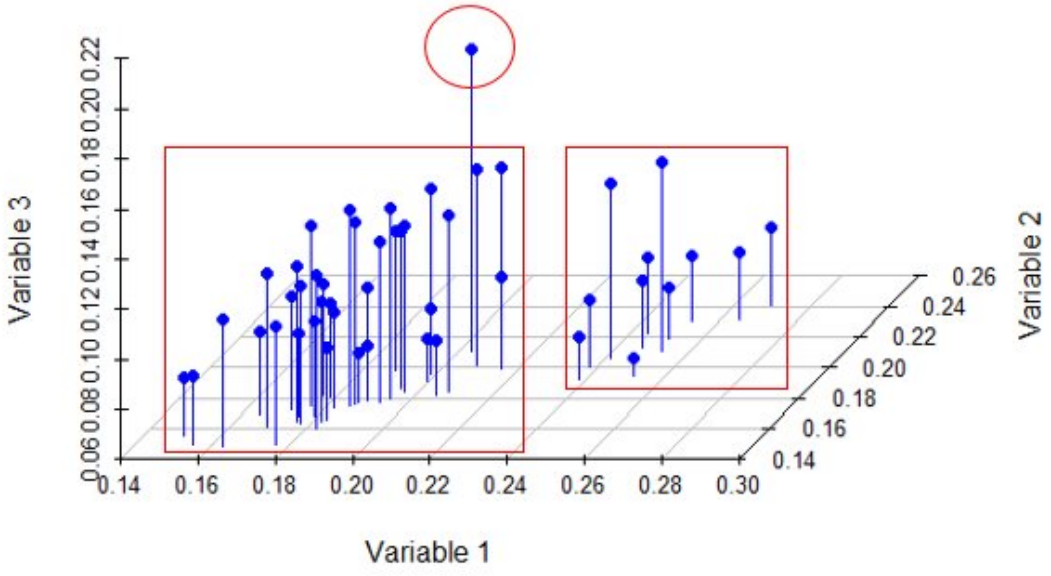


FIGURE 6.13 – Vision 3D des indicateurs d'instabilité de 700 capteurs, calculée sur la e de l'horodatage mensuel avec les paramètres $r = 4$, $jp = 2$ et $b = 1$. Chaque axe fait référence à une variable d'étude(voir chapitre 2) pour les variables.

En ce qui concerne les indications déterminées à partir du calcul de la stabilité, combinant la variabilité interne et externe sur 3 variables d'étude, nous obtenons la figure 6.13. Nous pouvons facilement identifier 2 grands groupes et un petit groupe plus isolé. Le groupe en à gauche regroupe les capteurs les plus stables sur les trois variables. Il pourrait être considéré comme un échantillon de récoltes de capteurs de confiance sur-lesquelles nous pouvons fonder nos analyses, pour par exemple, calculer les investissements publicitaires des annonceurs. Le second groupe, en haut à droite, est formé de capteurs ayant une instabilité plus élevée pour les variables 1 et 2. Le groupe plus isolé comporte lui des capteurs fournissant des récoltes plus instables pour les variables 2 et 3. Nous pourrions écarter les capteurs qui composent ces deux groupes pour en vérifier les configurations car ils sont potentiellement défaillants.

6.8 Conclusion

L'approche présenté permet de travailler sur la variabilité des séries temporelles de données imparfaites. Elle a notamment les avantages suivant par rapport à la littérature :

- Elle permet de comparer les données en se fondant sur le volume relatif
- Elle intègre les facteurs d'imperfection constatés sur nos données
- Elle permet la classification de l'absence de données afin de trouver des correspondances dans une récolte volumineuse
- Elle détecte les mouvements significatifs sur un ou plusieurs capteurs

— Elle fournit des indicateurs permettant de juger les données imparfaites sur plusieurs axes d'étude

Industriellement, elle vise à étudier le comportement des capteurs. L'approche QBA, nous a permis de construire des indicateurs permettant de vérifier la qualité de la récolte en termes de variabilité et de stabilité. Cette approche est paramétrique. Aussi, nous avons proposé une étude de sensibilité qui vise à comprendre l'impact des paramètres sur les scores de nos indicateurs.

Ainsi, nous pouvons à l'aide de cette approche obtenir des informations clés sur le fonctionnement de nos capteurs, sur les potentiels anomalies. Cette approche permet par exemple d'identifier des données de confiance mais aussi des données plus singulières. Cette approche est un outil d'étude de la qualité des séries temporelles de données imparfaites dans une masse de données.

En effet, l'approche QBA nous permet de construire un système de vérification de la qualité des séries temporelles et avoir des jugements sur le fonctionnement de nos capteurs, système qui sera présenté dans sa globalité dans la partie IV.

Le processus repose sur un système paramétrique. Différentes configurations de ces derniers peut donner différents jugements. Notre système est dépendant de la considération de nos imperfections et de la nature de nos données. Ce système représente les données par leur quantile. Cela amène un biais possible du fait que deux données proches peuvent être dans des quantiles différents, tandis que des données éloignées peuvent être placées dans le même quantile en fonction de la distribution des valeurs.

Ainsi, l'approche QBA repose principalement sur une découpe stricte des données dans des groupes différents (les quantiles). Une découpe en ensembles stricts de données imprécises présente l'inconvénient de ne pas tenir compte des intervalles de confiance de la donnée. En effet, la plage d'existence d'une donnée imprécise peut être étalée sur plusieurs groupes. Par exemple, soit $x = 60$ sur une variable d'étude v_i et deux groupes Q_1 et Q_2 englobant les valeurs de v_i [0,50] et [50,100] respectivement. Sur une découpe stricte $x \in Q2$ car $60 \in [50, 100]$. Cependant en réalité x est imprécise et pourrait par exemple avoir pour valeurs admissibles toutes valeurs dans l'intervalle [40,70]. Dans ce cas l'affectation de x à Q_2 n'est pas totalement juste, car le facteur d'imprécision n'est pas complètement considéré.

La représentation par la théorie des ensembles des données imprécises peut être le point faible de cette première approche. Comme vu dans le chapitre 4, les approches floues représentent un puissant moyen pour le traitement de l'imprécision dans les données. Nous positionnons nous dans ce cadre dans notre prochaine contribution pour rendre notre système plus générique et plus adapté au traitement des séries temporelles de données imparfaites.

Chapitre 7

Contribution 2 : Approche floue

7.1 Introduction

Dans le chapitre précédent, nous avons proposé une première approche pour l'étude de la variabilité fondée sur le passage par les quantiles. Cette démarche est intéressante, mais repose sur la représentation des données imprécises via leur quantile et donc en ensembles classiques. Cela implique que les éléments en bordure d'ensemble ne seront pas considérés comme appartenant, au moins partiellement, à l'ensemble voisin. Cette vision stricte des ensembles ainsi que le fait de devoir prédéfinir une échelle unique de valeurs (le nombre de quantiles) sont des limites à notre première contribution.

Dans notre seconde approche, qui est l'objet de ce chapitre, nous proposons de généraliser la première approche à l'aide de calculs par densité de groupements des données d'une part, et de fuzzification de ces groupes d'autre part afin de considérer des appartenances partielles des données aux groupes. En nous reposant sur ce principe, nous proposons une nouvelle approche pour le positionnement de chaque donnée dans leur ensemble fondée sur ces groupes flous. Nous étudions ensuite la variabilité au regard des variations de ces positions.

Ce chapitre est organisé comme suit. Nous exposons dans un premier temps les objectifs ainsi que les principes et hypothèses au centre de cette proposition. Nous continuons par la présentation de la démarche de positionnement flou. Puis, nous présentons sur un exemple comment l'indice de variabilité défini dans l'approche QBA peut être ré-exploité ici. Ensuite, une expérimentation est présentée. Enfin, nous discutons de l'approche et concluons ce chapitre.

7.2 Objectifs de l'approche

Dans ce chapitre, l'approche proposée (FBA) a les mêmes objectifs que l'approche QBA présentée dans le chapitre précédent. Nous cherchons à étudier la variabilité et la stabilité de séries temporelles de données imprécises et incomplètes. Aussi à l'instar de ceux de l'approche QBA, les objectifs de l'approche FBA sont :

— traiter et gérer des données imprécises et incomplètes ;

— considérer les possibles facteurs d'échelle (ou de puissance) présents dans les valeurs des données ;
— considérer la temporalité des séries étudiées en étudiant les passages remarquables ;
— considérer les possibles phénomènes de tendances locales (internes à chaque capteur) et globales (commune à l'ensemble des capteurs).

De plus, l'approche FBA cherche à généraliser QBA afin de traiter deux limites de l'approche QBA qui sont la détermination d'une échelle unique pour l'évaluation des positions des données (les indices de quantiles) et la considération de l'imprécision par regroupement en ensembles classiques (et donc strictes) ne permettant pas les appartenances partielles. FBA a pour objectif de :

— Composer avec des échelles variables de valeurs au cours du temps ;
— Gérer les données en considérant des appartenances partielles des données aux groupes.

7.3 Principes et hypothèses

Le principe général de notre proposition est de combiner les avantages de l'approche QBA avec ceux des approches floues présentées dans la section 5.2.3.

Afin de gérer l'imprécision de nos données et de permettre l'appartenance partielle des données, nous proposons de construire des groupes (clusters flous) sur lesquels seront fondés les positionnements de nos données. Les clusters flous permettent par définition à une donnée d'appartenir à plusieurs groupes avec un degré d'appartenance à chacun de ces groupes. Le positionnement de la donnée dans son ensemble se construit ensuite à partir de ces appartenances partielles aux différents groupes flous et des positions de ces différents groupes flous relativement aux autres. Ces positions prennent valeurs entre 0 et 1.

La méthode la plus connue de clustering flou est le fuzzy-c-means[Dun73] (les c-moyennes-floues). Cependant, cette méthode nécessite une valeur fixée du nombre de clusters flous à construire sur les données. Or nous souhaitons dans notre proposition alléger cette contrainte afin d'obtenir des échelles d'évaluation définies à chaque temporalité. Pour cela, nous proposons de construire des clusters flous selon les étapes suivantes :

1. Partitionnement ou regroupement par densité des données à chaque temporalité à l'aide d'une distance minimale, appelée d, séparant deux groupes [1] ;

2. Représentation de chaque donnée avec son imprécision à l'aide d'un ensemble flou (appelé i-set) ;

1. Soit une donnée appartenant à un groupe, si une autre donnée est à un distance inférieure à d alors elle fait partie du même groupe. Si une donnée est à une distance supérieure à d de tous les éléments du groupe alors elle fait partie d'un autre groupe.

	$vt^k_{(1)}$	$vt^k_{(2)}$	$vt^k_{(3)}$	$vt^k_{(3)}$	$vt^k_{(4)}$
VT^k	95	150	0	100	160

TABLE 7.1 – Exemple de série de données

3. Fuzzification des groupes fondée sur les différents ensembles flous qui composent les groupes.

L'appartenance partielle d'une donnée à un groupe est calculée par le ratio entre l'aire, formée par l'intersection entre la fonction d'appartenance du i-set et celle du groupe, et celle formée par la fonction d'appartenance du i-set.

Cette démarche à l'avantage de construire des positions sur lesquelles il est possible de calculer de manière analogue à la section précédente la variabilité et la stabilité. La seule différence est que le seuil sur les sauts (jp) devra être entre 0 et 1 car c'est l'espace de valeurs des positions.

7.4 Positionnement d'une donnée dans une série

Comme indiqué précédemment, l'indice de variabilité que nous proposons repose sur l'idée de prendre en compte l'imprécision des données, leur appartenance aux groupes de l'ensemble de données. Comme vu dans le chapitre précédent, chaque donnée peut être positionnée soit en fonction des données acquises par le même capteur (vision interne) soit par les données acquises au même moment par l'ensemble des capteurs (vision externe).

Ainsi, soit une variable d'étude v_i ($v_i \in V$), soit une temporalité t_k ($t_k \in T$), soit un capteur s_j ($s_j \in T$), l'ensemble des données dans lequel nous souhaitons positionner $v_i^{t_k}(s_j)$ à chaque t_k est :

— $\{v_i^{t_z}(s_j), z \in [1, m]\}$ pour la vision interne,

— $\{v_i^{t_k}(s_z), z \in [1, n]\}$ pour la vision externe.

Nous appellerons cet ensemble $VT^k = \{vt^k_1, ..., vt^k_l\}$ où l est le nombre d'éléments dans VT^k. Aussi, $l = m$ pour la vision interne, et $l = n$ pour la vision externe. $VT^{(k)} = \{vt^k_{(1)}, ..., vt^k_{(l)}\}$ est sa version ordonnée par ordre croissant.

Nous prendrons comme exemple illustratif de VT^k la série de données fournies dans le tableau 7.1 dans laquelle nous n'avons pas de données pour l'indice 3. Nous allons illustrer dans cette section notre approche par l'étude de la position de l'élément d'indice 4.

Comme indiqué précédemment, notre démarche pour le positionnement est constituée de quatre étapes, les trois premières formant la partie regroupement flou, la dernière pour le calcul de la position.

Pour le regroupement flou, les trois étapes sont :

— Groupement des données par densité à l'aide d'une distance d

— Représentation de chaque donnée par un intervalle flou de confiance utilisant une valeur γ d'écartement à la donnée.

— Fuzzification des groupes obtenus fondée sur les intervalles flous

Ainsi, notre approche pour le positionnement n'est plus dépendante du nombre de clusters (ou de quantile) mais de d la distance minimale séparant les clusters et de γ utilisé pour la fuzzification.

7.4.1 Groupement des données

Nous cherchons, dans un premier temps, à former des groupes par densité dans les données VT^k à l'instant t_k.

Soit d une valeur représentant la distance minimale souhaitée entre les groupes. Soit $\mathcal{C}$ l'ensemble des groupes recherchés. Le principe est le suivant : Soit une donnée $vt^k_{(z)}$ appartenant au cluster C_a d'indice a, si la valeur absolue de l'écart qui la sépare de $vt_{(z+1)}$ est inférieur à d alors $vt_{(z+1)}$ appartiendra à C_a sinon elle appartiendra à C_{a+1} (au groupe suivant).

$$\begin{aligned} vt^k_{(z)} \in C_a \text{ et } |vt^k_{(z)} - vt^k_{(z+1)}| \leq d \Longrightarrow vt^k_{(z+1)} \in C_a \\ vt^k_{(z)} \in C_a \text{ et } |vt^k_{(z)} - vt^k_{(z+1)}| > d \Longrightarrow vt^k_{(z+1)} \in C_{a+1} \end{aligned} \tag{7.1}$$

L'algorithme de construction des clusters est donc simple et est effectué en une passe (algo. 1). Cet algorithme simple à mettre en place est suffisant car il s'agit ici de trouver des groupes sur une seule variable ordonnée. Utiliser des algorithmes plus complexes comme DBSCAN par exemple n'est pas utile ici.

Data: VT^k, d
Result: $\mathcal{C}$: set of clusters
$\mathcal{C} = \{\}$;
$it = 2$;
$C_i = \{vt^k_{(1)}\}$;
while $it <= |VT^k|$ **do**
  if $|vt^k_{(it)} - vt^k_{(it-1)}| \leq d$ **then**
    $C_i = C_i \cup \{vt^k_{(it)}\}$;
    $it++$;
  else
    $\mathcal{C} = \mathcal{C} \cup \{C_i\}$;
    $C_i = \{vt^k_{(it)}\}$;
  end
end
$\mathcal{C} = \mathcal{C} \cup \{C_i\}$;

Algorithm 1: Algorithme de groupement par densité

Dans notre exemple, on obtient :

$$\begin{aligned} C_1 &= \{95, 100\} \\ C_2 &= \{150, 160\} \\ \mathcal{C} &= \{\{95, 100\}, \{150, 160\}\} \end{aligned} \tag{7.2}$$

7.4.2 Représentation des données avec leur imprécision

Nous cherchons maintenant à considérer les données avec leur imprécision. Pour cela, conformément à l'état de l'art, nous pouvons utiliser une représentation floue. Nous faisons le choix de représenter chaque donnée par une quantité floue qui peut être un nombre flou ou un intervalle flou.

La forme, le support et le noyau de l'intervalle flou sont dépendants de l'expertise sur les données. La principale contrainte logique est que le degré d'appartenance au nombre de la valeur initiale soit égal à 1.

En l'absence de ces informations, nous considérons une quantité floue ayant un support centré sur la valeur initiale dont l'étendue vaut $2 \times \gamma$, et dont le support est égal au noyau, i.e. le degré d'appartenance vaut 1 pour tout élément du support. Ainsi tout élément du support appartient totalement à la valeur floue. γ est un facteur d'extension de la donnée pour gérer son imprécision. Par simplification du paramétrage de l'approche, une seule valeur de γ est utilisée pour toutes les données.

Soit $\widetilde{VT}^k$ la série de quantités floues $\{\widetilde{vt}_1^k, ..., \widetilde{vt}_l^k\}$ associées aux données initiales. Alors, en suivant notre choix de modélisation.

$$\begin{aligned} &\forall z \in [1, l], \forall x \in support(\widetilde{vt}_z^k), \\ &\qquad \mu_{\widetilde{vt}_z^k}(x) = 1 \Leftrightarrow x \in noyau(\widetilde{vt}_z^k) \\ &\forall z \in [1, l], \forall x \notin support(\widetilde{vt}_z^k), \\ &\qquad \mu_{\widetilde{vt}_z^k}(x) = 0 \end{aligned} \tag{7.3}$$

Avec

$$\begin{aligned} support(\widetilde{vt}_z^k) = &[vt_z^k - \gamma, vt_z^k + \gamma] \text{ si } vt_z^k - \gamma \geq 0 \\ &[0, vt_z^k + \gamma] \text{ sinon.} \end{aligned} \tag{7.4}$$

La figure 7.1 montre les différentes fonctions d'appartenance pour nos données exemples.

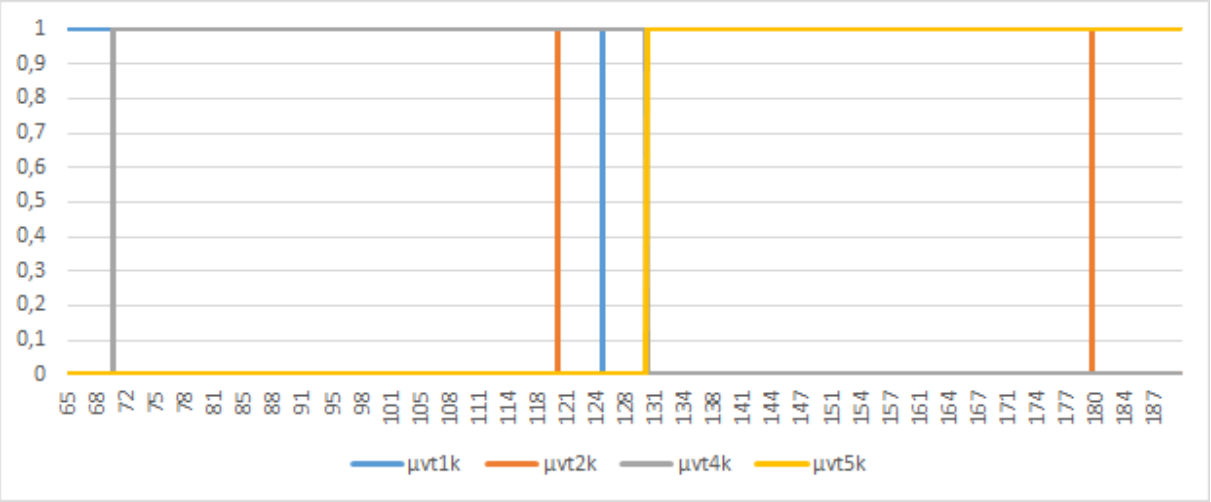


FIGURE 7.1 – Fonctions d'appartenance des éléments de $\widetilde{VT}^k$ pour l'exemple 7.1 avec $\gamma = 30$

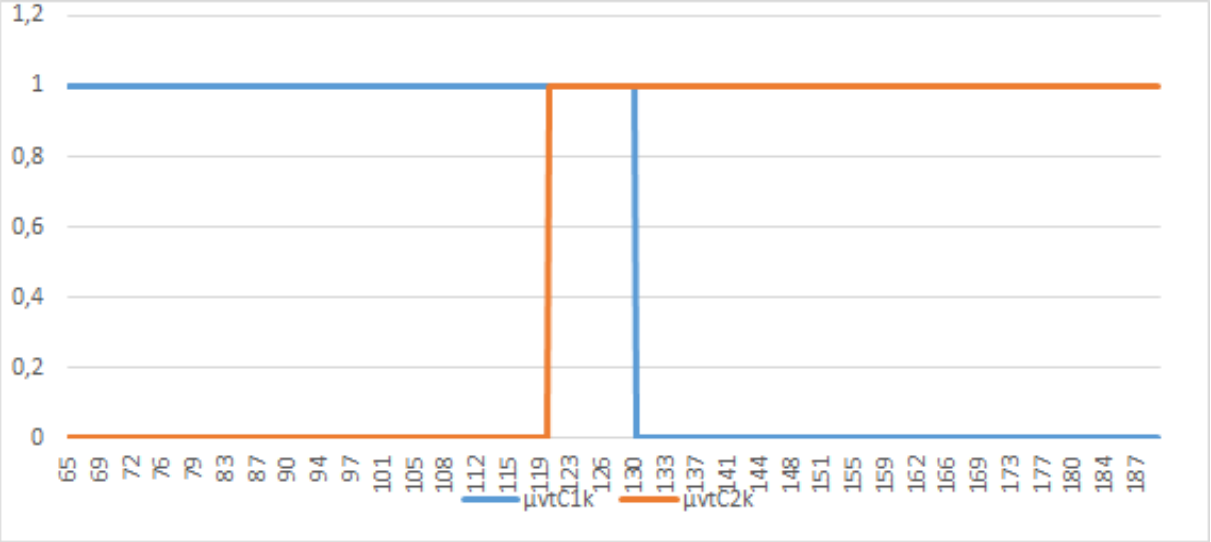


FIGURE 7.2 – $\widetilde{C_1}$ et $\widetilde{C_2}$ comme union des ensembles flous qui le compose.

7.4.3 Fuzzification des clusters

Comme nous avons d'un côté des clusters construits sur les valeurs initiales, et de l'autre des ensembles flous construits sur ces mêmes valeurs, nous pouvons considérer que les clusters flous associés aux clusters initiaux sont l'union des ensembles flous associées aux valeurs initiales. Soit $\widetilde{C_a}$ l'ensemble flou associé à C_a et $\mu_{\widetilde{C_a}}$ la fonction d'appartenance associée. Ainsi, cela donne, selon la t-conorme de Zadeh[Zad78], l'équation 7.5.

$$\forall x \in \mathbb{R}, \mu_{\widetilde{C_a}}(x) = max_{\{vt_z^k \in C_a\}}(\mu_{\widetilde{vt}_z^k}(x)). \tag{7.5}$$

où $\mathbb{R}$ est l'ensemble des réels.

Ce qui donne pour notre exemple, les fonctions d'appartenance présentées dans la figure 7.2.

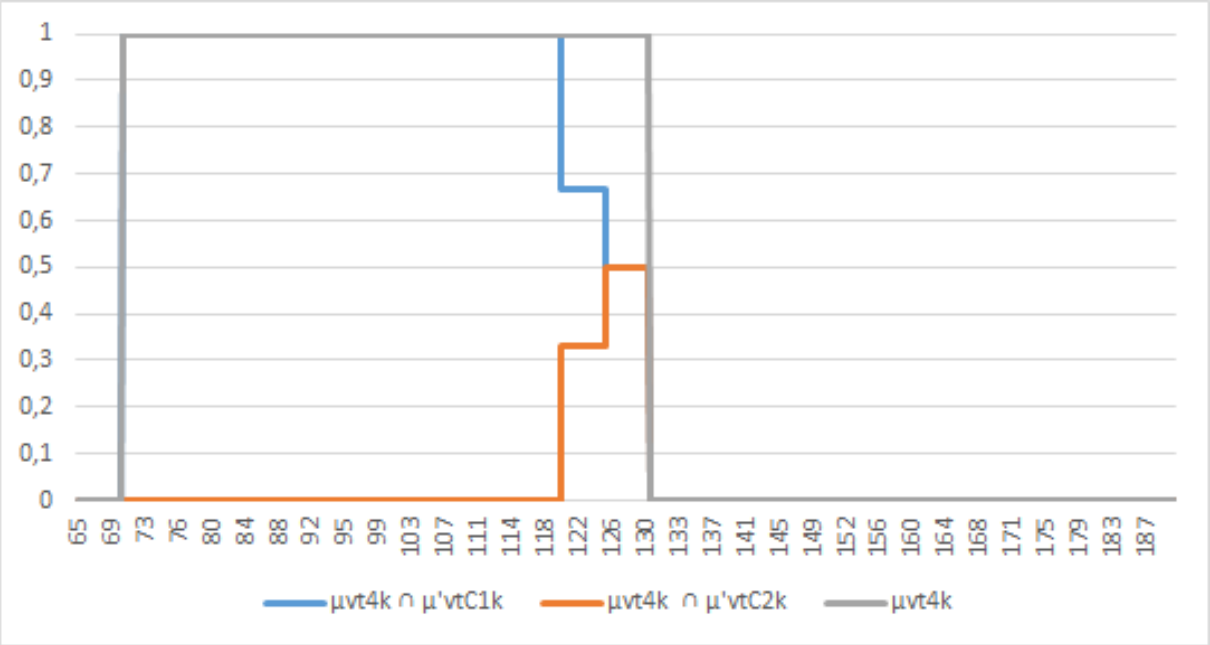


FIGURE 7.3 – Fonctions d'appartenance de $\widetilde{vt}_z^k$, $\widetilde{C_1} \cap \widetilde{vt}_z^k$ et $\widetilde{C_2} \cap \widetilde{vt}_z^k$.

Nous pouvons nous apercevoir que certains éléments peuvent appartenir pleinement à deux clusters flous. Aussi, afin de respecter les principes d'appartenance partielle (la somme des degrés d'appartenance doit au maximum être égale à 1) – principe exploité dans fuzzy-c-means [Dun73] – nous redéfinissons $\widetilde{C_a}$ à l'aide de l'équation 7.6. Dans ce contexte, nous considérons que le cluster flou est issu de la coopération de l'ensemble de ses membres.

$$\forall x \in \mathbb{R}, \mu_{\widetilde{C_a}}(x) = \frac{\sum_{\{vt_z^k \in C_a\}}(\mu_{\widetilde{vt}_z^k}(x))}{\sum_{\{vt_z^k \in VT^k\}}(\mu_{\widetilde{vt}_z^k}(x))}. \tag{7.6}$$

Si $\sum(\{mu_{\widetilde{vt}_z^k}(x), vt_z^k \in VT^k\})$ vaut 0 alors nous affectons 0 à la fonction d'appartenance afin de ne pas modifier son support par rapport à l'union.

Si, comme dans notre exemple, γ est plus grand que d alors nous pouvons avoir des chevauchements entre deux clusters flous successifs. C'est ce principe qui nous intéresse.

Soit $\mathcal{A}(\mu)$ l'aire sous la courbe d'une fonction d'appartenance μ. L'appartenance partielle d'un élément flou à un cluster flou est déterminée par le ratio entre l'aire de leur intersection et son aire (éq. 7.7). Les fonctions d'appartenance construites sont illustrées dans la figure 7.3).

$$\mu_{\widetilde{C_a}}(\widetilde{vt}_z^k) = \frac{\mathcal{A}(\widetilde{C_a} \cap \widetilde{vt}_z^k)}{\mathcal{A}(\widetilde{vt}_z^k)}. \tag{7.7}$$

Dans notre contexte,

— $\mathcal{A}(\widetilde{vt}_z^k) = 61$

— $\mathcal{A}(\widetilde{C_1} \cap \widetilde{vt}_z^k) = 56,33$

— $\mathcal{A}(\widetilde{C_2} \cap \widetilde{vt}_z^k) = 4,67$
— $\mu_{\widetilde{C_1}}(\widetilde{vt}_z^k) = 0,92$
— $\mu_{\widetilde{C_2}}(\widetilde{vt}_z^k) = 0.08$

7.4.4 Indice de positionnement

Nous cherchons ici à positionner une donnée v_j^k dans l'ensemble des données VT^k en considérant les données selon leurs appartenances partielles aux clusters flous.

L'indice de positionnement de vt_j^k à un instant donné k est défini à partir de l'appartenance de son ensemble flou associé $\widetilde{vt}_z^k$ au regard de son appartenance aux différents clusters flous $\widetilde{C_a}, C_a \in \mathcal{C}$ et d'un indice de position de ces clusters (ici donné par le rang maximal des éléments lui appartenant). Pour rappel, les clusters flous sont par construction ordonnés au regard de l'approche proposée dans [Run+10b].

L'indice de positionnement est défini dans l'équation 7.8.

$$\mathcal{P}_{vt_j^k} = \sum_{C_A \in \mathcal{C}} \mu_{\widetilde{C_a}}(\widetilde{vt}_z^k) * \frac{Rang(max(\{x, x \in C_a\}), VT^k)}{l} \tag{7.8}$$

Prenons notre exemple précédent, où l'on souhaite positionner vt_4^k.

$$\begin{aligned}\mathcal{P}_{vt_4^k} &= \mu_{\widetilde{C_1}}(\widetilde{vt}_4^k) * \frac{Rang(max(\{x, x \in C_1\}), VT^k)}{4} + \mu_{\widetilde{C_2}}(\widetilde{vt}_4^k) * \frac{Rang(max(\{x, x \in C_2\}), VT^k)}{4} \\ &= 0,92 * \frac{2}{4} + 0,08 * \frac{4}{4} \\ &= 0,54\end{aligned} \tag{7.9}$$

$0,54$ indique que la donnée est possiblement légèrement dans la partie supérieure de l'ensemble des valeurs VT^k en considérant sa possible imprécision.

Les données valant 0 ou NULL sont exclues de l'ensemble du processus précédent. Leur position par défaut vaut 0, pour les mêmes raisons que la valeur Q_0 leur avait été affectée dans le chapitre précédent.

À l'aide de ces indices de positions il est désormais possible de calculer la variabilité du processus.

7.5 Variabilité

7.5.1 Indicateurs de variabilité

Considérons désormais les séries temporelles des indices de positionnement au cours du temps T des données d'un capteur s_j pour une variable i calculées selon les sections précédentes. Soit $VT^T = \{vt_j^1, ..., vt_j^m\}$ le vecteur correspondant.

Sur ces indices de positionnement, nous pouvons calculer, comme pour toute série de données, des indicateurs de position de la série, par exemple la moyenne, et des indicateurs de dispersion comme la variance.

Nous pouvons aussi exploiter la variabilité définie dans le chapitre précédent de la manière suivante. Nous conservons les deux paramètres jp et b mais ici jp appartient à l'intervalle $[0, 1]$, car c'est l'espace des valeurs de l'indice de position déterminés dans la section précédente.

$$Sc_{ext}(i,k,j) = \begin{cases} 1, \text{si} & |P_{vt_j^k} - P_{vt_j^{k-1}}| \geq j_p \\ & \text{ou } P_{vt_j^{k-1}} > 0 \text{ et } P_{vt_j^k} = 0 \\ & \text{ou } P_{vt_j^{k-1}} = 0 \text{ et } P_{vt_j^k} > 0 \text{ et } cip \leq b \\ 0, & \text{sinon} \end{cases} \tag{7.10}$$

À l'instar de l'indicateur construit dans l'approche QBA, la variabilité (par exemple externe) de la série étudiée sur une période $T' = \{t_k, ..., t_{k+z}\}$ est définie par l'équation (7.11).

$$Var_{ext}(i,k,k+z,j) = Var_{ext}(i,T',j) = \frac{\sum_{x=k}^{x=t+z} Sc_{ext}(i,x,j)}{z - IPMD(i,j)} \tag{7.11}$$

où $IPMD(i,j)$ est le nombre d'inter-périodes sans données comprises dans des séquences consécutives sans données de longueur strictement supérieure à b.

Le même opérateur de stabilité peut être utilisé, bien qu'il ne sera exploité dans la suite. Nous nous concentrons dans la suite sur la variabilité externe.

7.5.2 Exemple

Dans la Figure 7.4, nous fournissons des informations permettant de déterminer le score de variabilité de l'élément (β) marqué en rouge au fil du temps.

La table 7.2 présente des exemples de résultats des valeurs de l'indice de positionnement externe de la série β.

Par exemple, la valeur de l'indice de positionnement pour t_2 est calculée comme suit :

$$\mathcal{P}_\beta^{t_2} = 0 * \frac{2}{6} + 0.8 * \frac{4}{6} + 0.2 * \frac{6}{6} \approx 0.73$$

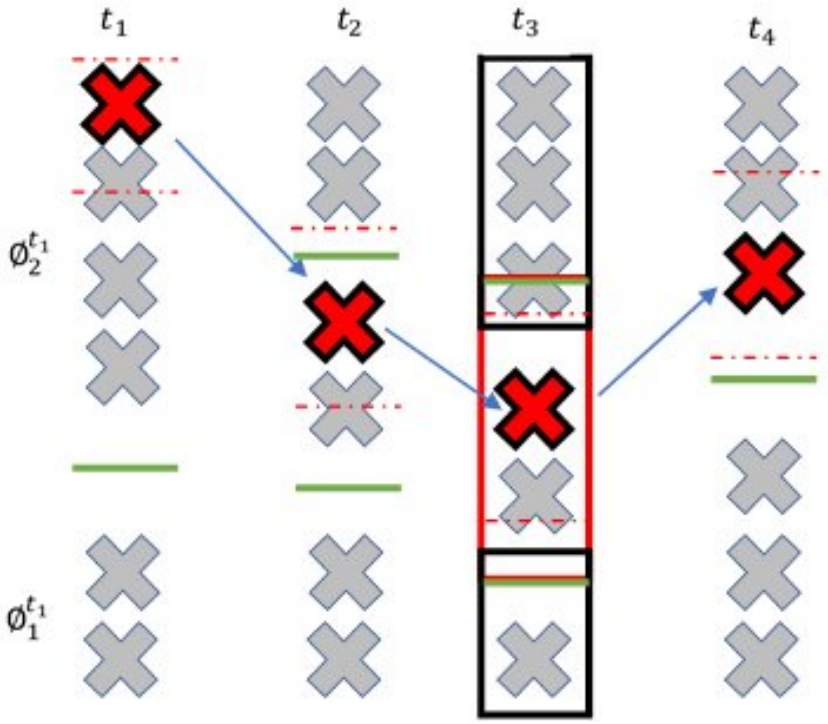


FIGURE 7.4 – Comportement de β sur une période de temps T

t_1	t_2	t_3	t_4
0	0.2	0.1	1
1	0.8	0.9	0
	0	0	

a

T	t_1	t_2	t_3	t_4
$\mathcal{P}_\beta^t$	1	0.73	0.55	1

b

TABLE 7.2 – (a) $MatFD_\beta^T$: Matrice des degrés d'appartenance de β aux clusters flous au cours de T. (b) $VecPos_\beta^T$: Vecteur des valeurs de l'indice de position de β au cours du temps.

En fonction de ces positions, nous pouvons calculer la moyenne et la variance :

$$\mathcal{M}_\beta^T \approx 0.82$$

$$\mathcal{V}_\beta^T \approx 0.22$$

Nous pouvons aussi calculer la variabilité en testant sur plusieurs valeurs de jp (avec $b = 1$) :

— si $jp = 0,1$, alors $Var_{ext}\beta^T = 1$,
— si $jp = 0,25$, alors $Var_{ext}\beta^T = 0.66$,
— si $jp = 0,33$, alors $Var_{ext}\beta^T = 0.33$,
— si $jp = 0,5$, alors $Var_{ext}\beta^T = 0$,

Ainsi, à l'instar de ce que nous avons observé dans la section 6.6, quand jp diminue la variabilité augmente. Nous ne referons, donc, pas dans ce chapitre d'analyse de sensibilité pour jp et b.

Cette approche fournit des scores relatifs pour mesurer le mouvement externe des données (c'est-à-dire que la volatilité du mouvement des données respecte la normalité du reste des ensembles appartenant au même domaine). Les scores générés permettent de détecter le comportement caché et significatif des données. La fusion des valeurs de l'indice de position temporel d'un élément (la variabilité) fournit un niveau de collecte qui représente des mouvements singuliers.

7.6 Expérimentation

Nous présentons dans cette section une expérimentation menée sur une base de 700 capteurs et sur une période de $T = 180$ jours. Les indicateurs ont été calculés par mois. Les valeurs figurant sur le tableau 7.3 représentent la moyenne des indices calculés par jour.

Nous comparons par une première expérience, la proposition actuelle FBA avec l'approche précédente QBA. Nous avons choisi $d = 50$ comme distance d'éloignement entre les clusters et γ égale à l'écart type des valeurs journalières. Pour l'approche QBA, le nombre de quantiles a été défini à 4 (quartiles), le paramètre pour la hauteur des sauts $jp = 2$, le paramètre de pause $b = 3$.

Pour cette première expérience, nous prenons la variance comme étant un indicateur de mesure de la variabilité pour la deuxième approche. Nous remarquons que le calcul de la variabilité par les deux approches suit la même trajectoire et a une même tendance. Le score de corrélation des deux approches est de 0.67 pour cet ensemble de données. En ajustant les paramètres, les deux courbes peuvent se rapprocher plus. Ceci nous laisse constater en premier lieu que les deux approches sont cohérentes en terme d'évaluation des séries temporelles de données imparfaites.

Par ailleurs, nous avons aussi effectué une autre expérience pour étudier les valeurs de variabilité données par d'autres approches et les comparer à la notre. Ceci nous permet de comprendre si nos jugements sur les flux de données temporelles imparfaites suivent la même logique que pour les autres approches.

Le tableau 7.3 compare les résultats de diverses approches en termes de mesures de variabilité. On a utilisé le même jeu de données que précédemment. Nous avons appliqué premièrement la variance $\mathcal{V}$ (FBA) sur le résultat du jeu de données. Ceci nous permet d'avoir des mesures de variabilité mettant en œuvre la dispersion des valeurs. Ensuite, nous avons calculé la variabilité. Nous avons étudié l'influence du paramètre jp dans ses valeurs. Nous avons enfin comparé notre indice avec une approche floue [CD09] et

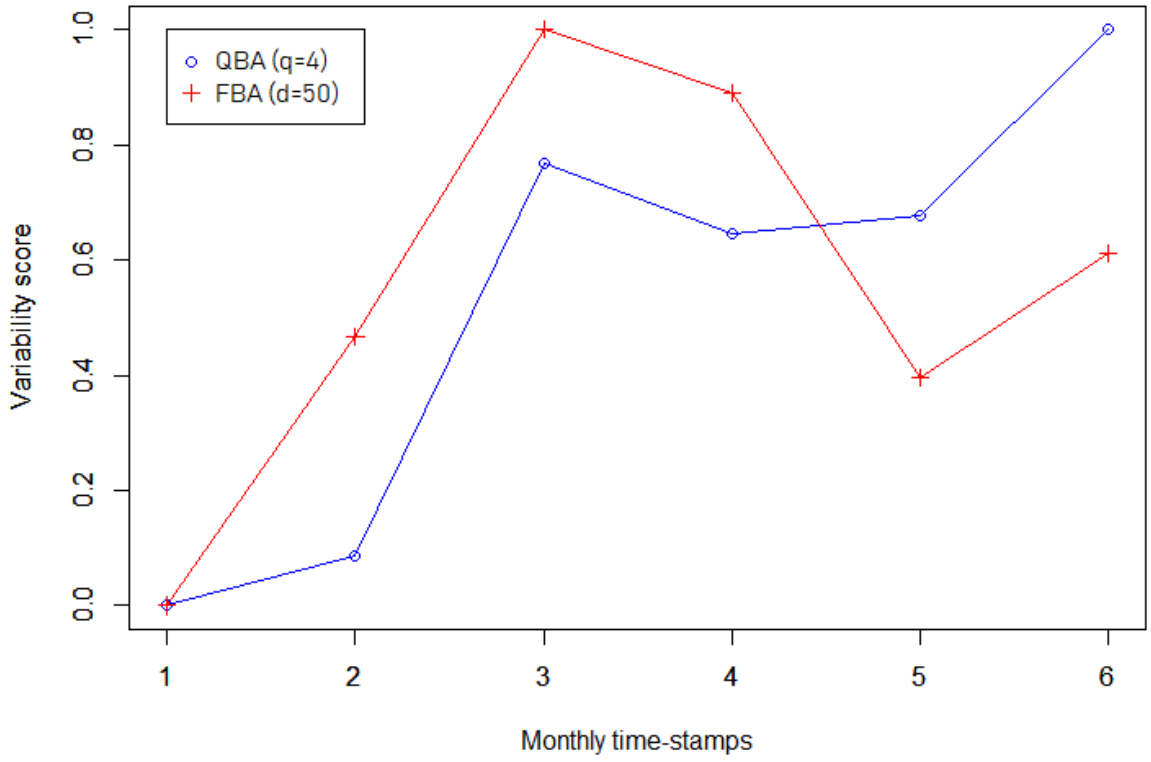


FIGURE 7.5 – Comparaison de la consistance de trajectoire entre FBA et QBA

une approche statistique [Fel15]. Toutes les approches étudiées adoptent la même vision d'étude, i.e., elles étudient la variabilité d'un ensemble de données (dans notre cas, les données issues de capteurs) par rapport au reste, en d'autre terme "la variabilité externe".

Le tableau 7.3 confirme que même par l'approche FBA, en diminuant jp la variabilité augmente. Par contre la convergence de cette algorithme vers 1 est plus lente que QBA, et ce, parce que $jp \in [0, 1]$ et $Var_{ext}(i, T', j) \in [0, 1]$, ainsi, un choix de jp va couvrir plus d'écarts entres les valeurs. Cet écart fait diminuer la variabilité et ralentir la convergence de FBA. Les expérimentations ont aussi montrées que l'utilisation de cette approche n'est pas rentable si la période d'étude est minime, ceci s'explique par le fait que jp va couvrir l'ensemble des écarts, ce qui rend le résultat final non significatif.

Pour vérifier si le mécanisme de calcul de la variabilité par FBA suit la majorité des approches citées dans le tableau précédent, nous avons étudié la dispersion des scores de variabilité [Sco15] par les différentes approches (voir figure 7.6).

Nous avons vu expérimentalement qu'en variant jp, nous pouvons obtenir des résultats qui se rapprochent des résultats de [CD09] et [Fel15]. Les tests confirment que le calcul de la variabilité par FBA est valide.

t_i	t_1	t_2	t_3	t_4	t_5	t_6
$\mathcal{V}$ (FBA)	0.0911	0.1146	6.0610	6.2081	0.0869	0.0814
Var(jp=0.33 & b =3)	0.42857	0.6000	0.4333	0.6428	0.3666	0.3548
Var(jp=0.25 & b =3)	0.42857	0.7333	0.4666	0.7142	0.354838	0.3548
Var(jp=0.10 & b =3)	0.60714	0.7666	0.7333	0.7857	0.7000	0.51612
Var(jp=0.05 & b =3)	0.71428	0.8333	0.8333	0.8571	0.612903	0.612903
Fuzzy [CD09]	0.01025	0.01296	0.6855	0.7021	0.009881	0.009233
Stat [Fel15]	90.60000	99.0333	83.8387	31.0967	13.6333	19.8709

TABLE 7.3 – Comparaison de la variabilité calculé par FBA vis à vis d'autres approches sur un T=12 mois et 700 medias

La figure 7.6 montre bien, qu'avec $jp = 0.33$ et $jp = 0.25$, la densité des scores de variabilité se rapproche de la densité de [CD09] avec des indices de similarité en appliquant Pearson de 0.91287 et 0.8944 respectivement. Ainsi, si on diminue encore ce paramètre, e.g. $jp = 0.10$ ou $jp = 0.05$, les courbes changent de trajectoires contrairement à la majorité, e.g., pour $jp = 0.05$ la similarité à [DHP03] s'affaiblie et devient de 0.8666. Ceci nous laisse situer l'intervalle surlequel le paramètre jp optimal peut exister.

Plusieurs méthodes de recherche des paramètres optimaux sont proposées dans la littérature, mise à part celle utilisée dans le chapitre précédent, dans cette partie nous n'allons pas chercher l'optimalité. Nous laissons ce travail en perspective de cette thèse.

Nous avons vérifié par la démarche précédente que cette approche est capable de donner des indices de variabilité significatifs. Par ailleurs, les expérimentations sur d'autres jeux de données ont confirmées la capacité de cette méthode à fournir des indicateurs précis. Les scores de variabilité obtenus par cette approche sont cohérents et que le système possède des caractéristiques solides permettant de juger la volatilité des flux de données imparfaits.

7.7 Discussion

Nous proposons dans cette section de revenir sur les paramètres de l'approche.

La variabilité, définie dans ce chapitre, fonctionnant sur les mêmes principes (hauteur du saut, période de pause) que ceux de l'approche QBA, la sensibilité de ces deux approches aux paramètres (jp,b) ont un comportement analogue : plus jp augmente plus la variabilité diminue, plus b augmente plus elle augmente.

En ce qui concerne les paramètres d et γ de la phase de fuzzification, nous proposons l'analyse suivante.

Le fait de découper en clusters (flous) sans prédéterminer leur nombre donne de la souplesse au procédé, mais complexifie grandement une étude fine de la sensibilité à ces paramètres. Cette étude est une des perspectives de cette thèse. Cependant, nous pouvons tout de même donner les grands principes.

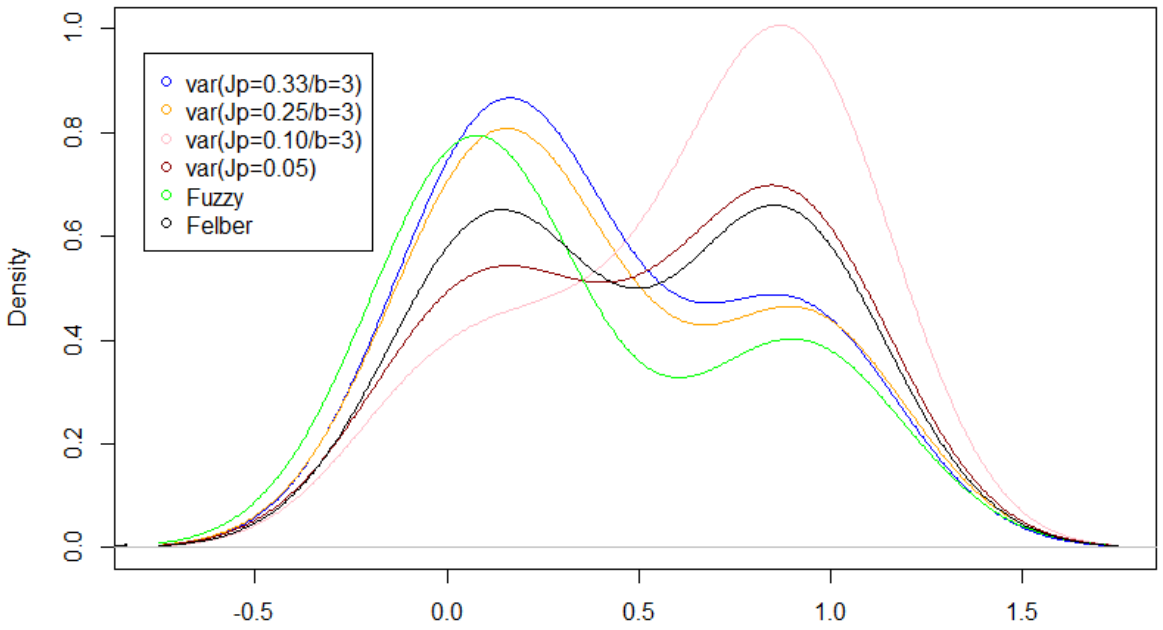


FIGURE 7.6 – Dispersion des scores de variabilité

Plus d est petit, plus le nombre de groupes sera grand. Pour ce qui est de l'impact sur la position et la variabilité, plus d est petit, plus grand sera le nombre de positions possibles ce qui, potentiellement en fonction de la valeur de jp, peut augmenter la variabilité.

Par ailleurs, si γ est inférieur à d alors il n'y a pas de chevauchement, et donc nous nous retrouvons avec des clusters classiques. Si γ est grand, les chevauchements seront possiblement importants, ce qui donne des nuances dans les positions des données. Aussi, la valeur de jp aura une grande influence. Nous avons utilisé un paramètre pour la fuzzification des données, nous aurions pu envisager des approches plus complexes, ou laisser à l'expert le soin de décider. Cette dernière solution est peu réaliste du fait du volume de données à traiter.

Par ailleurs, le calcul du positionnement selon la vision interne revient à déterminer la position de chaque donnée récoltée de la série vis-à-vis des autres acquises par le même capteur. Les positions au cours du temps sont toutes calculées sur le même vecteur. De manière schématique, ce vecteur est répété à chaque timestamp, et nous aurons la même découpe en clusters quel que soit le timestamp. C'est pourquoi nous ne l'avons pas détaillé dans les exemples.

7.8 Conclusion

Afin de lever les problèmes liés à la représentation par ensemble classique de l'approche QBA, nous avons proposé dans ce chapitre une approche fondée sur le flou.

Dans cette approche, nous proposons une classification floue par densité reposant sur deux paramètres : un paramètre de distance inter-clusters et un paramètre pour la fuzzification. À l'aide des différentes opérations effectuées, nous construisons un indice de position de chaque donnée dans son échantillon. À partir des valeurs de cet indice au cours du temps, nous pouvons étudier la variabilité des éléments. Nos différents exemples et expérimentations montrent la cohérence de notre démarche.

Enfin, nos propositions construisent un indice par variable. L'intérêt est certain pour suivre l'activité par l'intermédiaire d'un outil avec des tableaux de bords distincts. Autrement dit, cette solution est adaptée dans le contexte d'une étude dimension par dimension, et non multivariée. La présentation de cet outil est l'objet de la partie suivante.

Aussi, nous envisageons en perspective de travailler sur un indice de positionnement multi-varié en utilisant des méthodes de clustering flou en adéquation.

Quatrième partie

Visualisation interactive pour la gestion de la qualité de la récolte de données issues de capteurs sur le web

Chapitre 8

Etat de l'art sur la visualisation de flux de données imparfaites

8.1 Introduction

La visualisation de données et l'exploration visuelle qu'elle permet peuvent aider à mieux comprendre les données.

Dans ce chapitre, nous présentons différents travaux sur l'analyse visuelle dont nous nous sommes inspirés afin de construire l'outil présenté dans le chapitre suivant. Nous aborderons aussi les différentes formes de visualisation, à l'égard de l'interactivité et de l'acheminement des idées (visual thinking). Ces techniques ont été adoptées dans notre troisième contribution. Par ailleurs, nous étudierons la visualisation des données imparfaites. Enfin, des exemples d'outils issus de la recherche seront présentés.

8.2 Analyse visuelle

Thomas et Cook [TC06] définissent l'analyse visuelle comme étant la science du raisonnement analytique facilitée par des interfaces visuelles interactives. Elle combine des techniques d'analyse automatisées avec des outils interactifs. Elle peut faciliter la compréhension du comportement des objets, e.g signaux des capteurs, circulation des données. Elle nous permet, par exemple, par une simple visualisation de remarquer facilement des phénomènes, des incohérences, des aberrations etc. Ces phénomènes peuvent par exemple être une apparition saisonnière d'un pattern précis, des points extrêmes remarquables, etc.

Keim et al.[Kei+08a] considère que l'objectif de l'analyse visuelle est de transformer les informations en une opportunité. Le but est ainsi de rendre l'explication des données et des informations plus simple tout en utilisant un discours analytique.

La figure 8.1, proposée par Keim et al. dans [Kei+08a] montre les processus nécessaires qui doivent exister dans un projet d'analyse visuelle pour améliorer les connaissances sur un sujet particulier.

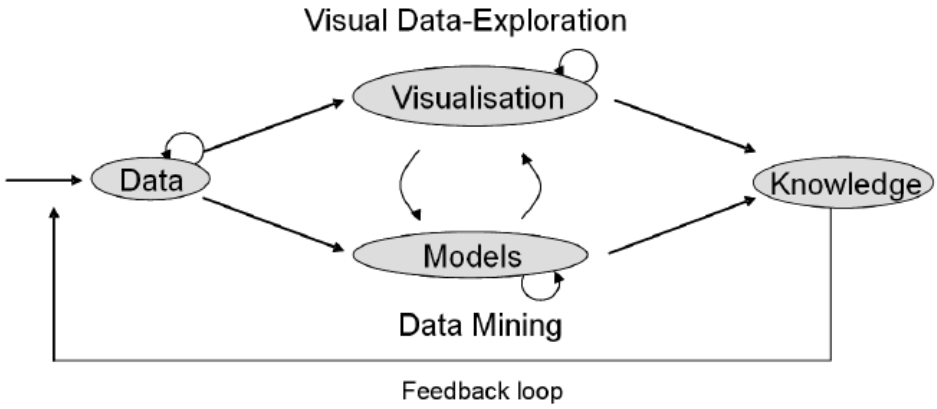


FIGURE 8.1 – Étapes d'une analyse visuelle pour un support décisionnel interactif et évolutif [Kei+08b]

Le système d'analyse visuelle proposé repose sur 4 composants : la collecte de données, le pré-traitement de données, le requêtage de données et l'analyse de données. Pour chacun de ces composants, une technique de visualisation analytique est appliquée. Par exemple, à l'issue de la détection des points atypiques dans une phase de pré-traitement de données, les données peuvent être visuellement distinguées. Il en va de même pour l'agrégation et le regroupement dans la phase d'analyse. L'utilisation des techniques de visualisations interactives peuvent améliorer certaines compréhensions sur les données, notamment en changeant les paramètres et les dimensions, permettant ainsi d'avoir des visions différentes des données.

Il existe diverses techniques permettant d'améliorer la compréhension des données reposant sur l'œil nu. Dans ce qui suit, nous en aborderons quelques unes.

8.2.1 Où placer des informations visuelles importantes ?

Pour améliorer le champs visuel de reconnaissance, Olshannikova et al. [OOK14] proposent une méthodologie de visualisation des données fondée sur la concentration direct de l'œil sur des champs spécifiques, comme, par exemple, le centre d'un graphique. La méthode utilisée consiste à grouper tous les angles de vision à fort intérêt pour l'utilisateur (déduits de ses interactions quotidiennes avec un objet visuel par exemple), puis à construire des zones visuelles graduées. Olshannikova et al. [OOK14] indiquent que la partie centrale est la zone la plus informative. Conceptuellement, elle représente la partie essentielle du champs visuel de l'utilisateur et, par conséquent, elle doit contenir le message important à passer (voir figure 8.2).

Ware [WKP14] trouve que la méthode de Olshannikova et al. [OOK14] nécessite de mettre en amont une série d'actions liées à l'attention de l'utilisateur. Ces actions doivent obliger les yeux à bouger et ajustent la concentration sur les champs d'intérêt. Cette action est appelée "la requête visuelle".

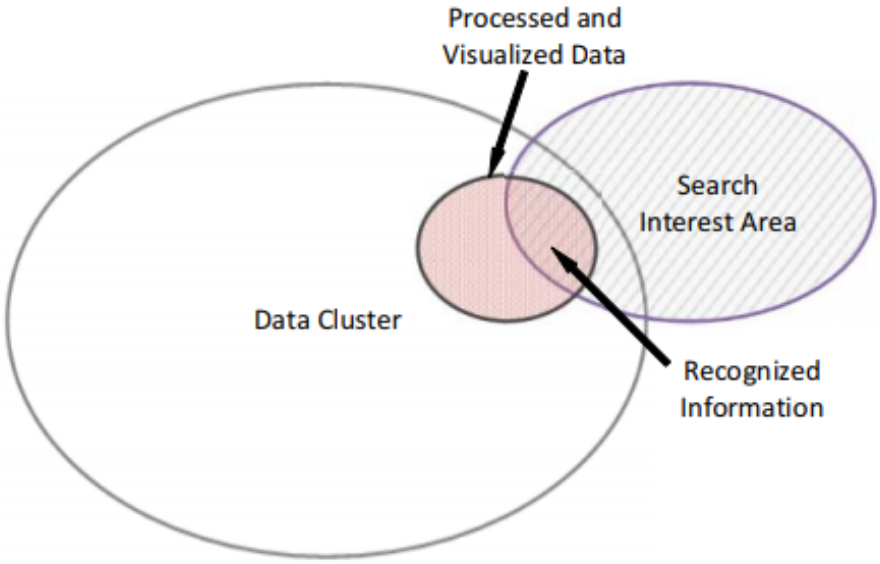


FIGURE 8.2 – Zone où le champs visuel est important [OOK14]

8.2.2 Comment visualiser les données multi-variées ?

L'exploration visuelle de données multivariées est un contexte important dans les études de visualisation des données. Il s'agit d'un des domaines de recherche scientifique qui visent à simplifier les données complexes i.e les données à divers dimensions [Aub+03]. Pour aboutir à une simplification des problèmes à traiter, il existe de nombreuses techniques de visualisations multivariées. Parmi les plus classiques se trouvent :

— Les **Scatter plots** (voir figure 8.3) sont des tableaux de panneaux présentant des diagrammes de dispersion adjacents d'un jeu de données multivarié.
— Les **Coordonnées parallèles** (voir figure 8.4) : Visualisation à base des axes parallèles pour tracer un jeu de données multivariées.

Quant à la recherche scientifique, Wong et Bergeron [WB97] présentent une méthode d'exploration de données multivariées via un aperçu de données sur des petites dimensions. L'approche consiste à une réduction d'échelle des composants en créant d'autres plus réduites par rapport aux composants principaux.

D'autres approches ont aussi été proposées comme les approches orientées-pixel [Kei00 ; BHL05] ou les visualisations par cartes de Kohonen. Certaines ont même été adaptées aux données floues [Run+08 ; Run+10a]. Afin d'obtenir des résumés visuels des données, des approches récentes portent sur le clustering visuel et interactif [Bou+16] ou encore sur la visualisation de flux de données à l'aide du subspace clustering [1] [LBT17].

1. Clustering effectué dans un sous-espace de dimensionalité réduite

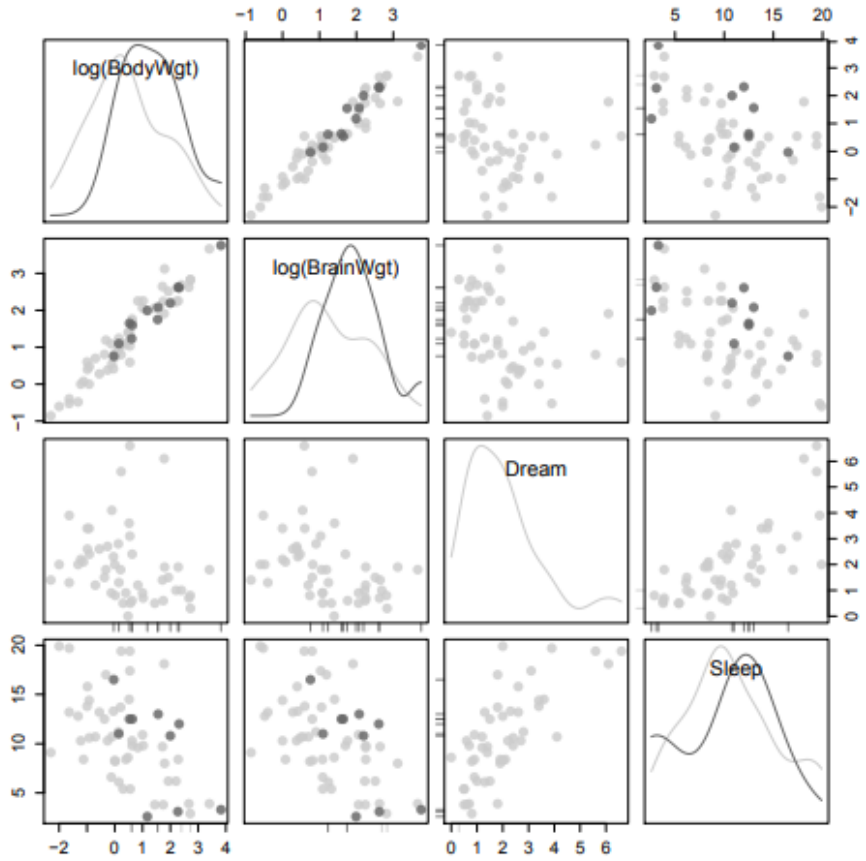


FIGURE 8.3 – Visualisation multivariée en utilisant Scatter plot

Cependant, Koo et al.[Koo+06] trouvent que les vues réduites ou filtrées peuvent déformer les données d'origines, ce qui donne une mauvaise interprétation des données. Ainsi, pour fournir à l'utilisateur des données précises, plusieurs angles et échelles, doivent être analysés.

Koo et al. [Koo+06] ont développé un environnement de visualisation de données multi-capteurs. Ils ont utilisé une méthode de fusion de données intégrant des graphiques et métriques. Leur système introduit des concepts visuels de complémentarité entre les indicateurs, et ce, pour unifier les connaissances sur un aspect particulier.

8.2.3 Comment visualiser les méta-informations ?

Les méta-informations ont été définies dans [AR14] comme étant des caractéristiques ou des qualifications des informations aidant à la prise de décision et proposant une vision générique sur des problématiques précises.

Par exemple, dans la classification de données venant de sources différentes, Sean L. Guarino et al. [Gua+09] définissent les méta-informations comme étant des vues sur l'incertitude, l'ambiguïté, la fiabilité de la source, la pertinence, le manque d'information, etc., autrement dit des nouvelles informations valorisant l'information de base.

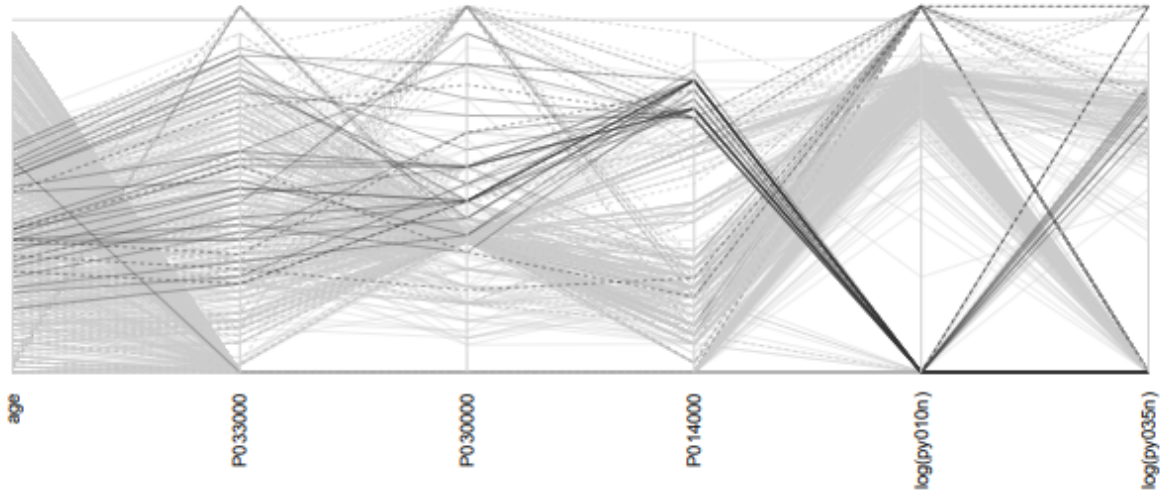


FIGURE 8.4 – Visualisation multivariée en utilisant les coordonnées parallèles [Oca85]

Différentes représentations des méta-informations ont été abordées dans la littérature. La plupart concernent la visualisation des données imparfaites [Mac+12]. A.M. MacEachren et al. [Mac+12] posent la question suivante : comment savoir si la visualisation de l'incertitude, et plus généralement les méta-informations, influent sur le raisonnement et la prise de décision dans des contextes visuels ?

Ces études comparent trois techniques principales de visualisation des données manquantes comme étant un facteur majeur engendrant l'incertitude dans la compréhension des données : La non-existence des données, la représentation floue des données manquantes et la complétude probable du vide par des études tendancielles.

L'effet de ces trois types de visualisation est évalué ensuite sur la base d'un score de confiance donné par un expert de visualisation. Il s'agit d'un score d'appréciation des résultats finaux mettant en œuvre des questions sur le risque de la liaison déterminée, la certitude envers cette détermination, etc. Cette comparaison permet de favoriser une représentation par rapport à une autre. Ainsi ces scores d'appréciation permettent de recommander certaines visualisations à adopter dans les tableaux de bord.

8.3 Technique de visualisation

8.3.1 Interactivité visuelle

Dans la littérature, plusieurs définitions de l'interactivité visuelle sont présentes. De ce fait, trouver une définition unique de cet aspect est difficile. D'un point de vue général, [He+07], décrivent simplement l'interaction comme "la communication entre l'utilisateur et le système" [WB04]. Becker et al. dans [BCW87] la définissent comme une manipulation

d'une direction. Beaudouin-Lafon [Bea04], trouve que le fait qu'une interaction peut se produire même avec une image statique est aussi une interaction visuelle, du fait que la personne peut comprendre différents sujets du messages passé par le peintre.

La définition technique de ce terme est différent. Dans un contexte numérique, Foley et al. [Fol+96] unifient les deux points de vue de Dix et al. [He+07] et Becker et al. [BCW87], ils trouvent que l'interaction visuelle est une technique d'IHM (interaction homme machine) qui a comme objectif d'effectuer une tache générique de communication.

Les techniques d'interactions visuelles sont utiles pour une meilleure compréhension des besoins des utilisateurs. Certains essaient de les catégoriser selon différents niveaux, i.e. bas niveau [BCS96] (sélection des variables,changement d'échelle, rotation, etc.), haut niveau (filtrer sur une catégorie [Rin+13] ou selon des dimensions [Twe97]). Toutes ces classifications permettent d'avoir différents points de vue sur l'interaction et la bonne pratique de son application.

D'autres travaux se concentrent sur la description des tâches utilisateurs et de leurs possibles comportements lors de l'interaction avec un système [AES05]. Pour cela, des représentations en cycles, présentées dans [Nor02], décrivent l'interaction dans un contexte formel, et ce, en utilisant plusieurs étapes : définition de l'objectif, formation de l'intention, spécification d'une action, exécution de l'action, interprétation du nouvel état et évaluation des résultats.

Afin d'exploiter au mieux l'interaction visuelle, Figueiras [Fig15] considère qu'une visualisation interactive peut comporter 11 procédés différents (cf. tableau 8.1). Ainsi, la phase de conception d'un outil visuel et interactif doit inclure une étude sur les procédés à mettre en œuvre.

Filtrer	Montrer uniquement les données sélectionnées
Sélectionner	Marquer ou suivre les éléments intéressants
Résumer - Élaborer	Ajuster le niveau d'abstraction des données
Donner un aperçu et explorer	Zoomer et filtrer, puis détailler la demande
Connecter - Mettre en relation	Montrer comment les données sont liées
Archiver	Retracer les étapes de l'exploration des données
extraction de fonctionnalités	Extraire les données d'intérêt
Reconfigurer	organiser les données
Encoder	Donner une représentation différente des données
Participer - Collaborer	Contribuer à l'évaluation des données
Gamification	Montrer les données de manière plus ludique

TABLE 8.1 – Techniques de l'interactivité visuelle [Fig15]

Visualisation des relations d'inter-connections

Il existe diverses techniques permettant de visualiser les relations entre les données. Ceci peut être intéressant pour déterminer les relations entre les attributs. Cela explique certains aspects cachés par l'application du principe de l'interactivité visuelle.

Ces relations peuvent être montrées en soulignant des liens entre les éléments déjà représentés dans la visualisation ou même en montrant des éléments qui sont pertinents pour un utilisateur donné. Selon Craft et Cairns [CC05] la découverte de relations est particulièrement importante lorsque des comparaisons entre les caractéristiques de différents objets sont manipulées.

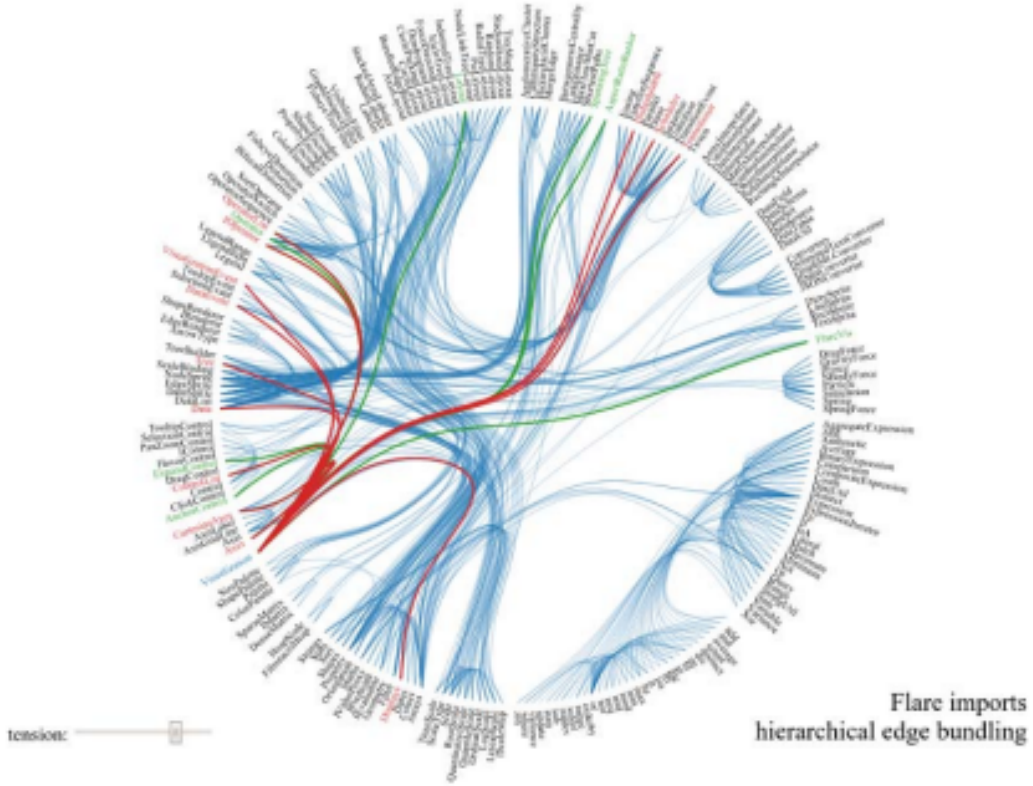


FIGURE 8.5 – Visualisation des inter-relations [Fig15]

À titre d'exemple, sur la figure 8.5, l'utilisateur peut suivre les données qui l'intéressent en cliquant sur des endroits spécifiques. Ensuite, les relations qui apparaissent seront mises en valeur, i.e. une apparition plus claire (voir la couleur rouge sur la figure) des relations possibles entre un attribut sélectionné et les autres.

En effet, même si la couleur aide à trouver les données d'intérêt dans les différentes vues, Figueiras [Fig15] trouve qu'il est difficile pour un utilisateur de faire des comparaisons s'il n'est pas en mesure de distinguer toutes les relations. Pour cela, ce type de visualisation doit être complété par des visualisations montrant les interactions par des illustrations complémentaires.

8.3.2 La pensée visuelle

Pour construire une idée à partir d'un ensemble de visualisations, la pensée visuelle est une technique spécifique qui permet d'ordonnancer les idées afin de trouver une nouvelle information ou approfondir une idée.

En d'autres termes, la pensée visuelle est l'action de rendre les échanges et les idées visuels pour aider à comprendre certaines logiques et/ou pensées. Elle représente un moyen d'organiser les pensées et d'améliorer la capacité à penser et à communiquer.

L'importance de la pensée visuelle est liée à l'idée de mettre à disposition de l'utilisateur tous les outils nécessaires pour exploiter une idée. Cet aspect est présent dans la conception des tableaux de bords de l'outil que nous avons développé pour la société.

En effet, dessiner, ou visualiser permet d'aider à trouver les idées facilement. Cette technique adopte le principe que si les idées ne peuvent pas être dessinées, elles ne peuvent pas être reprises facilement. Elle est donc une compétence essentielle pour développer de nouvelles idées et conceptions, communiquer efficacement ses idées et collaborer avec d'autres pour les concrétiser.

Cain [Cai19] affirme que l'objectif de ce concept est de passer un message de communication en mettant l'accent sur l'acheminement des idées (voir Figure 8.6). Un outil adoptant le principe de la pensée visuelle, doit donc, mettre en évidence les objets nécessaires qui orientent la recherche d'un utilisateur afin de trouver son souhait.

FIGURE 8.6 – Le concept du visual thinking

La pensée visuelle doit reposer sur des techniques de visualisation qui attirent l'observation et centralisent les intérêts de l'utilisateur. La technique du story telling [FBY05], ou raconter des histoires à partir des données, est un principe issu de cette technique. Ce dernier a pour objectif de relier les indicateurs complémentaires afin de construire une idée sur un sujet particulier.

8.4 Visualisation de données imparfaites

Nous nous intéressons ici à la représentation de l'incomplet puis de l'imprécis. Différents états de l'art dédiés à la visualisation de l'imperfection ou l'incertitude ont été proposés dans la littérature [Zuk08 ; Bon+14]. L'objectif de notre travail n'étant pas strictement la visualisation de l'incertitude, nous ne présentons dans la suite que les travaux qui nous ont paru les plus pertinents par rapport à notre travail.

8.4.1 Visualisation de l'incomplétude des données

Dans le cas où les données à étudier représentent des valeurs manquantes, comme dans le cas des séries temporelles avec des données absentes représentant des discontinuités de la récolte, plusieurs approches de visualisation ont été proposées.

Sjöbergh et Tanaka [ST17] proposent de coordonner différentes vues afin d'obtenir des informations exploitables. Pour ce faire, ils proposent différentes visualisations dont leur agrégation donne une nouvelle information synthétique, et ce, pour éviter les informations trompeuses. L'adéquation des vues est considérée comme essentielle en vue de l'aide puissante qu'elle représente pour comprendre les éléments et leurs propriétés. Visualiser la discontinuité par plusieurs angles d'analyse aide également à réduire les mauvaises interprétations et reconnaître des modèles de trafic pour un contrôle de qualité plus intelligent.

La figure 8.7 présente une agrégation de deux visualisations qui montre le nombre de valeurs manquantes trouvées par variables d'étude et leurs fréquences d'apparition.

Le diagramme en bar (à gauche dans la figure 8.7) montre la non-existence des données par des fréquences et des proportions, tandis que la figure à droite montre toutes les combinaisons existantes des valeurs manquantes et non manquantes. Dans cette figure, les rectangles foncés indiquent l'absence dans la variable correspondante et les rectangles gris clair représentent les données disponibles. De plus, les fréquences des différentes combinaisons sont représentées par une petite barre horizontale (voir à l'extrémité droite). Ce type de graphique permet ainsi de reconnaître des comportements par le ré-ordonnancement du cadre visuel.

Templ et al.[TAF12] visualisent l'absence de données en adaptant des histogrammes. Cette adaptation met en évidence deux classes de données (les données observées et les non observées). Dans cette approche, la quantité de valeurs manquantes sur un histogramme

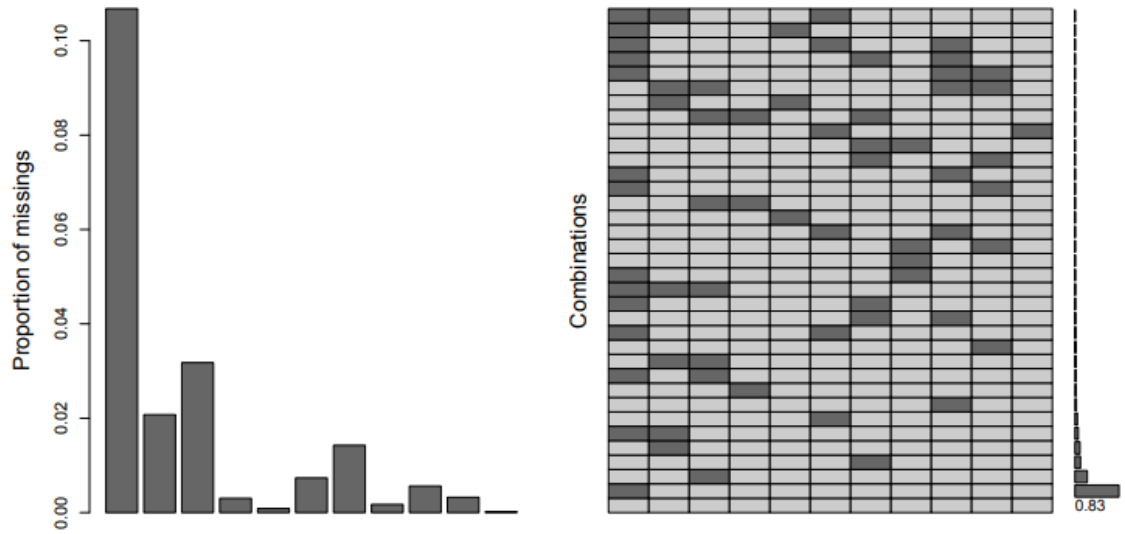


FIGURE 8.7 – Visualisation des volumes de données manquantes selon [TAF12]

est visualisée par un marqueur sur les barres de l'histogramme ou séparée par une barre qui s'éloigne du reste du graphique. Elle peut être aussi placée au dessus des barres de l'histogramme et ce, en prenant compte les fréquences des valeurs observées (voir figure 8.8). Ce type de visualisations peut mener à des visualisations multi-variables ou multi-catégories.

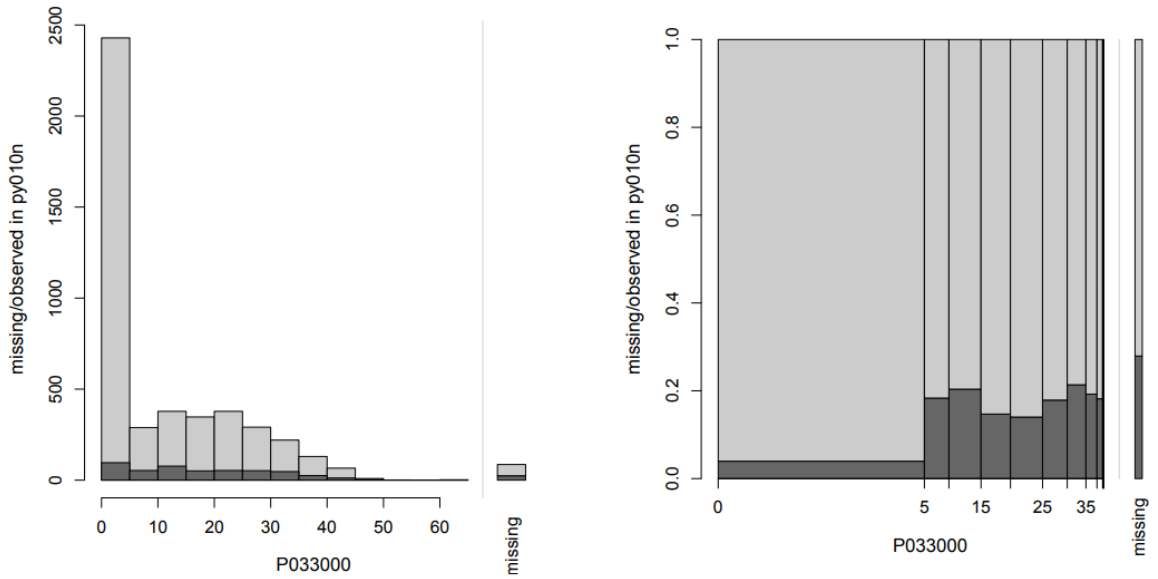


FIGURE 8.8 – Spinogramme pour la visualisation actfis/inactifs [HT05]

Pour une visualisation plus développée des valeurs manquantes, Hofmann et Theus [HT05] proposent les spinogrammes. Ce type de visualisations est étroitement liés aux histogrammes. L'axe horizontal est mis à l'échelle en fonction des fréquences relatives, i.e. la largeur des barres reflète les fréquences plutôt que leur hauteur. Ainsi, la hauteur correspond à la proportion de valeurs manquantes et ou observées. Par cette visualisation,

il est maintenant possible de comparer les proportions de valeurs manquantes à travers les différents bacs. Des différences significatives dans ces proportions peuvent indiquer des éventuelles situations à traiter. La figure 8.8 (à droite) contient un spinogramme de la variable P033000 (année d'étude). Les valeurs observées sont des valeurs binaires du marché de l'emploi (chômeur présenté en gris clair, gris foncé sinon). Nous pouvons facilement remarquer, via le spinogramme, que le marché du travail est dominé par les inactifs.

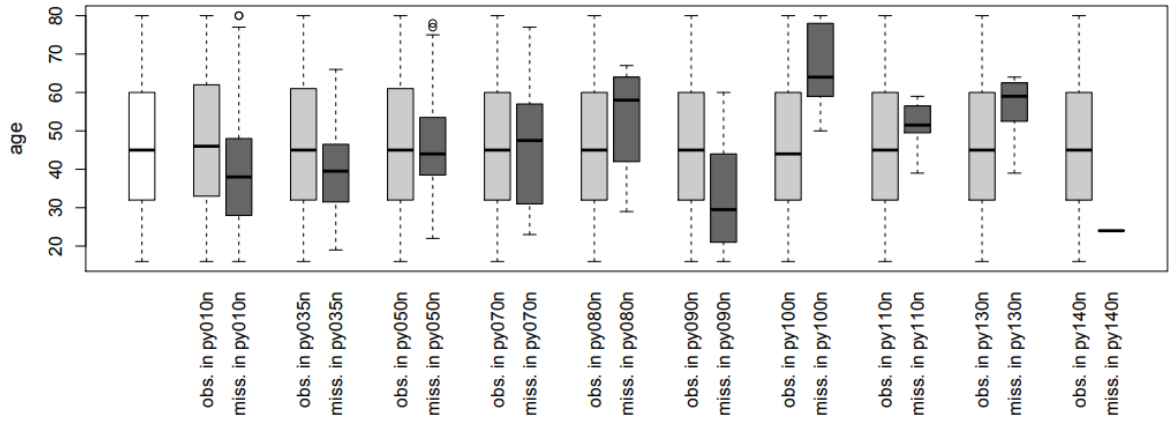


FIGURE 8.9 – Boîtes à moustaches parallèles [Hei+97]

Quand il s'agit des variables continues, les distributions des valeurs observées et manquantes peuvent être comparées par des boîtes à moustaches parallèles (Paralell Boxplot). Ce graphique est particulièrement utile pour déterminer si une solution reposant sur des variables continues explique la distribution des valeurs manquantes. La figure 8.9 montre un exemple de cette présentation sur une variable continue, ici l'âge, (Ordonnées) au fil d'une sous composition catégorielle (abscisse). Cette visualisation fournit une boîte à moustaches standard (à gauche en blanc) et décomposée en sous boîtes à moustaches (droite) selon des catégories. Dans ces boîtes, les valeurs observées sont gris clair et les valeurs manquantes sont en gris foncé selon les sous catégories.

Heike et al. [Hei+97] proposent de prendre en considération les largeurs des boîtes par rapport à l'importance des sous catégories. Cette représentation peut ne pas être pertinente dans le cas où le nombre des valeurs manquantes est proche de 0, ce qui laisse l'explication de certains phénomènes impossible. Par contre une interaction visuelle appliquée sur cette visualisation est intéressante en permettant de naviguer d'une variable d'étude à l'autre et de zoommer sur les boîtes à moustaches. Un clic sur une boîte pourrait mener à une nouvelle visualisation sur la catégorie visée (en reprennant possiblement le même type d'approche).

8.4.2 Visualisation de l'imprécision des données

Nathan Yau [2] part de l'idée que les données sont une représentation de la vie réelle. Une abstraction exhaustive des données est ainsi impossible car on ne peut pas tout modéliser. En effet, échantillonner un jeu de données crée forcément des doutes sur l'information que l'on en tire. En effet, l'échantillon obtenu peut ne pas être représentatif. Aussi, quelle confiance pouvons-nous avoir envers les valeurs ? Quels sont les taux d'erreur possibles ? etc.

Dans ce qui suit nous mettons l'accent sur des visualisations permettant d'essayer de réduire l'incertitude dans l'interprétation des résultats en visualisant par exemple les données avec leur imprécision.

Visualisations floues

L'idée générale est de proposer des visualisations rendant compte des degrés d'appartenance issus des processus de fuzzification [ZC07 ; Zuk08 ; BP03]. En effet, comme pour tout traitement de données, il est important de visualiser ces données fuzzifiées afin d'exploiter les résultats et comprendre les phénomènes.

MacEachren [Mac92] affirme que, plus une donnée est imprécise, plus il est difficile de la visualiser. Pour la visualiser il propose de jouer sur la visibilité via une échelle allant du moins visible au plus visible (voir figure 8.10). Cette échelle peut être en fonction de la couleur ou de l'opacité de la vision.

FIGURE 8.10 – Visualisation de l'imprécision en utilisant l'opacité [Mac92]

Un autre type de visualisation a été proposé dans [PWL97]. L'idée est de projeter des prévisions d'une donnée à partir d'un état actuel précis sur des états possibles futurs. Cette visualisation permet de donner une idée sur les possibles valeurs qui viennent sur un axe du temps, par exemple.

2. `https://flowingdata.com/2018/01/08/visualizing-the-uncertainty-in-data/` vue le 31/10/2019

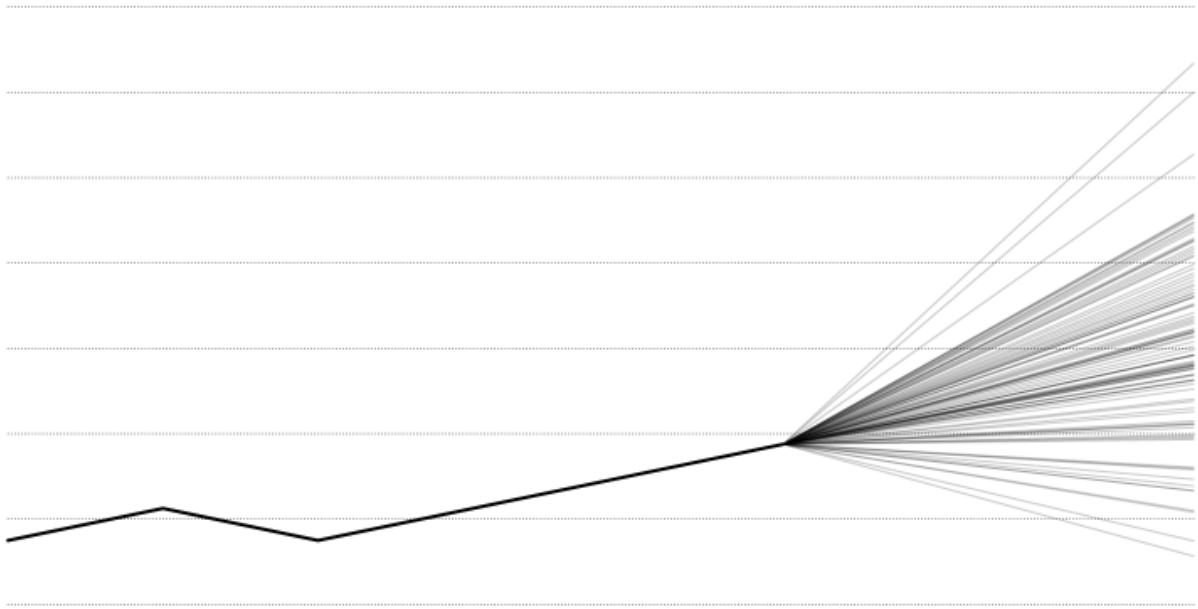


FIGURE 8.11 – Visualisation de l'imprécision en utilisant une projection sur les possibles valeurs [PWL97]

FIGURE 8.12 – Visualisation de l'évolution du caractère imparfait [PWL97]

Ceci est important pour connaître les différents résultats possibles afin de se préparer aux différents scénarii d'études. Dans cette visualisation, l'imprécision est affichée plus explicitement car il n'y a pas un seul chemin mais plusieurs chemins possibles.

Cependant, l'inconvénient possible de cette visualisation est que si le niveau de bruit est élevé ou s'il y a beaucoup de possibilités, il se peut que le graphique devienne dense et ne contienne aucune information utile.

L'exemple de la figure 8.11 illustre cette approche en affichant plusieurs résultats possible à partir d'un moment temporel.

Par ailleurs, dans [PWL97], les auteurs proposent de visualiser l'évolution du caractère imparfait (incertain ou imprécis). Ce type d'affichage présente des résultats multiples à l'issue d'un traitement sur une ou plusieurs données. L'idée est que le degré de confiance sur la valeur soit visualisé au cours du temps en jouant sur l'opacité ou la couleur mais aussi sur la taille. Par exemple, plus la valeur est de confiance, plus sa taille sera grande et sa couleur sera foncée (voir figure 8.12).

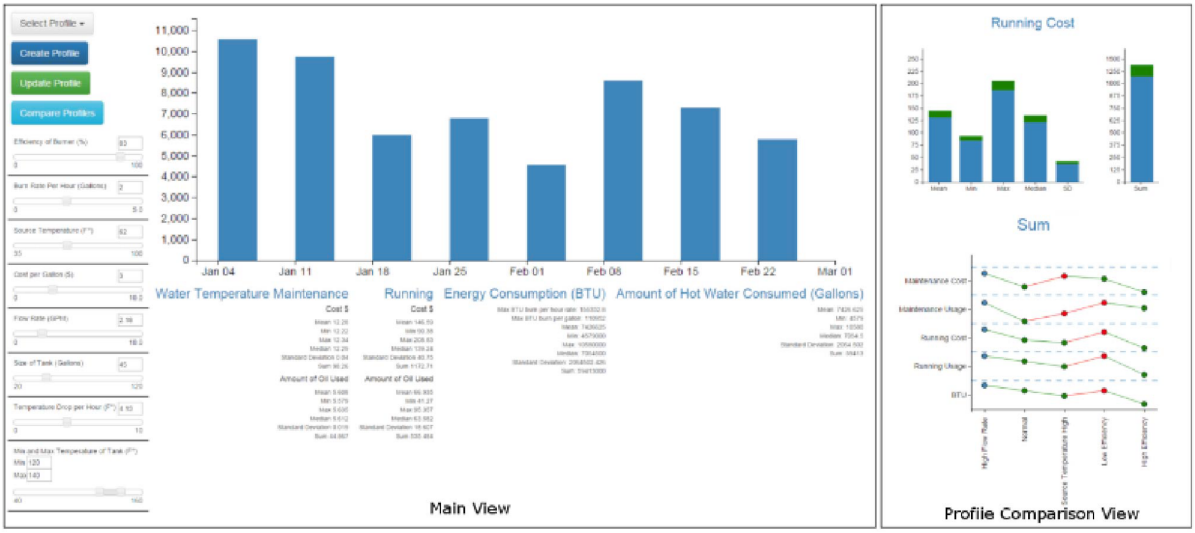


FIGURE 8.13 – Interface de l'outil Quick Vis [AL15]

8.5 Outils de visualisation

L'analyse visuelle vise à la création d'outils et de techniques permettant de :

— Synthétiser l'information et tirer des conclusions sur les données étudiées, e.g., sur des données bruitées et atypiques.

— Détecter ce qui est attendu et découvrir l'inattendu.

— Fournir des évaluations et des compréhensions sur les données.

L'objectif de la conception d'un outil est de rassembler, en son sein, des indicateurs visuels et des tableaux de bord afin d'offrir une navigabilité importante. Cela permet ainsi d'obtenir de nouvelles connaissances sur les données manipulées. La manipulation des graphiques, de leurs paramètres et de leurs collaborations au travers de la navigation permet de donner naissance à plusieurs outils d'analyse visuelle.

Quick Vis

Cet outil repose sur l'idée de construire et de comparer les profils des utilisateurs et de leurs navigations afin d'en comprendre les comportements. L'outil utilise des métriques de similarité adaptées à la notion du profil. Pour pouvoir répondre aux besoins, Agnello et al. [AL15] intègrent dans cet outil les concepts suivants :

— Moteur d'analyse fondé sur le calcul de similarité entre profils. Ce moteur reçoit un ensemble de valeurs et de propriétés et affecte une valeur unique par profil.

— Une technique interactive où les propriétés des profils peuvent être modifiées, ce qui déclenche automatiquement une mise à jour du calcul précédent.

— Une approche interactive qui permet de manipuler les profils par ensemble.

— Sélection de profils multiples à des fins de comparaison

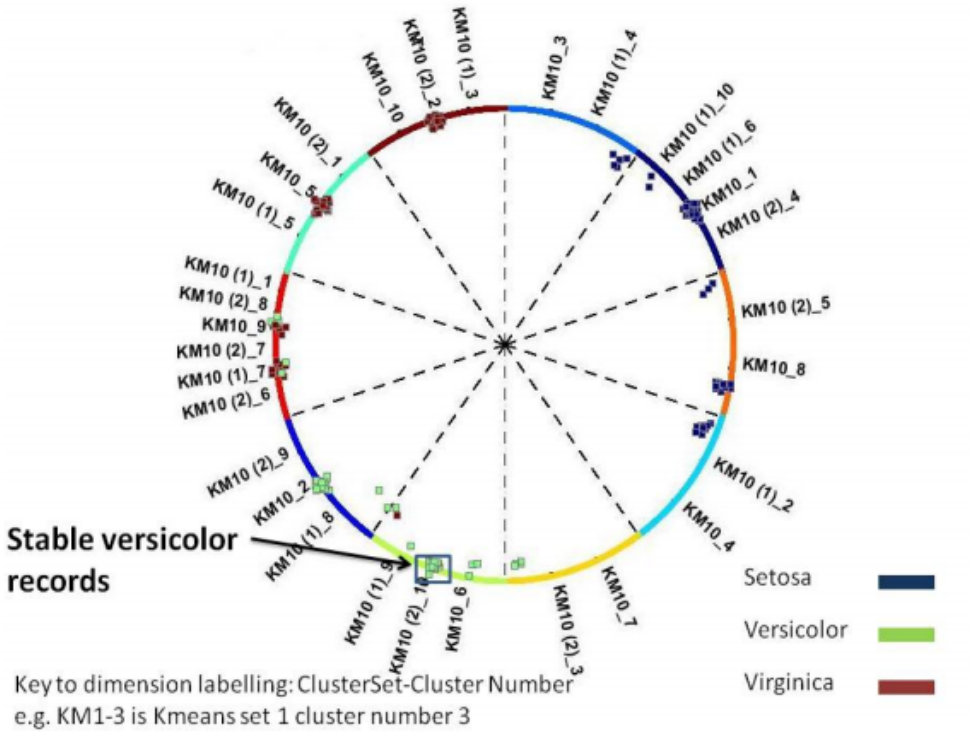


FIGURE 8.14 – Zones de concentrations sur une images [SG09]

L'interface de l'outil, illustrée dans la figure 8.13 repose sur deux vues principales (tableaux de bord). Chacune est étendue sur une interface. Une première vue permet l'analyse et la création des profils et une deuxième permet de les comparer. Ces vues utilisent principalement cinq composantes : Histogrammes, tableaux croisés dynamiques, curseurs, graphiques empilés dynamiques et tableaux de tendances.

L'idée principale de la conception est donc de consacrer chacune des vues à un objectif particulier. Chaque vue correspond à un ensemble d'objets visuels répondant à l'objectif du sujet.

RadViz

Cet outil développé par Sharko et Grinstein [SG09] utilise la logique floue pour visualiser les clusters associés à des enregistrements de données. Il permet aussi de donner des jugements sur la stabilité de l'affectation des données à des groupes, i.e. si la donnée appartient bien à un groupe. Il adopte donc des visualisations spécifiques pour ce propos. Par exemple, dans la figure 8.14 à l'issue d'un paramétrage des clusters, si un point tend vers le centre ça veut dire que la donnée qu'il représente est mal catégorisée ou possiblement bruitée. La figure montre un jeu de données défini sur trois variables d'étude où une des données est bien catégorisée et jugée stable à la vue de son rapprochement au centre du domaine associé à son groupe. La mesure pour dire qu'une donnée est stable vis-à-vis de sa catégorie est donc la distance vers le centre du domaine.

8.6 Conclusion

Dans ce chapitre, nous avons étudié plusieurs travaux sur la visualisation, et plus particulièrement ceux traitant des flux de données imparfaites. Nous avons vu comment on peut placer efficacement les objets visuels sur les tableaux de bord, utiliser les techniques de visualisation qui peuvent enrichir la présentation d'un tel résultat et finalement, étudier différentes méthodes de visualisation de l'imparfait.

Ces travaux sont en concordance avec notre objectif de fournir un outil de visualisation pour la supervision de la qualité de nos données. Aussi, nous nous en sommes inspiré pour proposer un prototype, répondant à notre objectif, présenté dans le chapitre suivant.

Chapitre 9

Contribution 3 : MMS Explore

9.1 Introduction

Nous avons proposé, précédemment, deux contributions permettant d'évaluer la qualité des séries temporelles tout en utilisant des indicateurs adéquats à la nature des données manipulées. Les deux contributions sont paramétriques, uni-variées et issues d'une source agrégée. Nous constatons, donc à minima, qu'un outil de visualisation interactive peut s'avérer nécessaire ne serait-ce que pour manipuler les paramètres des précédentes approches.

Par ailleurs, comme nous avons pu le remarquer dans l'état de l'art, la qualité des données peut être étudiée selon différentes dimensions. Aussi offrir un outil permettant de naviguer dans ces différentes dimensions est tout aussi utile.

Ainsi, fournir un prototype simple pour une manipulation destinée à un usage professionnel est souhaitable pour faciliter la compréhension des résultats des approches précédentes et tirer intérêt de leurs objectifs. Il faut que cet outil offre un mixte entre simplicité d'utilisation et possibilité d'axes d'observations permettant de découvrir de manière interactive le plus de connaissances possibles sur les récoltes de données.

Dans cette partie nous présentons un prototype d'outil, appelé MMS Explore (MMSE), qui fournit les moyens nécessaires à l'évaluation de la qualité de la récolte et, ainsi, qui permet de mieux comprendre le comportement des capteurs.

L'outil met en relation des concepts clés de la visualisation, par l'intermédiaire de tableaux de bord, et permet de mieux explorer les résultats des indicateurs et des dimensions de la qualité. L'outil exploite principalement l'approche QBA. L'intégration de l'approche FBA est une des perspectives possibles.

Compte tenu que l'objectif principal est d'évaluer la qualité de la récolte, nous présentons à la fin de ce chapitre divers cas d'utilisation pour montrer l'utilité de cet outil dans l'étude des comportements des capteurs.

9.2 Problématique et objectifs

Afin de permettre aux experts de Kantar de superviser la qualité de leurs récoltes, ce chapitre propose de répondre à la question : Quels outils de visualisation et d'interaction faut-il déployer pour permettre une exploration interactive de la qualité des flux de données ?

À partir de cette question et dans l'objectif de permettre à l'utilisateur de mieux appréhender la qualité de ces données en gérant leur imperfection, nous nous posons les questions suivantes :

— Quels sont les concepts à visualiser pour mieux appréhender les données ?
— Quels principes de la fouille visuelle interactive peut-on exploiter ? Et pourquoi ?
— Quels sont les tableaux de bord à utiliser pour unifier la compréhension d'une dimension de la qualité ?
— Quels sont les indicateurs visuels à mettre en place pour informer l'utilisateur sur la qualité des données ?
— Quelles modélisations et architectures logicielles choisir pour développer un tel prototype ?

9.3 Principes et hypothèses

Afin de répondre aux questions précédentes, nous proposons un prototype d'outil de visualisation interactive pour la supervision de la qualité des données et des récoltes. Cet outil s'inscrit dans une stratégie d'établissement de la qualité dirigée par les données (cf. 3.2.2).

Les données manipulées sont orientées sur plusieurs axes d'études, comme, par exemple, les données brutes [1], l'audience, le contenu d'une publicité, etc. Ces données sont multi-variées, multi-sources et multi-capteurs. Aussi, la première hypothèse, que nous posons, est que l'analyse de la qualité des données doit pouvoir se faire selon plusieurs axes d'étude. On doit donc pouvoir étudier les données, acquises et agrégées selon une échelle temporelle, par sources, par média, par catégorie et méta-catégorie, ainsi que par combinaison de ces informations.

Nous avons vu dans le chapitre 3 que la qualité des données peut être étudiée selon différentes dimensions. Nous proposons dans ce chapitre un outil dont l'objectif est de permettre d'étudier la qualité des récoltes principalement selon les dimensions suivantes :

— La complétude
— La variabilité
— La stabilité

1. les données récoltées et enregistrées directement par les robots

Dans cet objectif, et en suivant le principe du visual thinking (cf. chapitre précédent), notre prototype propose plusieurs tableaux de bord ayant chacun un objectif particulier, afin d'étudier :

— Les valeurs récoltées agrégées selon une échelle temporelle afin d'avoir une première idée de l'état de la récolte.
— Les volumes de données manquantes et à l'opposé les volumes présents afin d'avoir une information sur la complétude des données d'une part et sur l'état de la récolte d'autre part.
— Les quantiles internes et externes d'appartenance des données récoltées afin de pouvoir positionner l'état de la récolte d'un ou plusieurs capteurs vis à vis des autres. L'intégration de l'approche QBA permet de fournir des indicateurs sur la qualité des données temporelles imprécises, lacunaires et aide à évaluer l'irrégularité des flux temporels.
— Les variabilités et stabilités obtenues via l'approche QBA afin d'obtenir des informations quantitatives sur la volatilité des flux.

Les tableaux de bord s'appuient sur la visualisation d'indicateurs clés de performance (Key Performance Indicator – KPI). La combinaison de ces indicateurs dans des tableaux de bords offre, pour chaque dimension d'étude, une vision synthétique et complémentaire des données étudiées. Ces indicateurs sont statiques i.e. des indices précis à un temps d'étude bien défini, ou bien dynamiques, i.e. évoluant dans une période de temps.

Nos différents objets visuels associés sont tous issus du domaine. Nous avons cependant détourné leur utilisation afin de mettre en avant du contenu. Nos visualisations n'exploitent pas le principe de transparence pour l'imprécision et l'incertitude car nous n'exploitons pas dans ce chapitre les approches floues.

Par ailleurs, afin de permettre à l'utilisateur d'interagir avec l'outil en vue d'une meilleure exploration des données, les principes suivants sont exploités :

— Zoom in/zoom out : l'utilisateur a la possibilité de faire une sélection/dé-sélection dans les données, à l'instar de la sélection d'un capteur ou d'un ensemble de capteurs. Cette sélection peut être faite ou combinée avec une sélection des sources et/ou des catégories/méta-catégories. Ce principe permet de pouvoir observer plus en profondeur (de manière plus individualisée) un sous ensemble de capteurs, ou au contraire de généraliser afin de comprendre des phénomènes plus généraux. La sélection peut aussi être faite sur la période à étudier.
— Réglage des paramètres : l'utilisateur peut interagir avec l'outil pour définir les paramètres souhaités notamment l'échelle et le nombre de quantiles pour l'approche QBA. Les autres paramètres nécessaires au calcul de la variabilité ne sont pour l'instant pas modifiables par l'utilisateur. Cela permet à l'utilisateur de jouer sur les échelles et donc de pouvoir affiner son appréciation de la qualité d'une récolte.

— La continuité dans la navigation : en suivant le principe que la visualisation doit permettre une continuité dans l'analyse des résultats, les sélections faites à un moment donné se répercute sur l'ensemble des tableaux de bord. Cela permet à l'utilisateur de poursuivre son interprétation en naviguant entre les différents tableaux de bord.

La figure 9.1 montre le pipeline du prototype développé. Les données venant des différentes sources et capteurs sont analysées par un moteur fondé sur QBA, puis elles sont intégrées dans l'outil.

9.4 Modélisation

MMS Explore est développé de manière à guider l'utilisateur pour avoir une vision unifiée sur une ou plusieurs dimensions de la qualité des données. En intégrant le principe de la pensée visuelle, les tableaux de bord et les KPI orientent la recherche d'une connaissance sur un sujet particulier afin de trouver une conclusion.

Le processus de la recherche d'une information est simple. L'utilisateur choisit d'abord la dimension sur laquelle il souhaite travailler. Il sélectionne les données à étudier. Les indicateurs le guident alors afin qu'il comprenne l'irrégularité du flux sur la sélection. L'outil permet aussi de faire varier les paramètres de l'approche QBA. Ceci permet à l'utilisateur de comprendre certains comportements dans les flux imparfaits.

Par ailleurs, MMS Explore (MMSE) répond à l'usage suivant. Le processus examine des ensembles de données afin de trouver des informations utiles à l'évaluation de la qualité. Il s'agit généralement d'utiliser les résultats issus de l'approche QBA. Cette vue est destinée à des utilisateurs ayant des connaissances préalables sur les données.

9.4.1 Modèle général

La figure 9.2 présente un aperçu du modèle général de fonctionnement de l'outil. Un utilisateur peut choisir un axe de travail, e.g une sélection de données sur laquelle il souhaite obtenir des informations sur la qualité. Il doit pouvoir effectuer des personnalisations liées à l'orientation de sa recherche, c'est-à-dire choisir un axe de travail particulier avec une ou plusieurs sources. Il peut également spécifier un nombre de capteurs à étudier. À ce stade, nous lui proposons la possibilité de vérifier les problèmes liés à ces récoltes. Il peut pour ce faire, utiliser des techniques telles que les graphiques, les indices statistiques, etc. ou bien les indicateurs fournis par QBA.

Nous proposons trois concepts différents pour aider l'utilisateur à mieux percevoir la qualité d'une sélection particulière :

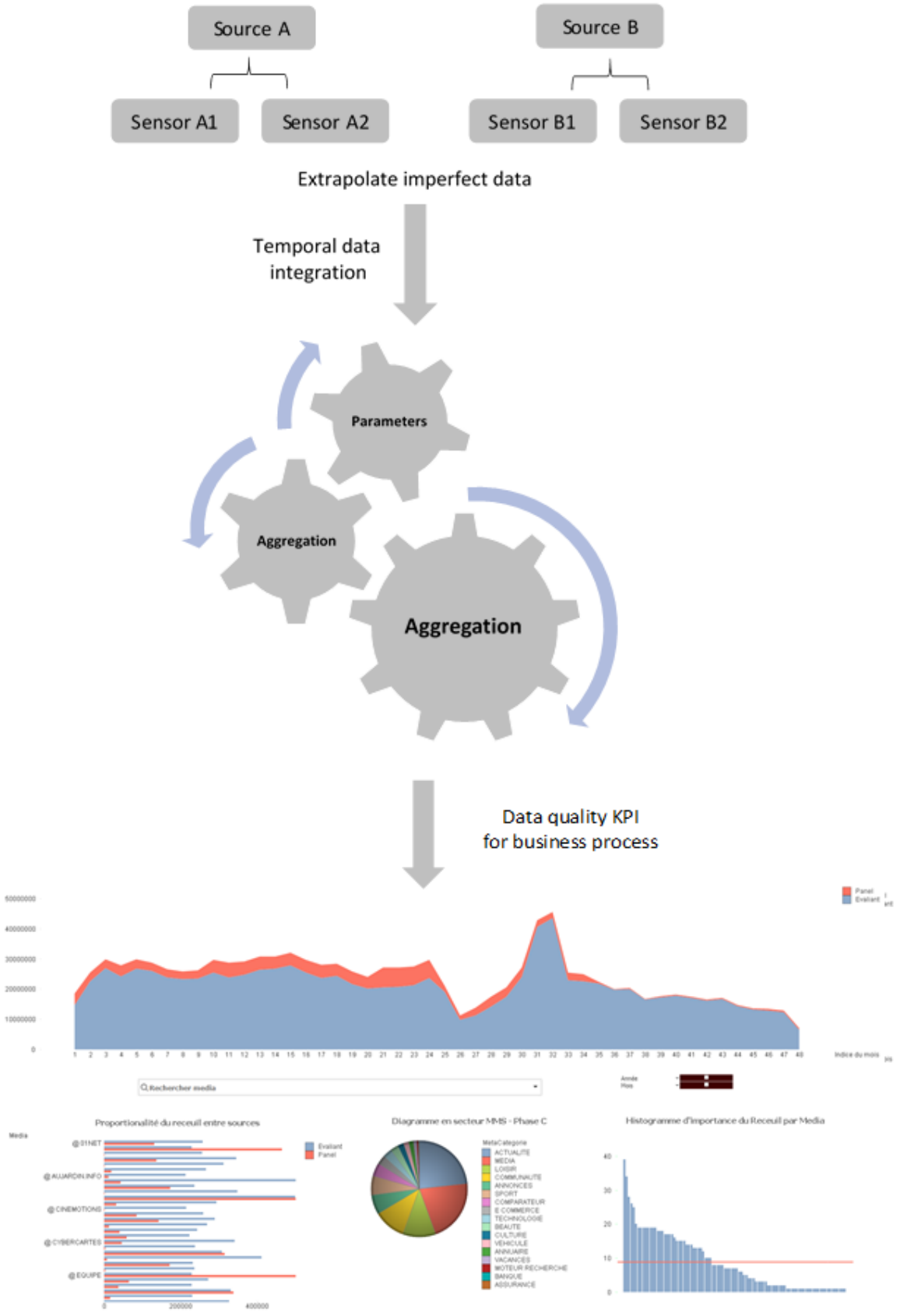


FIGURE 9.1 – De l'acquisition des données imparfaites au développement des tableaux de bord

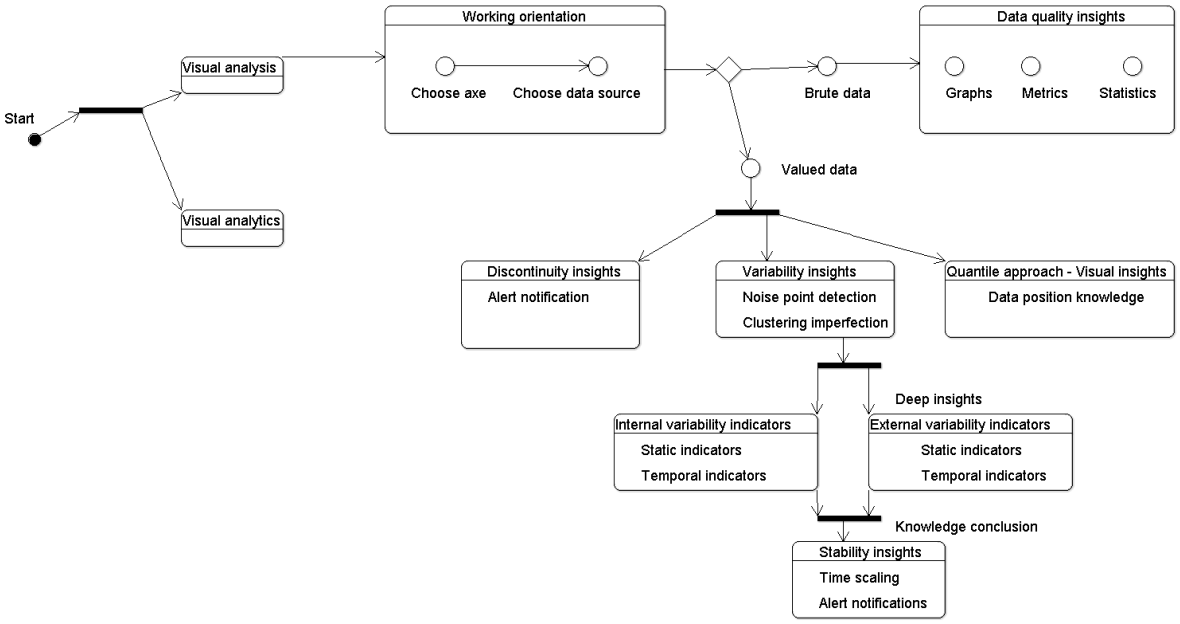


FIGURE 9.2 – Modèle général de supervision de la qualité des données temporelles imparfaites

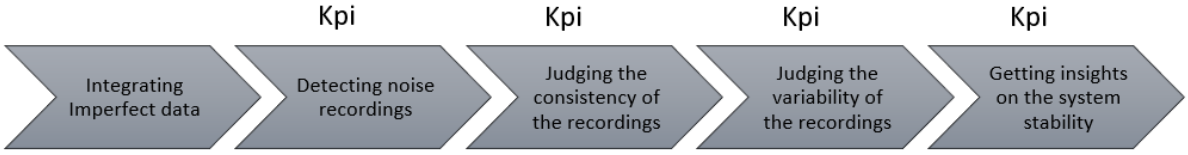


FIGURE 9.3 – Exemple de chaînage d'utilisation des outils visuels pour mieux appréhender la stabilité des récoltes

1. Le premier est la discontinuité : ce groupe d'indicateurs clés de performance vise à fournir des informations sur la qualité en quantifiant l'absence dans les données temporelles. Ce modèle adopte des métriques en comparant différents flux de données pour fournir des vues synthétiques sur une sélection de données.

2. Le deuxième aspect utilise des vues développées reposant sur l'approche QBA. Ceci permet d'évaluer les positions en quantiles des données et distinguer les comportements aberrants, et ce, en fonction de plusieurs variables d'études.

3. Le troisième concept consiste à adopter des indicateurs particuliers en rapport avec les dimensions citées auparavant e.g indicateur de surveillance temporelle de la stabilité des flux.

Les KPI, que nous présenterons dans les sections suivantes, aident l'utilisateur à évaluer plus précisément la qualité des données selon les diverses dimensions.

Cette structuration est faite pour aider l'utilisateur à mieux comprendre la qualité des données, à l'instar par exemple de la stabilité. La figure 9.3 présente les étapes enchaînées dans ce but. Les flèches expliquent la possibilité de faire un zoom-in jusqu'à l'établissement d'une compréhension sur la dimension stabilité, i.e avoir des résultats sur cette dimension suite à des manipulations précédentes. En pratique, la solution repose sur la manipulation des objets visuelles en utilisant le zoom-in/zoom-out.

Afin de permettre à l'utilisateur de pouvoir voir les données sous plusieurs dimensions, notre outil est organisé en plusieurs tableaux de bord. Chaque tableau de bord permet à l'utilisateur d'appréhender une dimension particulière.

9.4.2 Dimensions d'étude intégrées

Afin d'analyser la qualité du recueil, notre outil exploite les dimensions suivantes :

1. Étude de la discontinuité : Au vu que nous sommes dans un contexte de données temporelles volumineuses, la détection des arrêts et reprises de la récolte d'un ensemble des capteurs est délicat. Notre outil intègre des graphiques aidant à juger la continuité des enregistrements (présence/absence par exemple). Il adopte aussi un système sur la criticité de la récolte indiquant le niveau d'urgence et le degré de la nécessité d'une intervention rapide.
2. Étude de comportements : Nous proposons des visualisations pour étudier le comportement des capteurs dans des périodes d'études choisies par un utilisateur. Ce type de visualisations est donné par la projection des données dans leur quantiles soit par rapport à eux mêmes ou bien par rapport à l'ensemble. Cette dimension d'étude nous permet de visualiser les positionnements temporels de la récolte soit sur l'aspect interne et/ou sur l'aspect externe. L'étude du positionnement interne permet de vérifier la cohérence de placement des publicités sur un média donné, sur une variable et durant une période d'étude. L'étude du positionnement externe consiste à identifier le positionnement temporelle d'une donnée par rapport à toutes les autres appartenant à une sélection bien définies. Ces informations nous permettent de construire des indicateurs de suivi de tendance de chaque capteur. Cette dimension d'étude permettent, par ailleurs, de prendre en considération l'imprécision des données dans les flux puisque les informations à visualiser ont été analysées en amont par l'approche QBA. Le choix des paramètres e.g. le nombre de quantiles, l'échelle, etc. sont paramétrables dans notre outil.
3. Étude de la variabilité : la variabilité est calculée selon la vision interne et externe. Le paramètrage du calcul de la variabilité n'est pas encore intégré dans le *Front-End* de l'outil mais est possible dans le *Back-End* et les scripts d'analyses associés. La dimension variabilité permet de donner des scores relatifs au mouvements internes (Sc_{int}) et externes (Sc_{ext}). Ces scores nous ont permis de détecter de possibles

problèmes au niveau du recueil. Ces derniers sont représentés dans notre outil par un nuage de points. Les points singuliers i.e les points qui s'éloignent de la masse, expliquent l'existence de possibles dysfonctionnements (anomalies) au niveau des capteurs.

4. Étude de la stabilité : Cette dimension d'étude met en relation la variabilité externe avec la variabilité interne (dite cohérence). Son objectif est de présenter des indices quantifiant la stabilité totale par QBA, i.e. une mesure globale informant sur la qualité du fonctionnement d'un ou plusieurs capteurs. Les scores fournis peuvent se présenter sous la forme d'indicateurs statiques ou dynamiques, sur un temps précis ou sur une période du temps, sur une catégorie, une méta catégorie ou un ensemble personnalisé.

Pour faciliter l'analyse des données, les tableaux de bord sont organisés d'une manière consécutive permettant de comprendre les relations entre leurs contenus e.g Accueil → Données brutes → Présence/absences des données → Quantiles, etc. . Cette organisation permet à l'utilisateur d'aller en profondeur dans sa recherche d'informations. Pour chacune des dimensionsx les tableaux inclus plusieurs KPI et objets visuels.

9.4.3 Modèle des tableaux de bord

Dans cette section, nous présentons le modèle (voir figure 9.4) sur lequel repose la conception de nos différents tableaux de bord. Ce modèle unifie les différentes vues présentées a l'utilisateur. Il montre comment les objets visuels sont positionnés dans les tableaux de bords.

Ce modèle est conçu sur un arbre divisé en deux parties.

La première partie contient des objets en relation avec le processus métier e.g les sources, les capteurs, etc. Dans cette partie l'utilisateur peut sélectionner des données en fonction de ses critères de recherche, principalement en relation avec l'expertise métier.

La deuxième partie contient les objets faisant référence à l'évaluation de la qualité des données e.g les indicateurs, les indices, etc. L'utilisateur utilise certains KPI en rapport avec la dimension visée. Il peut aussi profiter des métriques de QBA pour étudier en profondeur les données imparfaites.

Le tableau suivant illustre les types d'indicateurs clés de performance utilisés dans les tableaux de bord de MMS Explore.

9.5 Présentation de l'outil

L'outil fonctionne sur la base de divers tableaux de bord, complémentaires et interagissant entre eux, intégrant les dimensions d'études cités auparavant. Chaque tableau de bord répond à un questionnement spécifique et vise à apporter des informations complémen-

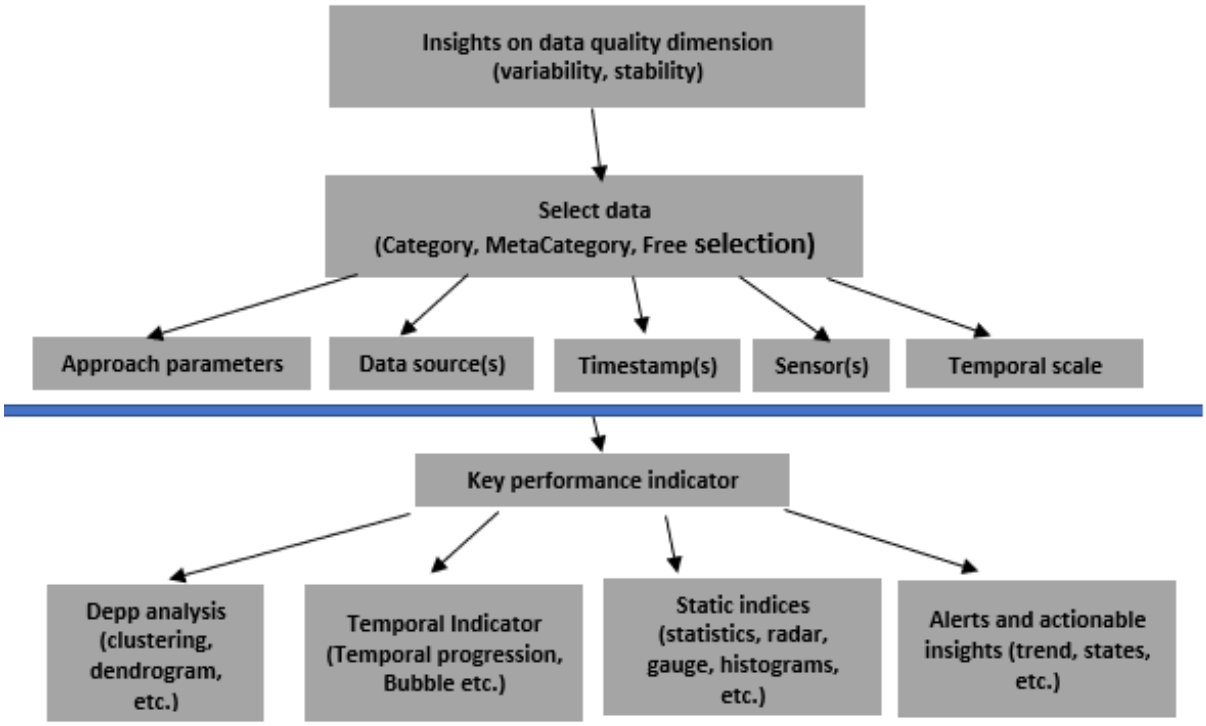


FIGURE 9.4 – Modèle des tableaux de bord de MMS Explore

Catégorie des KPI	Motivation
Dynamique	Traite l'irrégularité des flux de données temporelles imparfaites
Statique	Examine une information sur une période bien définie
Alerte	Alerte sur des situations critiques détectées
Actionnable	Motive un utilisateur à agir sur des situations particuliéres

TABLE 9.1 – Catégories des KPI en fonction de leurs intérêts

taires tout en répondant à une problématique bien définie. L'affichage des informations est donné par rapport à l'information recherchée, i.e par rapport à une sélection de données à étudier. MMSE offre des visualisations de méta-compréhension de l'information e.g par l'utilisation d'un méta-plan quantile, ce qui donne un jugement différent sur les comportements. Les différentes modalités d'interactions utilisées dans notre outil permettent d'explorer en profondeur les informations sur la qualité des données et des capteurs.

Voici une liste de sélections possibles dans MMS Explore :

— Multi-axes : Choix d'un axe de travail approprié
— Multi-sources : Navigation entre les sources de la récolte
— Multi-capteurs : Choix du nombre des données médias à analyser
— Temporalité : Choix de l'échelle temporelle (Mensuelle, Journalière)
— Catégorisation : Choix d'une catégorie ou méta-catégorie
— Typologie : Choix d'un ensemble précis de médias

L'outil propose des tableaux de bord incluant des traitements sur les données brutes, les données valorisées (en variables d'études) et les résultats fournis par QBA.

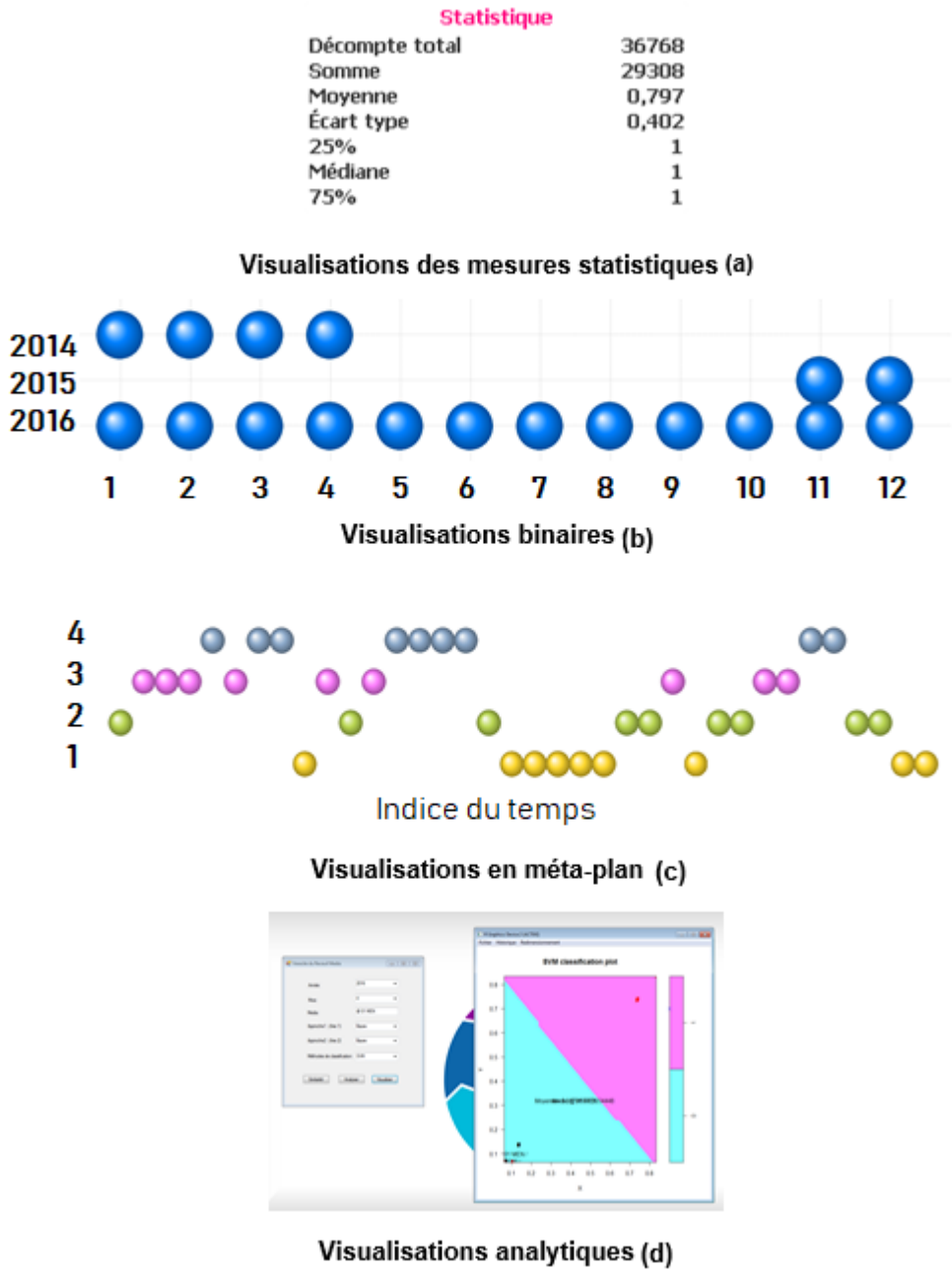


FIGURE 9.5 – Liste des visualisations possibles proposées par MMS Explore

Voici une liste de visualisation proposées par MMS Explore :

— Visualisations des mesures statistiques (figure 9.5 - a) : des mesures qui se changent automatiquement à la suite d'une sélection d'un changement d'axe de travail.

— Visualisations binaires (figure 9.5 - b) : un mode de représentation d'existence/absence de la donnée selon un axe de temps ou une catégorisation appropriée.

— Visualisations en méta-plan (figure 9.5 - c) : un type de visualisation informant sur les positionnements des données.

— Visualisations analytiques (figure 9.5 - d) : appels externes à des scripts de fouille de données offrant des visualisations analytiques.

La figure 9.6 [2] montre les différents choix et personnalisations possibles.

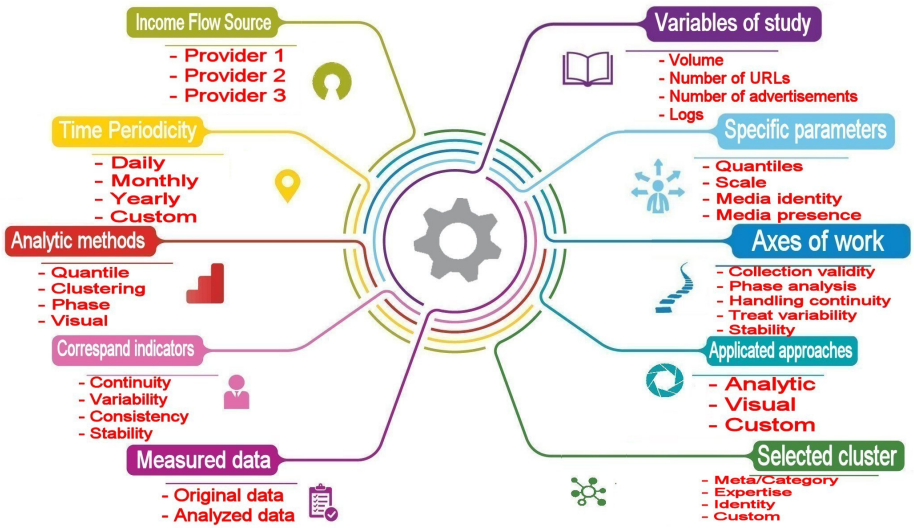


FIGURE 9.6 – Ensemble des fonctionnalités et paramètrages possibles de MMS Explore

MMS Explore offre également la possibilité d'interagir avec des scripts externes à l'outil. On peut donc faire appel à d'autres indicateurs de fouille de données pour élargir notre vision d'étude.

9.5.1 Environnement technique

Notre prototype exploite les données issues des pré-traitements présentés dans le chapitre 2. Afin d'exploiter les données, l'ensemble de nos analyses statistiques et de nos approches de fouille de données sont implantées dans un outil R qui est le noyau ou le moteur de MMS Explore. Il permet de lancer des analyses selon les besoins et produit des résultats destinés aux utilisateurs de notre outil. Ces résultats représentent les données alimentant un outil de visualisation interactif développé en utilisant QlikView.

Nous avons utilisé l'interactivité (et la programmation événementielle) présente dans QlikView pour offrir une manipulation simple et puissante des résultats au travers des tableaux de bords. Afin que notre outil permette l'ensemble des modalités d'interaction voulues, plusieurs scripts *ad hoc* ont été développés et ajoutés à l'outil QlikView.

Les tableaux de bord représentent les interfaces visuelles de MMS Explore. Ces dernières interagissent directement avec le noyau d'analyse pour, par exemple, varier les paramètres de QBA et faire appel à des scripts d'analyse exterieurs.

2. Image que nous avons développé pour illustrer les différentes orientations possibles dans MMSE

Enfin, le prototype a été développé en version monoposte mais aussi en version web déployée sur un serveur web. Ceci facilite les discussions entre les agents sur la qualité des données.

9.6 Techniques de visualisation utilisées

MMS explore adopte deux techniques de visualisation : technique d'interactivité et technique de la pensée visuelle. La première sert à faire des zoom-in et zoom-out sur les objets visuels, i.e avoir des informations précises ou agrégés pour unifier une compréhension globale. La deuxième technique consiste à acheminer une recherche d'information pour fournir les indicateurs et les orientations nécessaires.

9.6.1 Interactivité

Cette technique est appliquée sur les différentes dimensions d'études. On peut par exemple travailler sur la dimension "discontinuité" pour faire une sélection sur une catégorie des sites qui ont plus de valeurs manquantes, puis entrer en détails pour voir les capteurs influant dans cette catégorie i.e ceux qui contiennent plus de discontinuité.

Nous pouvons ainsi, continuer à chercher plus en profondeur pour isoler ces derniers et visualiser ses comportements temporels. L'interactivité nous permet, dans ce cas, d'avoir des informations sur d'autres dimensions pour les capteurs dernièrement isolés, i.e., nous pouvons vérifier leurs variabilité, stabilité, etc. Par ailleurs, nous pouvons aussi faire un zoom-out, i.e., faire un retour arrière de la sélection, par exemple pour former des visions agrégées soit sur la dimension "Discontinuité" ou bien sur d'autres dimensions.

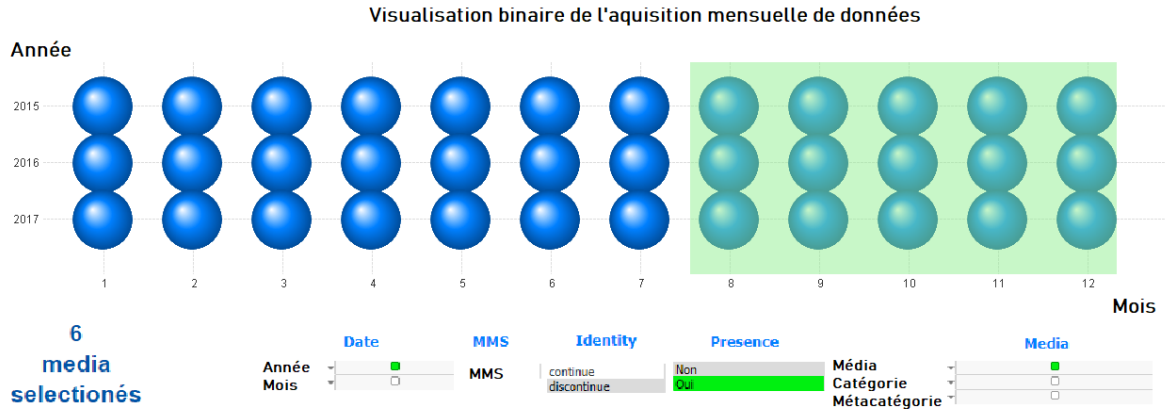


FIGURE 9.7 – Information sur la variabilité de la récolte

La figure 9.7 montre une information visualisée par un indicateur binaire mettant en œuvre l'aspect "Présence/Absence" i.e si une donnée existe, nous lui attribuons un poids de 1, 0 sinon, où une bulle représente un présence. Nous montrons qu'une première sélection

par l'application des filtres permet de donner une visualisation mettant en œuvre un cumul des poids i.e une somme sur les 0 ou 1. Pour explorer les détails de cette sélection, nous pouvons utiliser le principe de zoom in/zoom out sur la période du temps qui représente visuellement le plus de soucis. Ainsi, si la somme des poids est inférieure aux autres, il est largement remarquable. L'outil permet aussi d'analyser cet effet sur les autres dimensions et objets visuels.

9.6.2 Pensée visuelle

La technique de la pensée visuelle a pour objectif d'acheminer la pensée d'un utilisateur jusqu'à ce qu'il trouve ces souhaits. En d'autres termes elle permet de mettre en œuvre tous les objets visuels facilitant l'aboutissement à l'information.

MMS Explore fournit les moyens et techniques dans ce sens soit sur le plan *inter-dashboards*, i.e lors d'une navigation entre les tableaux de bord, ou bien dans la dimension d'étude même, i.e. dans un tableau de bords.

FIGURE 9.8 – KPI de la dimension stabilité

La figure 9.8 montre le tableau de bord associé à la dimension "Stabilité", nous remarquons que les objets visuels présents entourent bien cet aspect. Ils offrent à l'utilisateur tous les moyens nécessaires pour analyser cette dimension, e.g. il peut faire une étude sur diverses variables, changer les paramètres, changer l'échelle, être alerté en cas de pro-

blèmes. Divers indicateurs clés de performance sont ainsi présents. Des KPI informent sur la progression temporelle de la stabilité. D'autres, de nature statique, fournissent des pourcentages sur les capteurs les plus stables, etc.

À noter que si une sélection est faite sur un endroit précis, tous les objets visuels changent en fonction de cette personnalisation e.g le niveau d'alerte, les indices statistiques, etc.

9.7 Indicateurs et tableaux de bord

Dans cette partie nous donnons un aperçu sur quelques indicateurs clés de performance développés et intégrés dans MMS Explore ainsi que les tableaux de bord qu'ils contiennent.

9.7.1 Indicateurs visuels associés

Afin de permettre aux experts d'évaluer la qualité des données, nous avons construit un outil de visualisation fondé sur deux types d'indicateurs :

— Indicateurs statiques fondés sur les indices cités auparavant (variabilité interne, variabilité externe, stabilité) et mesurant des données sur un intervalle de temps précis.

— Indicateurs dynamiques reposant sur une agrégation des indices précédents au cours du temps permettant de visualiser des données temporelles.

Leur utilisation a pour objectifs de :

— Informer sur l'absence des données dans les séries temporelles de données imparfaites. Autrement dit, nous cherchons à trouver les capteurs qui souffrent de discontinuité dans leur fonctionnement dans une période T et à en superviser leur comportement durant cette période en nous reposant sur des scores.

— Avoir des valeurs clés sur la variabilité du recueil, soit dans une période précise ou à un instant t, où t est divisible en sous-périodes. L'idée est de juger le comportement d'un capteur durant un mois donné en considérant les données acquises à l'échelle du jour. Un score d'agrégation sur le mois est ainsi calculé.

— Surveiller au cours du temps un recueil des données.

— Construire des indicateurs pour de possibles déclenchements d'alerte, par exemple si la récolte dévie de son fonctionnement normal ou, tout simplement, vis à vis de l'ensemble.

— Prévoir le prochain jugement sur la qualité des données.

— Informer sur les moments critiques.

— Juger la qualité sur plusieurs variables d'étude et selon diverses sources.

Dans ce qui suit, nous montrons certains indicateurs développés.

Indicateurs de discontinuité

Les méthodes de découverte de données fondées sur la visualisation permettent aux utilisateurs d'être informés sur l'absence de données d'une série temporelle. La non-détection des données est une information importante pour juger de la qualité de la récolte et des robots. La considération de cette information à travers des indicateurs peut aider l'utilisateur à détecter des biais dans la récolte.

La figure 9.9 présente un tableau contenant un décompte de la détection complète dans une période donnée, conformément à un paramétrage en amont. La figure montre un histogramme indiquant l'évolution du nombre de sites Web non analysés par mois. Cela peut être utile, en cas de cycle périodique dans le processus de récolte, en déterminant une période dans l'année où la récolte souffre toujours de ce phénomène.

Par ailleurs, nous proposons aussi un indicateur d'alerte informant sur la détection d'une situation critique compte-tenu des analyses précédentes. L'indication affiche la couleur appropriée pour indiquer le degré de criticité (alerte de 1 à 4). L'idée est que plus il y a d'absences, plus le niveau d'alerte sera élevé. Soit $A = a_{t_1}, ..., a_{t_m}$ la variable qui indique pour chaque timestamp, le nombre de valeurs absentes. a_{t_m} représente le nombre de valeurs absentes à l'instant actuel. Notre indicateur d'alerte utilise directement la fonction quantile (éq. 9.1) avec $r = 4$. Ainsi les données dont le nombre d'absences est dans Q_1 seront en niveau 1, celles dans Q_2 seront de niveau 2, celles dans Q_3 de niveau 3, et celles dans Q_4 de niveau 4. D'autres répartitions sont possibles mais nous avons choisi celle-ci pour la simplicité de son interprétation.

$$AQ_r(a_{t_m}, A) = \left\lfloor \frac{Rang(a_{t_m}, A)}{(m/4)} \right\rfloor + 1 \tag{9.1}$$

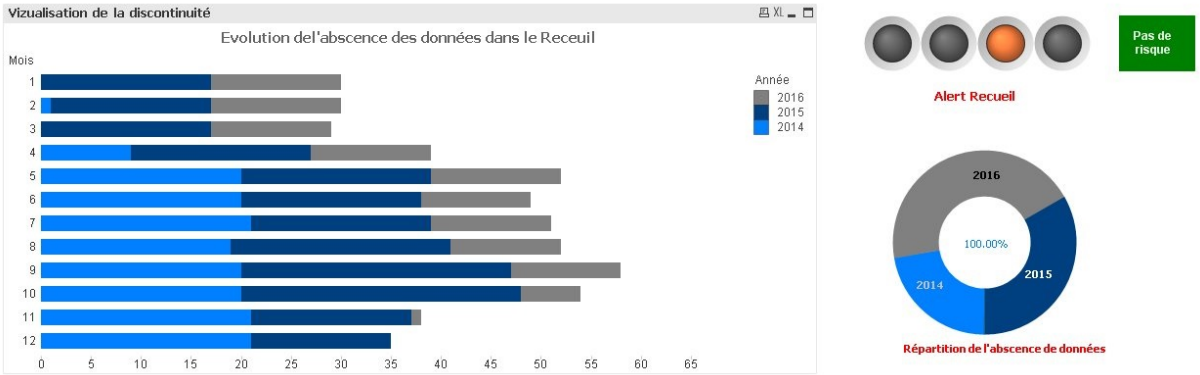


FIGURE 9.9 – Indicateurs développés informant sur l'absence de données

La figure 9.10 fournit une visualisation binaire de la présence de données par une présentation en bulles de même taille. Lorsqu'une bulle est affichée, cela signifie que des données existent à ce timestamp, et ce, quelle que soit la quantité de données récoltées. Lorsqu'aucune bulle n'est affichée, cela signifie qu'aucune donnée n'a été récoltée. Nous

avons suivi une représentation projetée sur une période mensuelle par année pour chaque média. Cette représentation permet aussi de visualiser le comportement d'un groupe de capteurs en sommant les tailles des bulles (voir figure 9.11), i.e. si la bulle est grande cela signifie qu'il y a beaucoup de sites web sur lesquels des données ont été récoltées au cours de la période. Cette présentation permet de déterminer, de manière simple, l'existence ou non de données sur un ou plusieurs sites web en même temps. En conséquence, cela permet de détecter les variations et les discontinuités dans les récoltes d'une collection de capteurs.

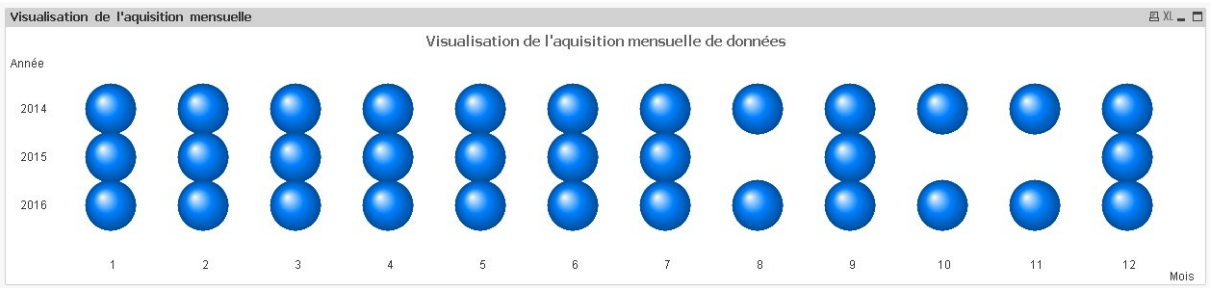


FIGURE 9.10 – Affichage de l'absence de données par mois durant 3 ans d'études

De plus, différentes échelles temporelles peuvent être considérées sur cette représentation : échelles mensuelles ou journalières. Une bulle affichée sur une échelle mensuelle peut masquer certaines informations sur la récolte pendant la période. Lorsque nous entrons dans une bulle mensuelle, nous pouvons découvrir plus de données manquantes dans les jours de ce mois. Cette possibilité de zoom dans les périodes temporelles offre la possibilité à l'utilisateur d'affiner ses analyses.

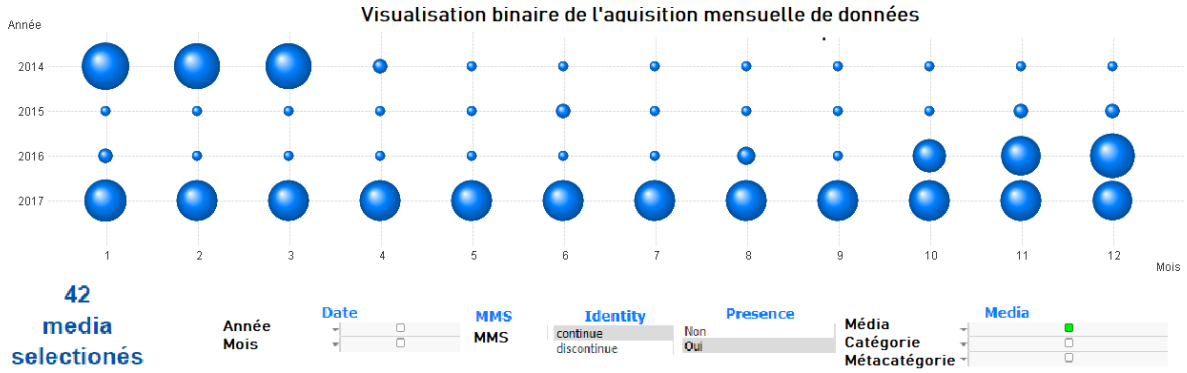


FIGURE 9.11 – Visualisation binaire de l'acquisition mensuelle de données pour un ensemble de capteurs. Les tailles des bulles sont proportionnelles au nombre de médias.

Indicateur sur le comportement de la récolte Afin de comprendre les comportements des capteurs tout en prenant en compte les imperfections dans la récolte, nous proposons un indicateur visuel fondé sur l'utilisation des quantiles. La visualisation des positions relatives, fournies par les quantiles d'appartenance, permet d'avoir une vision plus robuste des données.

La figure 9.12 présente cet indicateur. MMS Explore met en œuvre deux visions pour cet indicateur, la vision interne et l'externe.

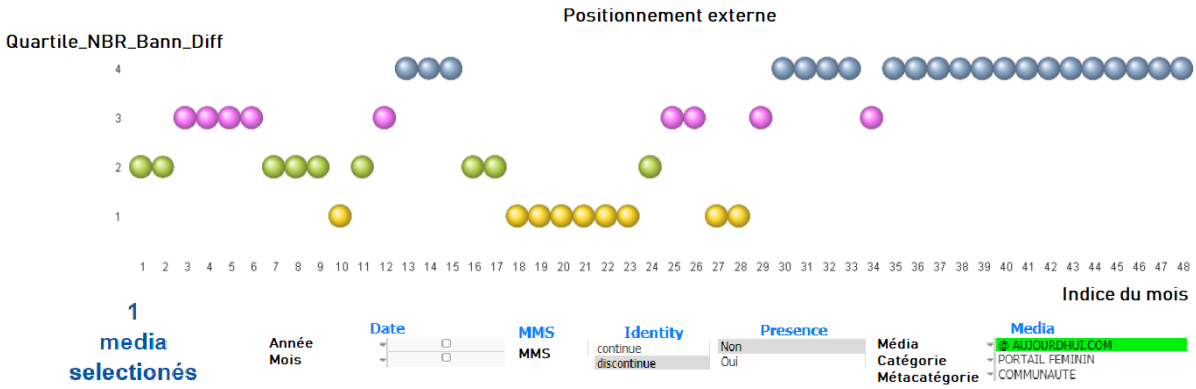


FIGURE 9.12 – Comprendre le comportement externe d'un capteur

Dans la figure 9.12, l'axe des ordonnées donne la valeur du quantile dans $[0, r]$, ici r est égal à 4 donc il s'agit d'une partition quartile. L'axe des abscisses représente le temps. La présence d'une bulle dans une position indique que pour cette variable il y a une donnée enregistrée par le capteur. La figure montre que sur une période de 36 mois consécutifs, on peut déterminer que sur la période du 18 au 23, ce capteur récolte moins de données par rapport aux autres. Son comportement dans cette période est cohérent, par contre, nous voyons qu'à partir du moment t=28 ce dernier a changé de comportement. On comprend aussi que le site web sur-lequel ce capteur pointe devient parmi les grands sites recevant le plus de publicité par rapport à l'ensemble de la récolte. Nous pouvons ainsi exploiter l'approche QBA pour quantifier cette variabilité.

Indicateurs de variabilité

La figure 9.13 montre un ensemble d'indicateurs qui traitent de la variabilité externe d'un groupe de capteurs sur une période T et selon trois variables d'étude différentes. Pour un instant t_k de la période, l'agrégation est faite par la moyenne des valeurs de la variabilité des trois variables.

Cet indicateur fondé sur Q_{ext} montre comment on détermine la variation d'un ensemble de capteurs. Le diagramme radar donne les mesures moyennes d'une collection par mois. Lorsque les traits se resserrent vers le centre, cela signifie qu'il y a moins de variabilité dans les données et inversement. Sur le graphique progressif, nous pouvons obtenir des

informations sur l'évolution de la variabilité. Lorsque la courbe baisse cela signifie que le processus de collecte des données devient plus robuste. Cela donne, donc, des informations sur la qualité du fonctionnement des robots. Les indicateurs statiques sont également présents et sont calculés automatiquement en suivant les personnalisations appropriées et les sélections des paramètres.

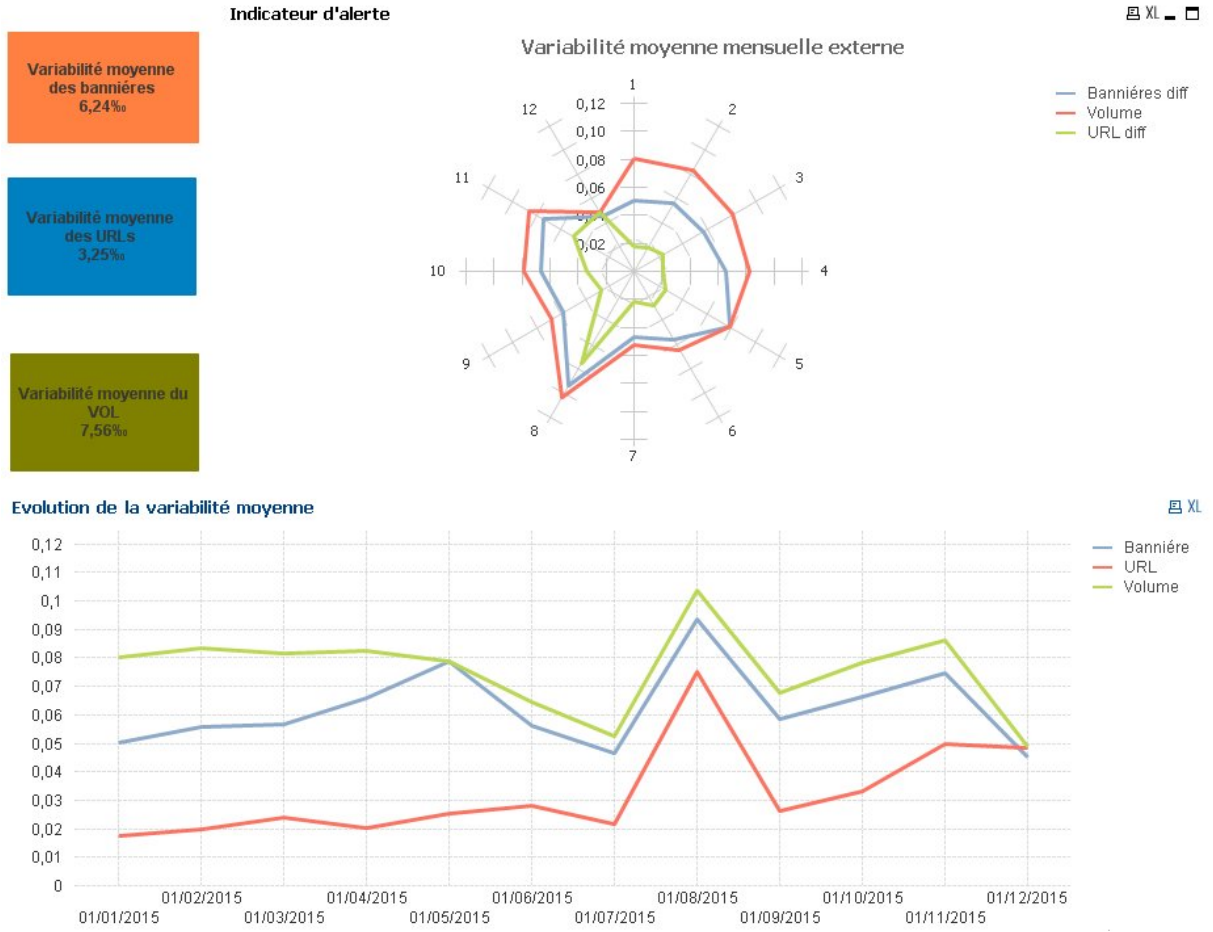


FIGURE 9.13 – Ensemble d'indicateurs informant sur la variabilité d'une récolte

D'autres indicateurs visuels de variabilité peuvent être, ainsi, construits. Par exemple, la figure 9.14 représente des nuages de points se référant à un score calculé de chaque série temporelle mettant en relation la variabilité interne et externe dans T. La distance de chaque point au centre représente la stabilité du capteur. L'indicateur est aussi personnalisé par des paramètres, e.g. choix des variables d'étude. Le nuage présenté montre ainsi la distribution de la variabilité de la récolte sur une période du temps pré-définie pour les différents capteurs étudiés.

Vu que cet indicateur est personnalisable, d'autres informations peuvent être distinguées comme par exemple des comportements singuliers, i.e des capteurs dont la récolte a une forte variabilité et donc qui est potentiellement défaillante.

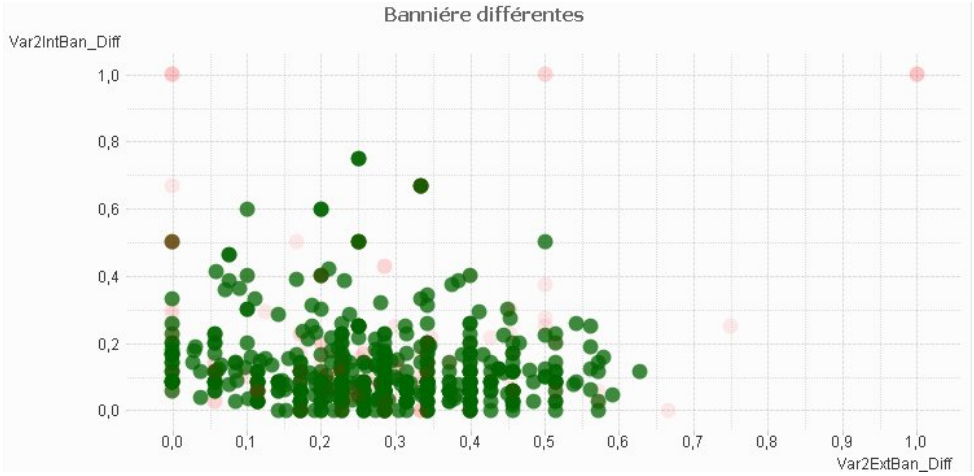


FIGURE 9.14 – Détection de la variabilité atypique dans la récolte

Ainsi, en suivant les techniques d'ingénierie visuelle, nous mettons cet indicateur en relation avec d'autres de même nature traitant les données sur d'autres variables d'études. Ainsi, lorsqu'un ensemble de points est sélectionné dans ce graphique, ces points varient dans les autres graphiques, ce qui permet d'avoir une vision multi-variée sur la distribution des points.

Nous voyons, dans la figure 9.14, un exemple dans lequel nous avons sélectionné des points sur une première variable. Ces données sont ensuite transformées automatiquement en vert. En cherchant si des changements apparaissent sur d'autres variables, nous pouvons ainsi obtenir une vision transversale de variabilité et comprendre certaines caractéristiques des capteurs, e.g. les points isolés et extrêmes pourraient être des capteurs défaillants.

En suivant cette logique, nous associons en complément à ces visualisations, des boîtes à moustaches (voir figure 9.15). Ces graphiques font référence à des méta-catégories et à des catégories de capteurs sélectionnés. Pour une catégorie ou une méta-catégorie donnée, la boîte à moustaches indique la distribution de l'indice de variabilité sur les capteurs sélectionnés. Cela permet, par exemple, de déterminer quel groupe de capteurs a la plus grande variabilité par rapport à l'ensemble des groupes.

Indicateur sur l'instabilité

En nous fondant sur l'indice d'instabilité St, nous pouvons aussi construire des indicateurs pour évaluer la qualité des données. Étant donné qu'un indice d'instabilité est une mesure de jugement sur le comportement d'un capteur dans une période T, nous pouvons ainsi concevoir un indicateur temporel sur une période plus large. Pour minimiser les erreurs d'interprétation, nous pouvons encadrer les valeurs par une enveloppe, pouvant être définie par l'utilisation de plus ou moins un ratio de l'écart type de l'ensemble à chacun des points.

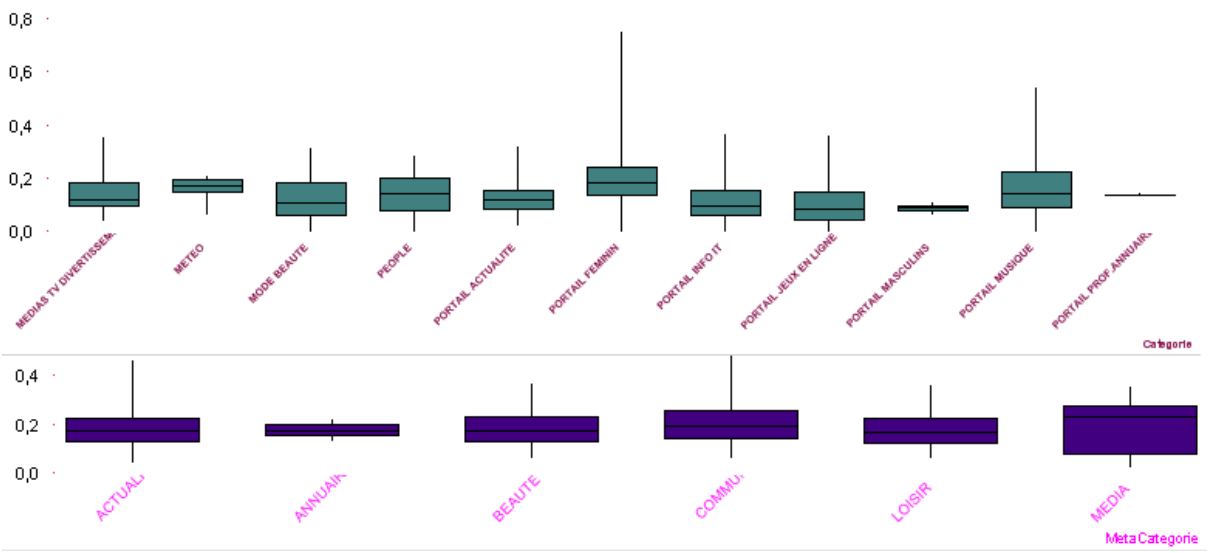
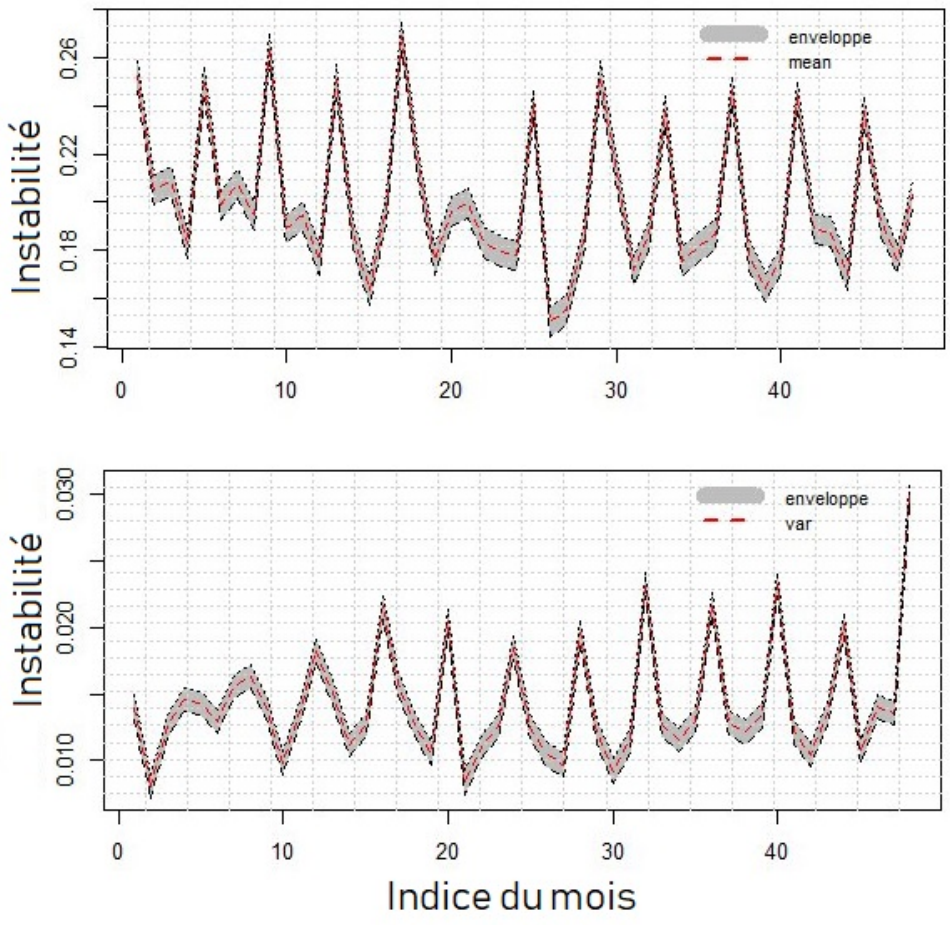


FIGURE 9.15 – Classification de la variation dans les catégories des sites web



Le calcul est effectué sur une base mensuelle. Dans chaque t, la mesure statistique appropriée (moyenne / variance) est appliquée. Les paramètres sont $r = 4$, $jp = 2$, $b = 1$ et $var = \{v_i\}$.

FIGURE 9.16 – Évolution de l'instabilité englobée par une enveloppe qui est égale à $\pm\frac{\sigma}{5}$.

Au vue que MMS Explore est en relation directe avec le noyau de fouille de données, ce dernier fournit des indicateurs à un usage scientifique. L'outil fait le lien avec le noyau en faisant appel à des scripts externes.

La figure 9.16 montre un indicateur visuel de l'instabilité de la récolte fournit par un prototype externe. Le signal visualisé est calculé pendant chaque instant dans T en suivant l'équation 6.7 et en associant une mesure statistique (ici on a choisi la moyenne et la variance). L'enveloppe qui entoure ce signal fait référence aux taux d'erreurs possibles dans les mesures.

Ce type d'indicateur est fondé sur les différents indices, introduits dans ce chapitre et gérant les imperfections de nos données. Il nous permet de distinguer des périodes d'instabilité et de stabilité dans le recueil, i.e, déterminer les moments critiques dans les cycles de la récolte des publicités.

9.7.2 Présentation de quelques tableaux de bord

Les tableaux de bords sont les conteneurs des objets visuels utilisés par notre outil. Ils rassemblent des indicateurs clé de performance, des filtres, des statistiques, etc. afin de fournir les aspects nécessaires permettant d'étudier une dimension.

La figure 9.17 montre un tableau de bord de la dimension variabilité. Les objets visuels offrent une vision sur le positionnement de la récolte sur une période du temps. Nous pouvons ainsi effectuer une sélection dans le nuage pour isoler les capteurs au comportement aberrant, et ce, en se fondant sur les scores Var_{int} et Var_{ext} données en abscisse et ordonnées respectivement.

Une telle sélection dans le nuage engendre à son tour des changements, non uniquement sur tous les objets visuels de ce tableau de bord, mais aussi sur les autres. La figure montre l'apparition des catégories et des méta-catégories plus importantes que d'autres, les KPI utilisés pour ce propos étant des boîtes à moustaches. Ces dernières donnent des informations sur les différentes variables d'études. Nous pouvons par ailleurs détailler cette variabilité en vérifiant chacun des constituants à part, i.e. en navigant dans le tableau de bord associé à la variabilité interne ou externe, etc.

MMS Explore adopte des tableaux de bord donnant la possibilité d'avoir un aperçu sur les données brutes et des statistiques usuelles. Ceci aide l'utilisateur à avoir une idée à l'issue d'une manipulation dans l'outil. Un utilisateur peut évaluer la qualité des capteur en utilisant QBA (e.g. pour le calcul des variabilités – Figure 9.17) ou aussi des approches statistiques plus classiques (e.g. figure 9.18), d'où la complémentarité entre les tableaux de bord.

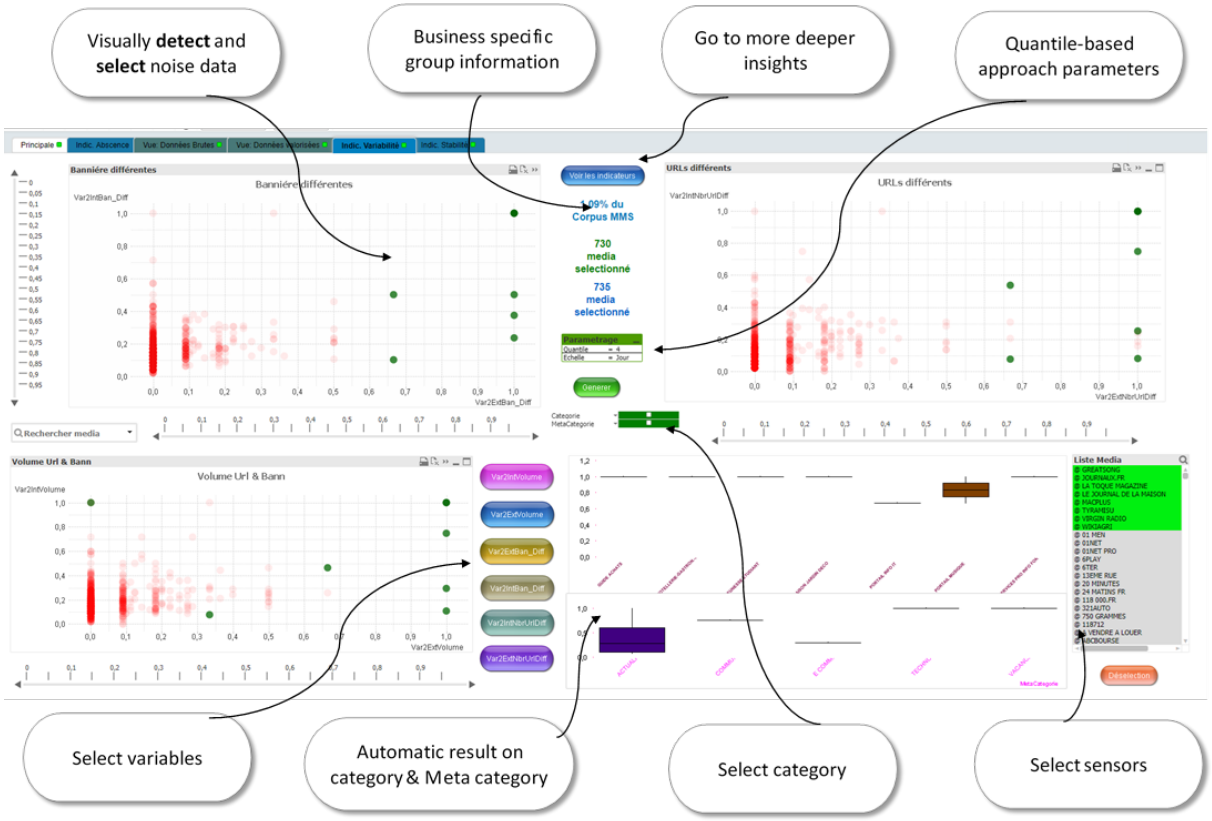


FIGURE 9.17 – Tableau de bord de la dimension variabilité

FIGURE 9.18 – Tableau de bord de la dimension présentant les valeurs brutes

La figure 9.18 présente un premier tableau de bord contenant des mesures sur les données brutes. Il permet de donner des mesures simples sur un recueil de données. L'utilisateur peut obtenir des informations plus précises en naviguant sur les objets visuels qu'il présente, e.g. en regardant l'outil *TOP 10*, on peut voir les 10 capteurs fournissant le plus de données, etc.

9.8 Cas d'utilisations

Dans cette partie nous présentons quelques cas d'utilisation pour montrer comment on peut tirer des conclusions sur le comportement de nos capteurs en utilisant MMSE.

9.8.1 Cas d'utilisation 1

Ce premier cas d'utilisation référant à la figure 9.19 a pour objectif de détecter les capteurs les moins efficaces et où leur utilisation peut influer négativement sur la production des chiffres des investissements publicitaires.

Nous pouvons commencer par faire une sélection dans le nuage de points du tableau de bord consacré à l'étude de la variabilité. Nous remarquons l'existence des points singuliers dans la figure 9.19a, i.e. des points qui s'éloignent bizarrement de tous les autres. Ces points sont en effet les capteurs de fortes variabilité sur la période et la variable d'étude choisies.

Par le biais des modes d'interactions visuelles intégrées dans l'outil, nous pouvons voir la dispersion de ce nuage des points sur les autres variables d'étude.

Cette interaction donne aussi lieu à une nouvelle construction automatique de boites à moustaches (Apparition de deux boîtes importantes sur la figure 9.19 b). Nous pouvons remarquer ainsi que l'une d'entre elles est plus importante, i.e. elle contient beaucoup de capteurs très variables en terme du comportement. En utilisant le principe de zoom-in nous pouvons cliquer dessus pour avoir plus de détails.

Cette méta catégorie contient les catégories et les sites Web les plus variables de la sélection du début. Nous pouvons voir leur importance dans la figure 9.19(c) respectivement.

Nous pouvons donc comprendre qu'une grande partie des capteurs qui pointent sur ces sites web sont l'objet d'une forte variation et que cette catégorie elle même nécessite certaines corrections. Nous pouvons essayer de comprendre certains comportements de ces robots associés. Pour cela, on va aller plus en profondeur pour en juger les fonctionnalités internes.

Nous pouvons voir dans la figure 9.19(d), que nous avons choisi un capteur important qui doit être examiné. Grâce aux graphiques de visualisation et l'approche QBA, nous pouvons évaluer la progression de sa récolte et alerter sur les possibles problèmes.

En fait, conformément à la projection en quantile, nous nous assurons que ce capteur contient des sauts importants pendant le processus de récolte. Cela peut déclencher une alerte indiquant la criticité de ce capteur.

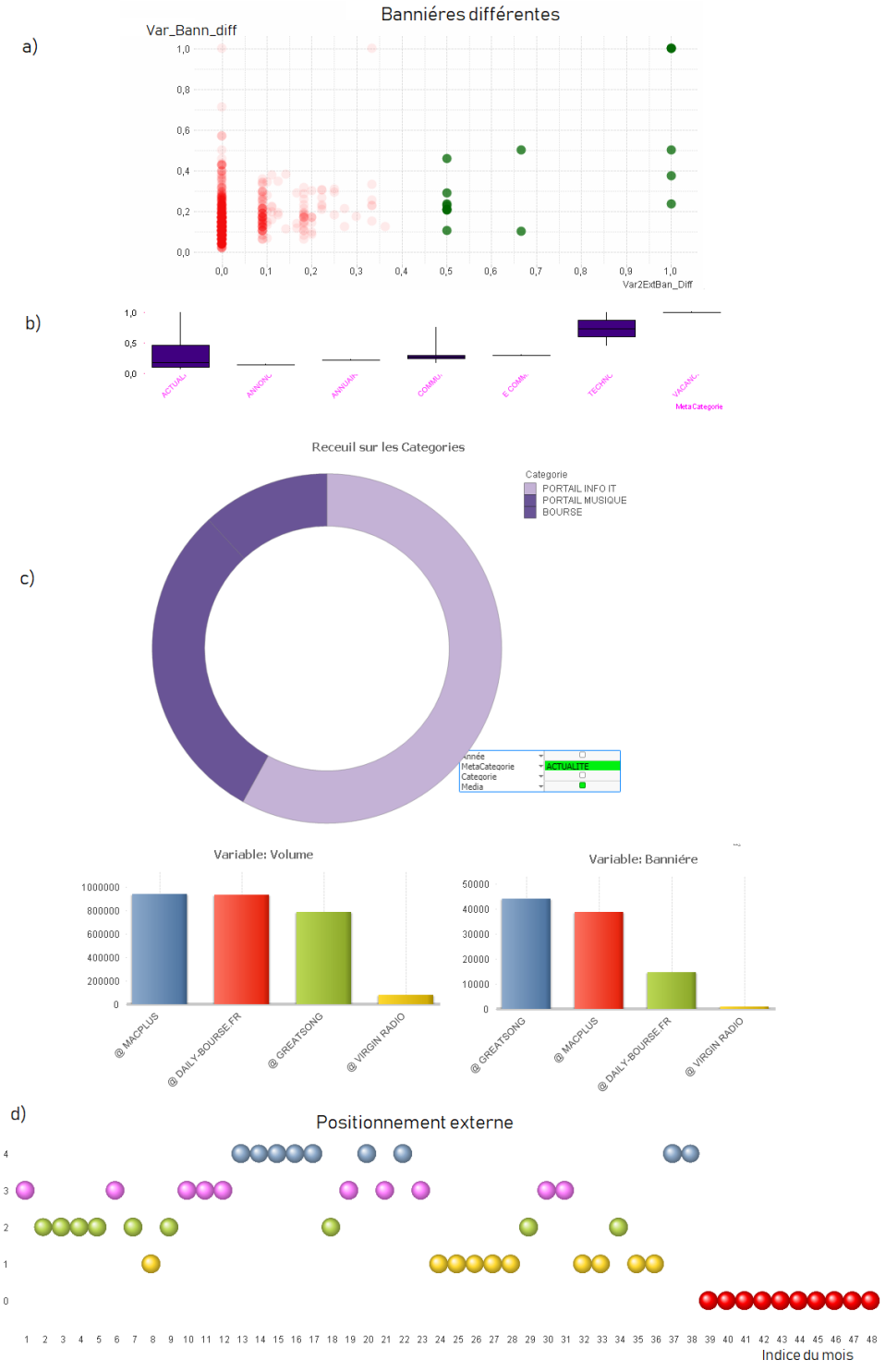


FIGURE 9.19 – Trouver les capteurs qui présentent une potentielle anomalie dans une catégorie

9.8.2 Cas d'utilisation 2

Dans ce cas d'utilisation, nous présentons des étapes de recherche acheminées sur la provenance des données manquantes dans une récolte volumineuse. Nous commençons par choisir une période, une source et un ensemble de médias. Visuellement nous pouvons

déterminer qu'il existe une baisse remarquable dans la récolte totale (voir figure 9.20(a)). Pour ce propos, nous pouvons ensuite faire un zoom-in pour voir les provenances temporelles, i.e. chercher s'il y des phénomènes saisonniers. Dans la figure 9.20(b), nous remarquons que l'année 2017 (en jaune) est l'année qui contient le plus d'absence de données par rapport aux années précédentes. En choisissant cette année, nous remarquons sur l'indicateur dynamique de variabilité (voir figure 9.20(c)) que la variabilité augmente fortement dans les derniers mois de l'année, et ce, sur les trois variables d'études. L'indicateur radar (voir figure 9.20(d)) confirme bien cette constatation visuelle en donnant plus d'indices sur la variation de cette sélection par une grille moyennant les résultats.

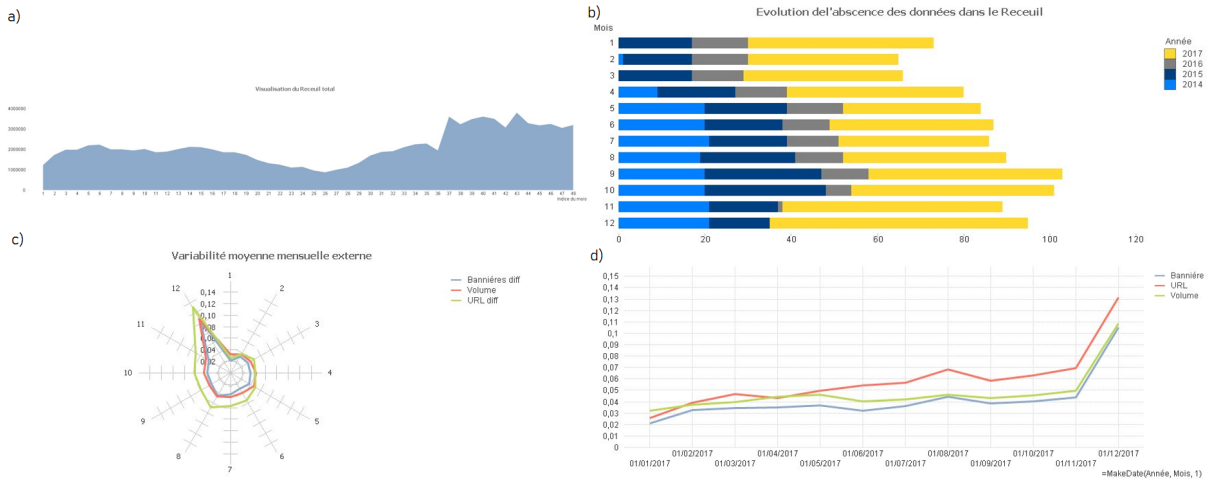


FIGURE 9.20 – Évaluation de la qualité de la récolte de 2017 par MMS Explore

9.9 Conclusion

Dans cette contribution, nous avons proposé un outil de visualisation interactive permettant d'aider un expert métier à comprendre et évaluer le comportement des capteurs. L'outil exploite un moteur d'analyse mettant en oeuvre différentes fonctionnalités. Il guide l'utilisateur dans son exploration de la qualité des données.

L'outil utilise les techniques de visualisations afin de facilité l'étude de la qualité de la récolte. MMS Explore est conçu à partir d'un modèle visant à présenter les différentes dimensions étudiées de la qualité au travers de tableaux de bord dédiés à chacune d'elles. Chaque tableau de bord porte sur un sujet particulier lié à une dimension particulière et traite ce sujet par l'intermédiaire d'un ensemble d'indicateurs visuels dédiés.

L'utilisateur peut donc étudier la qualité de la récolte selon plusieurs dimensions en utilisant les modalités d'interaction et la continuité de présentation du contenu de l'outil entre les différents tableaux de bord.

Les cas d'utilisations présentés, dans ce chapitre, montrent le potentiel de MMS Explore à fournir des résultats pertinents par le biais des indicateurs fournis notamment par notre approche QBA.

Notre outil a pour destination principale les experts métiers, et ne fournit pour l'instant que des fonctionnalités réduites d'utilisation de méthodes de fouille de données. Cette mise en relation de l'outil avec ces dernières est une possibilité d'amélioration technique de notre outil. L'introduction de l'approche FBA pourra aussi être intégrée afin de permettre d'exploiter ces positionnements flous.

Ces deux pistes peuvent être complétées par l'utilisation de méthodes de visualisation interactive de données multivariées présentées dans l'état de l'art afin d'avoir une vision résumant les données en combinant les différentes dimensions de la qualité.

Ce chapitre a présenté la dernière contribution de cette thèse. Dans la partie suivante, nous ferons le bilan des travaux que nous avons menés et nous en présenterons les perspectives envisagées.

Cinquième partie

Conclusion générale

Conclusion

Face à la volumétrie grandissante des données venant de différentes sources et à la leur mise en production, par les systèmes d'information des entreprises, sans une vérification préalable de leur qualité, le besoin d'avoir des moyens aidant à qualifier les enregistrements est primordial. En effet, les possibles imperfections des données peuvent influer directement sur la pertinence des prises de décisions finales et indirectement sur la véracité des informations valorisées par les entreprises.

Notre travail de thèse s'inscrit dans une démarche ayant pour objectif de mieux appréhender les données récoltées et leur qualité afin d'apporter des explications qui peuvent améliorer la véracité et implicitement la valeur des résultats finaux. Dans ce contexte, nous nous sommes focalisés sur la problématique de l'analyse de la qualité au travers notamment de la variabilité et de la stabilité des flux multi-capteurs.

Ces capteurs, chez Kantar, fournissent des données formant, une fois agrégées, des séries temporelles incomplètes, imprécises et difficilement comparables. Les approches de la littérature d'analyse de la qualité, et plus précisément de la variabilité et de la stabilité, ne considèrent pas l'ensemble des caractéristiques de nos données.

Nous dressons, ici, le bilan des principales contributions de ce travail de recherche. Ce travail de thèse a permis de faire avancer la connaissance sur l'étude de la qualité des séries de données temporelles et imparfaites. Notre recherche ouvre vers des perspectives qui pourront être considérées dans le futur.

Bilan des contributions

Nous avons trois principales contributions dans ce travail de thèse : l'approche QBA (approche par quantiles), l'approche FBA (approche par la logique floue) et l'outil de visualisation. Dans ce bilan, nous proposons une présentation transversale de nos contributions.

Étude de la qualité des données et de la récolte

À partir de l'état de l'art, nous avons identifié les dimensions clés dans l'étude de la qualité de séries de données récoltées sur le web :

- l'imperfection des données :
  - la discontinuité et l'incomplétude associée,
  - l'imprécision ;
- la cohérence des récoltes :
  - la variabilité,
  - la stabilité.

Au regard de ces dimensions et de l'étude de nos données, nous avons proposé dans cette thèse deux approches, FBA et QBA pour l'analyse de la variabilité et de la stabilité des séries temporelles de données imparfaites, et un outil de visualisation interactive permettant à l'utilisateur de mieux comprendre la qualité des données.

Ces approches et cet outil exploitent l'idée de positionner chaque donnée à la fois dans la série temporelle du capteur qui l'a fournie, et dans la série des données récoltées par l'ensemble des capteurs au même moment. L'étude des trajectoires des positionnements des données donne la possibilité de mieux appréhender la cohérence de la récolte pour ce flux.

Indices de positionnement des données considérant leur imprécision et incomplétude

Nos données sont imprécises et incomplètes. De plus, elles sont en l'état peu comparable entre elles pour des questions de rapport d'échelle. C'est pourquoi nous avons proposé dans cette thèse deux approches pour le positionnement des données imparfaites dans les séries temporelles : les approches QBA et FBA.

Dans ces deux approches, les données manquantes des séries incomplètes (souffrant de discontinuité) sont représentées par une valeur de positionnement particulière. Par ailleurs, elles sont exclues des valeurs considérées dans les calculs de positionnements présentés dans nos approches.

Dans l'approche QBA, la représentation des données tenant compte de leur imprécision se fait à l'aide d'une représentation ensembliste en leur associant les quantiles auxquels elles appartiennent. Cette représentation en quantile à l'avantage de construire une échelle de valeur réduite où toutes les données deviennent comparables entre elles. L'autre avantage est que chaque valeur de l'échelle représentera un volume de données initiales similaire. Dans cette approche la finesse du positionnement dépend du nombre de quantiles à considérer. Nous pouvons, d'ailleurs, faire varier la valeur de ce paramètre dans notre outil de visualisation.

Cependant, la représentation en ensembles stricts pose le problème de l'appartenance pleine des éléments aux frontières. En effet, ces derniers sont possiblement plus proches de données d'un autre quantile que de celles du sien.

Pour gérer cela, notre seconde approche (appelée FBA), qui peut être vue comme une généralisation de la première, repose sur la notion d'appartenance partielle à l'aide de représentations floues des valeurs et des ensembles, formant l'échelle de valeurs représentées de manière imprécise. Dans FBA, le positionnement flou d'un élément dans un ensemble est défini à l'aide du degré d'appartenance de l'élement flou associé aux clusters flous, en considérant aussi la position de chaque cluster flou. Nous avons proposé de regrouper les

données à l'aide d'une distance maximale entre deux données de valeurs successives dans chaque cluster. Cette approche dépend non plus du nombre de quantiles (ou du nombre de cluster) mais de la distance et de la fonction de fuzzification pour chaque élément.

Afin d'étudier les positionnements des données et leur évolution dans nos récoltes, nous tenons compte, dans ces deux approches, de deux visions : interne et externe. La vision interne permet d'observer le comportement d'un capteur envers lui-même, i.e. observer la cohérence intrinsèque de la récolte fournie par ce capteur. La vision externe permet d'étudier le comportement d'un capteur à partir de son positionnement par rapport aux autres, i.e. étudier la cohérence de la récolte fournie par ce capteur vis-à-vis des récoltes fournies par l'ensemble des capteurs. Notre outil de visualisation permet une analyse visuelle de l'évolution de ces positionnements.

Notre méthodologie permet aussi de considérer différentes agrégations temporelles. En permettant de naviguer entre ces échelles d'agrégations temporelles, notre outil offre la possibilité de mener une étude visuelle et interactive permettant de spécialiser ou de généraliser la vision d'un phénomène.

Indices de variabilité et de stabilité

Afin d'étudier la variabilité et la stabilité de nos données en considérant l'ensemble de leurs caractéristiques, nous nous sommes appuyés sur les positionnements précédemment obtenus.

Nous considérons, en premier lieu, qu'une transition (passage) entre deux timestamps est remarquable si la variation entre les positions des données récoltées à ces deux timestamps est suffisamment notable. Pour cela, nous utilisons un score binaire qualifiant la transition. Ce score vaut 1 si la variation est supérieure à un seuil, 0 sinon. Ce score tient aussi compte des valeurs manquantes dans les données, en différenciant les périodes longues d'interruption des périodes courtes. Cette définition des périodes longues est paramétrique.

La variabilité sur une période donnée est alors calculée comme le ratio entre le volume des transitions remarquables sur celui de l'ensemble des transitions. Nous voyons donc la variabilité comme la mesure des passages notables. Plus une série est composée de données dont les positions varient d'un timestamp au suivant, plus la variabilité sera grande.

Nous avons aussi gardé les deux visions (interne et externe) pour l'étude de la variabilité. Les variabilités internes et externes informent donc sur le comportement du capteur. Par exemple, une faible variabilité interne mais une forte variabilité externe, peut indiquer que le capteur en question n'est pas assujetti aux mêmes stratégies de récolte que les autres capteurs. À l'inverse, une variabilité interne forte et externe faible indiquent que la variabilité interne est surtout due aux tendances générales de la récolte de données.

La stabilité est construite par agrégation des variabilités interne et externe à l'aide d'un opérateur d'agrégation. Nous avons proposé d'utiliser la distance euclidienne et normalisée du point de coordonnées (variabilité interne, variabilité externe) à l'origine (0,0). Une comparaison sur un exemple de cet opérateur d'agrégation avec d'autres opérateurs de la littérature a montré, pour cet exemple, que notre opérateur propose une valeur au centre des valeurs issues des différentes agrégations. Cependant, il serait sans doute pertinent de faire une étude spécifique sur le choix de l'opérateur d'agrégation à utiliser.

Dans notre outil de visualisation, nous offrons la possibilité d'exploiter ces indices en construisant des indicateurs statiques indiquant la valeur de ces indices pour une période donnée, et des indicateurs dynamiques caractérisant l'évolution de ces indices dans le temps.

MMS Explore

Afin de permettre aux agents de Kantar de mieux appréhender la qualité des données, nous avons proposé, dans cette thèse, un outil d'analyse visuelle interactive.

Cet outil permet une exploration allant du plus général sur l'ensemble des données vers du plus spécifique sur un sous ensemble de données, voire sur un flux particulier pour une sous-période donnée.

Il repose sur des techniques d'analyses visuelles permettant de juger de la qualité. Pour cela, nous avons utilisé deux techniques principales : l'interactivité visuelle au travers principalement du zoom in/zoom out et la pensé visuelle. La première permet d'avoir des visualisations représentant des informations plus détaillées sur une sélection de données ou en plus générales sur l'ensemble des données. La seconde permet de fournir les objets visuels nécessaires à l'étude d'une dimension de qualité, e.g variabilité, stabilité. Chaque dimension de la qualité peut être étudiée par un tableau de bord spécifique. Les visualisations présentées sur les différents tableaux de bord fournissent des informations complémentaires entre-elles.

Tout en mettant en évidence les résultats de notre approche QBA au travers de tableaux de bord dédiés à l'étude des positionnements en quantiles, à la variabilité et à la stabilité, l'outil fournit des indicateurs informant sur les absences/lacunes dans la récolte. Il permet aussi d'exploiter la classification, faite par la société, organisant les médias et publicités en méta-catégories et catégories.

Notre outil a pour but d'aider les agents à valider un recueil de données et de les informer sur les capteurs ayant un comportement anormal. Il permet de chercher une information précise sur les flux de données imparfaites et sur leur qualité, d'évaluer la récolte selon une personnalisation spécifique, e.g. temporelle ou catégorielle, tout en fournissant des indicateurs clés de performance pour chacune des dimensions étudiées.

Cet outil mériterait d'être complété par des fonctionnalités de fouille de données plus avancées.

Perspectives

Nous présentons ici des possibles perspectives à notre travail.

Vers une recommandation du calibrage des paramètres des approches QBA et FBA

Nos deux approches QBA et FBA sont paramétriques. Notre étude de sensibilité a montré que les paramètres ont clairement une influence sur la valeur de la variabilité. Une étude plus approfondie des impacts est une première perspective, par exemple, pour savoir si ces paramètres peuvent modifier les classements des séries selon leur variabilité.

Nous envisageons, de plus, de définir une méthode de calibration automatisée des paramètres de nos approches. Cet objectif pose premièrement la question de qu'est-ce qu'une bonne variabilité. C'est à dire, quels indicateurs de performance de la mesure de variabilité devons nous considérer pour déterminer si un paramétrage est meilleur qu'un autre ? Qu'est-ce qu'un paramétrage idéal ? Si de tels indicateurs de performances n'existent pas, lesquels devons-nous proposer ?

Nous pouvons d'ores et déjà considérer qu'un indicateur unique de performance n'est pas envisageable. Aussi, plutôt que de paramétrer automatiquement notre système, il pourrait être pertinent de proposer à l'utilisateur des paramétrages pertinents. Cette recommandation pourrait reposer sur les utilisations fréquentes de l'outil mais aussi sur les indicateurs de performance et des problématiques de "fitness for use". On se retrouve alors sur des problématiques de recommandations multicritères et multi-parties prenantes, qui sont d'actualité [SBM18 ; Bur+19].

Analyse multivariée, multi-sources et multi-axes

Nos deux approches de la variabilité, proposées dans ce manuscrit, étudient les séries temporelles de données imparfaites sur une seule variable d'étude. Or, nos données d'études sont d'ores et déjà définies sur trois variables. De plus, les données publicitaires sont multi-axes : elles peuvent être étudiées sur des sujets différents, e.g. la volumétrie, l'audience, le contenu. Elles sont aussi multi-sources. Dans notre étude, nous avons traité principalement l'axe "volumétrie" pour une source agrégée. Ainsi, considérer ces aspects peut amener à augmenter le nombre de variables et surtout à construire des séries ayant des comportements autres que ceux observés actuellement. L'idée d'avoir une approche multi-variée fait donc sens.

L'analyse entre les différentes variables peut, actuellement, se faire grâce aux possibilités de navigation offertes par l'outil de visualisation. Cependant, ne pas avoir d'indicateurs et d'indices directement multivariés est une limite pour tendre à une meilleure perception globale des possibles problèmes.

Une première piste pourrait être de travailler sur une composante résumant l'ensemble des variables. Pour cela, différentes approches de réduction de la dimensionnalité, à l'instar de l'Analyse en Composantes Principales, ou de sélection de variables peuvent être utilisées. Dans ce contexte, plusieurs questions se posent. Quelle serait l'approche la plus pertinente ? La réduction doit elle se faire en amont du traitement dans nos approches ou après le positionnement dans chacune des variables ? N'a-t-on pas un risque trop important d'une perte d'information ?

Une autre piste serait de travailler à la construction d'un score sur les passages non plus univarié mais multi-varié. Ici, les questions suivantes se posent encore. Quel opérateur d'agrégation choisir ? Le max ? La moyenne ? Une moyenne pondérée ? Si cette dernière est choisie, quels seraient les paramètres idéaux et pourquoi ?

Ainsi, l'analyse multivariée pose de nouvelles questions que nous envisageons de traiter dans nos futurs travaux.

Étude de l'influence de la qualité des données sur la véracité des résultats finaux

Nous rappelons que l'objectif de Kantar est de proposer à ses clients des estimations les plus fiables possibles sur les montants des investissements publicitaires.

Dans cette thèse, nous nous sommes concentrés sur l'étude de la qualité des données d'entrée, et nous avons pu isoler de potentielles anomalies dans les fonctionnements des capteurs. Les cas aberrants, confirmés par les experts, ont possiblement une influence négative sur les résultats finaux. Une perspective de ce travail est donc d'analyser l'impact de la qualité observée des données sur la véracité des estimations finales obtenues à partir de ces données.

L'hypothèse est qu'il existe un lien entre la qualité des données d'entrée et la véracité des valeurs en sortie. Nous souhaitons donc mener une étude expérimentale sur les possibles corrélations et relation de causalité.

À partir de cette étude d'impact, nous pourrions essayer de recommander des stratégies à mettre en place pour améliorer la véracité des résultats au regard de la qualité des données. Cette étude pourrait ainsi fournir des pistes pour ajuster les valeurs fournies par Kantar.

Ouverture vers d'autres usages

Ce travail de thèse s'est déroulé en partenariat avec une entreprise qui a pour activité principale la veille concurrentielle sur le marché médiatique. Kantar repère les messages publicitaires diffusés, les valorise et les catégorise afin de concevoir des campagnes publicitaires.

Le travail actuel propose des réponses pour l'étude de la qualité des données. Ces approches peuvent aussi être utilisées dans l'étude du positionnement des concurrents dans le marché médiatique. L'étude de la variabilité et de la stabilité pourrait alors être utilisée : pour étudier le comportement d'une activité ou une campagne, pour les classifier par exemple par émergence ou importance, pour recommander des actions à mettre en place pour faire face à une activité concurrente, pour informer sur une activité importante ou pour établir des prévisions stratégiques.

Enfin, les approches et l'outil proposé sont aussi exploitables dans d'autres domaines ou l'analyse de la qualité de données multi-capteurs est importante. Nous pourrions, par exemple, exploiter nos méthodes et outils pour un autre projet du CReSTIC pour l'analyse de la qualité des données de communications véhiculaires [Fou+17 ; NLF18 ; Leb+19].

Publications personnelles

Notre travail de thèse a donné lieu à trois communications dans des conférences internationales [Ben+19b ; Ben+19d ; Ben+18b] et trois dans des conférences ou ateliers francophones [Ben+19c ; Ben+19a ; Ben+18a].

Liste des publications effectuées

[Ben+18a] Zied Ben Othmane, Damien Bodénès, Amine Aït Younes et Cyril De Runz, « Vers une nouvelle interface visuelle dédiée à l'analyse des récoltes multisources de données », *dans* : *AtelierVIF@EGC2018*, Paris, France, 2018, p. 1–4, URL : https://hal.archives-ouvertes.fr/hal-01924255.

[Ben+18b] Zied Ben Othmane, Damien Bodenes, Cyril de Runz et Amine Aït Younes, « A Multi-sensor Visualization Tool for Harvested Web Information : Insights on Data Quality », *dans* : *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*, sous la dir. d'Ebad Banissi et al., IEEE Computer Society, 2018, p. 110–116, DOI : 10.1109/iV.2018.00029, URL : https://doi.org/10.1109/iV.2018.00029.

[Ben+19a] Zied Ben Othmane, Cyril De Runz, Amine Ait Aït Younes et Vincent Mercelot, « MMS Explore : un outil de visualisation interactive pour l'analyse qualité de flux données temporelles », *dans* : *Conférence Extraction et Gestion de Connaissances 2019*, t. RNTI-E-35 (Revue des Nouvelles Technologies de l'Information-E-35)), EGC 2019, Metz, France, 2019, p. 445–448, URL : https://hal.archives-ouvertes.fr/hal-02050016.

[Ben+19b] Zied Ben Othmane, Cyril De Runz, Amine Aït Younes et Vincent Mercelot, « Effect of Imprecise Data Income-Flow Variability on Harvest Stability : A Quantile-Based Approach », *dans* : *International Conference on Database and Expert Systems Applications - DEXA 2019*, sous la dir. d'Hartmann S., Küng J., Chakravarthy S., Anderst-Kotsis G., Tjoa A. et Khalil I., t. 11706, Database and Expert Systems Applications. Lecture Notes in Computer Science, Linz, Austria : Springer, Cham, 2019, p. 248–257.

[Ben+19c] Zied Ben Othmane, Cyril De Runz, Amine Aït Younes et Vincent Mercelot, « Quantifier la variabilité de séries temporelles de données imprécises », *dans* : (2019).

[Ben+19d] Zied Ben Othmane, Cyril De Runz, Amine Aït Younes et Vincent Mercelot, « Quantify the Variability of Time Series of Imprecise Data », *dans : International Conference on Flexible Query Answering Systems - FQAS2019*, t. 11529, Flexible Query Answering Systems, Amantea, Italy, 2019, p. 203–214, DOI : 10.1007/978-3-030-27629-4_20, URL : https://hal.archives-ouvertes.fr/hal-02293136.

Bibliographie

Références

[AES05] Robert A. Amar, James Eagan et John T. Stasko, « Low-Level Components of Analytic Activity in Information Visualization », *dans* : *IEEE Symposium on Information Visualization (InfoVis 2005)*, sous la dir. de John T. Stasko et Matthew O. Ward, Minneapolis, MN, USA : IEEE Computer Society, 2005, p. 111–117, doi : 10.1109/INFVIS.2005.1532136, url : https://doi.org/10.1109/INFVIS.2005.1532136.

[AFG98] Robert J. Adler, Raisa E. Feldman et Colin Gallagher, « A Practical Guide to Heavy Tails », *dans* : (1998), sous la dir. de Robert J. Adler, Raisa E. Feldman et Murad S. Taqqu, p. 133–158, url : http://dl.acm.org/citation.cfm?id=292595.292606.

[AFN98] AFNOR, *Prestations de veille – Prestations de veille et prestations de mise en place d'un système de veille*, FR, avr. 1998, url : https://www.boutique.afnor.org/norme/xp-x50-053/prestations-de-veille-prestations-de-veille-et-prestations-de-mise-en-place-d-un-systeme-de-veille/article/708892/fa047502.

[AL15] A. Agnello et H. Levkowitz, « Quick Vis : A Web-Based Visualization Delivering Flexible Exploration of User-Driven Analytics », *dans* : *2015 19th International Conference on Information Visualisation*, Barcelona, Spain, juil. 2015, p. 468–473, doi : 10.1109/iV.2015.84.

[AMS01] Paolo Atzeni, Paolo Merialdo et Giuseppe Sindoni, « Web site evaluation : Methodology and case study », *dans* : *International Conference on Conceptual Modeling*, Springer, Los Angeles, LA, USA, 2001, p. 253–263.

[AR14] *Effects of visualizing missing data : an empirical evaluation*, IEEE, Paris, France, 2014, p. 132–138.

[Aub+03] David Auber, Yves Chiricota, Fabien Jourdan et Guy Melançon, « Multiscale Visualization of Small World Networks », *dans* : *Proceedings of the Ninth Annual IEEE Conference on Information Visualization*, INFOVIS'03, Seattle, Washington : IEEE Computer Society, 2003, p. 75–81, isbn : 0-7803-8154-8, url : http://dl.acm.org/citation.cfm?id=1947368.1947385.

[Bat+09] Carlo Batini, Cinzia Cappiello, Chiara Francalanci et Andrea Maurino, « Methodologies for Data Quality Assessment and Improvement », *dans* : *ACM Comput. Surv.* 41.*3* (juil. 2009), p. 1–52, issn : 0360-0300, doi : 10.1145/1541880.1541883, url : http://doi.acm.org/10.1145/1541880.1541883.

[Bat18] Carlo Batini, « Data Quality Assessment », *dans* : *Encyclopedia of Database Systems, Second Edition*, sous la dir. de Ling Liu et M. Tamer Özsu, Springer, 2018, doi : 10.1007/978-1-4614-8265-9_107, url : https://doi.org/10.1007/978-1-4614-8265-9%5C_107.

[BB15a] Laure Berti-Équille et Javier Borge-Holthoefer, *Veracity of Data : From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics*, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2015, doi : 10.2200/S00676ED1V01Y201509DTM042, url : https://doi.org/10.2200/S00676ED1V01Y201509DTM042.

[BB15b] Laure Berti-Équille et Javier Borge-Holthoefer, « Veracity of Data : From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics », *dans* : Synthesis Lectures on Data Management (2015), doi : 10.2200/S00676ED1V01Y201509DTM042, url : https://doi.org/10.2200/S00676ED1V01Y201509DTM042.

[BB96] David B Buller et Judee K Burgoon, « Interpersonal deception theory », *dans* : *Communication theory* 6.*3* (1996), p. 203–242.

[BBK90] A Bardossy, I Bogardi et WE Kelly, « Kriging with imprecise (fuzzy) variograms. I : Theory », *dans* : *Mathematical Geology* 22.*1* (1990), p. 63–79.

[BC14] Giorgio Battistelli et Luigi Chisci, « Kullback–Leibler average, consensus on probability densities, and distributed state estimation with guaranteed stability », *dans* : *Automatica* 50.*3* (2014), p. 707–718.

[BCS96] Andreas Buja, Dianne Cook et Deborah F Swayne, « Interactive high-dimensional data visualization », *dans* : *Journal of computational and graphical statistics* 5.*1* (1996), p. 78–99.

[BCW87] Richard A Becker, William S Cleveland et Allan R Wilks, « Dynamic graphics for data analysis », *dans* : *Statistical science* (1987), p. 355–383.

[Bea04] Michel Beaudouin-Lafon, « Designing interaction, not interfaces », *dans* : *Proceedings of the working conference on Advanced visual interfaces, AVI 2004*, sous la dir. de Maria Francesca Costabile, Gallipoli, Italy : ACM Press, 2004, p. 15–22, doi : 10.1145/989863.989865, url : https://doi.org/10.1145/989863.989865.

[Bea05] Jeffrey Beall, « Metadata and data quality problems in the digital library », *dans* : *Journal of Digital Information* 6.*3* (2005).

[Ben+18a] Zied Ben Othmane, Damien Bodénès, Amine Aït Younes et Cyril De Runz, « Vers une nouvelle interface visuelle dédiée à l'analyse des récoltes multisources de données », *dans* : *AtelierVIF@EGC2018*, Paris, France, 2018, p. 1–4, URL : https://hal.archives-ouvertes.fr/hal-01924255.

[Ben+18b] Zied Ben Othmane, Damien Bodenes, Cyril de Runz et Amine Aït Younes, « A Multi-sensor Visualization Tool for Harvested Web Information : Insights on Data Quality », *dans* : *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*, sous la dir. d'Ebad Banissi et al., IEEE Computer Society, 2018, p. 110–116, DOI : 10.1109/iV.2018.00029, URL : https://doi.org/10.1109/iV.2018.00029.

[Ben+19a] Zied Ben Othmane, Cyril De Runz, Amine Ait Aït Younes et Vincent Mercelot, « MMS Explore : un outil de visualisation interactive pour l'analyse qualité de flux données temporelles », *dans* : *Conférence Extraction et Gestion de Connaissances 2019*, t. RNTI-E-35 (Revue des Nouvelles Technologies de l'Information-E-35)), EGC 2019, Metz, France, 2019, p. 445–448, URL : https://hal.archives-ouvertes.fr/hal-02050016.

[Ben+19b] Zied Ben Othmane, Cyril De Runz, Amine Aït Younes et Vincent Mercelot, « Effect of Imprecise Data Income-Flow Variability on Harvest Stability : A Quantile-Based Approach », *dans* : *International Conference on Database and Expert Systems Applications - DEXA 2019*, sous la dir. d'Hartmann S., Küng J., Chakravarthy S., Anderst-Kotsis G., Tjoa A. et Khalil I., t. 11706, Database and Expert Systems Applications. Lecture Notes in Computer Science, Linz, Austria : Springer, Cham, 2019, p. 248–257.

[Ben+19c] Zied Ben Othmane, Cyril De Runz, Amine Aït Younes et Vincent Mercelot, « Quantifier la variabilité de séries temporelles de données imprécises », *dans* : (2019).

[Ben+19d] Zied Ben Othmane, Cyril De Runz, Amine Aït Younes et Vincent Mercelot, « Quantify the Variability of Time Series of Imprecise Data », *dans* : *International Conference on Flexible Query Answering Systems - FQAS2019*, t. 11529, Flexible Query Answering Systems, Amantea, Italy, 2019, p. 203–214, DOI : 10.1007/978-3-030-27629-4_20, URL : https://hal.archives-ouvertes.fr/hal-02293136.

[Ber07a] Laure Berti-Equille, « Measuring and modelling data quality for quality-awareness in data mining », *dans* : *Quality measures in data mining*, Springer, 2007, p. 101–126.

[Ber07b] Laure Berti-Équille, « Data Quality Awareness : A Case Study for Cost Optimal Association Rule Mining », *dans* : *Knowl. Inf. Syst.* 11.*2* (fév. 2007), p. 191–215, ISSN : 0219-1377, DOI : 10.1007/s10115-006-0006-x, URL : https://doi.org/10.1007/s10115-006-0006-x.

[Ber15] Laure Berti-Equille, « Data veracity estimation with ensembling truth discovery methods », *dans* : (2015), p. 2628–2636.

[Ber99] Laure Berti-Équille, « Qualité des données multi-sources et recommandation multi-critère », *dans* : *Actes du congrès francophone INFormatique des ORganisations et Systèmes d'INformation Décisionnels (INFORSID'99)*, Toulon, France, juin 1999, p. 185–204, URL : https://hal.inria.fr/hal-01856327.

[BHL05] Frédéric Blanchard, Michel Herbin et Laurent Lucas, « A new pixel-oriented visualization technique through color image », *dans* : *Information Visualization* 4.*3* (2005), p. 257–265, DOI : 10.1057/palgrave.ivs.9500104, URL : https://doi.org/10.1057/palgrave.ivs.9500104.

[Bhu+14] Nishi Bhuvandas, PV Timbadiya, PL Patel et PD Porey, « Review of downscaling methods in climate change and their role in hydrological studies », *dans* : *Int J Environ Chem Ecol Geol Geophys Eng* 8.*10* (2014), p. 648–653.

[Bon+14] Georges-Pierre Bonneau, Hans-Christian Hege, Chris R. Johnson, Manuel M. Oliveira, Kristin Potter, Penny Rheingans et Thomas Schultz, « Overview and State-of-the-Art of Uncertainty Visualization », *dans* : *Scientific Visualization : Uncertainty, Multifield, Biomedical, and Scalable Visualization*, sous la dir. de Charles D. Hansen, Min Chen, Christopher R. Johnson, Arie E. Kaufman et Hans Hagen, London : Springer London, 2014, p. 3–27, ISBN : 978-1-4471-6497-5, DOI : 10.1007/978-1-4471-6497-5_1, URL : https://doi.org/10.1007/978-1-4471-6497-5_1.

[Bou+16] Lydia Boudjeloud-Assala, Philippe Pinheiro, Alexandre Blansché, Thomas Tamisier et Benoît Otjacques, « Interactive and iterative visual clustering », *dans* : *Information Visualization* 15.*3* (2016), p. 181–197, URL : https://doi.org/10.1177/1473871615571951.

[Bou93] Bernadette Bouchon-Meunier, *La logique floue*, QUE SAIS-JE ?, PUF, 1993.

[Bou95] Bernadette Bouchon-Meunier, *La logique floue et ses applications*, EddAddison-Wesley, 1995.

[BP03] R. Brown et B. Pham, « Visualisation of fuzzy decision support information : a case study », *dans* : *The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ '03.* T. 1, United States of America, mai 2003, 601–606 vol.1, doi : 10.1109/FUZZ.2003.1209432.

[BP95] Donald P. Ballou et Harold L. Pazer, « Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff », *dans* : *Info. Sys. Research* 6.*1* (mar. 1995), p. 51–72, issn : 1526-5536, doi : 10.1287/isre.6.1.51, url : http://dx.doi.org/10.1287/isre.6.1.51.

[BT99] Donald P. Ballou et Giri Kumar Tayi, « Enhancing Data Quality in Data Warehouse Environments », *dans* : *Commun. ACM* 42.*1* (jan. 1999), p. 73–78, issn : 0001-0782, doi : 10.1145/291469.291471, url : http://doi.acm.org/10.1145/291469.291471.

[Bur+19] Robin Burke, Himan Abdollahpouri, Edward C. Malthouse, K. P. Thai et Yongfeng Zhang, « Recommendation in multistakeholder environments », *dans* : *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019.* Sous la dir. de Toine Bogers, Alan Said, Peter Brusilovsky et Domonkos Tikk, ACM, 2019, p. 566–567, doi : 10.1145/3298689.3346973, url : https://doi.org/10.1145/3298689.3346973.

[Cac04] Serge Cacaly, *Dictionnaire de l'information*, Dunod, 2004.

[Cai19] Alan J Cain, « Visual thinking and simplicity of proof », *dans* : *Philosophical Transactions of the Royal Society A* 377.*2140* (2019), p. 20180032.

[Cap15] Cinzia Cappiello, « On the Role of Data Quality in Improving Web Information Value », *dans* : *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, Florence, Italy : ACM, 2015, p. 1433–1433, isbn : 978-1-4503-3473-0, url : http://dl.acm.org/citation.cfm?id=2740908.2778845.

[CC05] Brock Craft et Paul Cairns, « Beyond guidelines : what can we learn from the visual information seeking mantra ? », *dans* : *Ninth International Conference on Information Visualisation (IV'05)*, IEEE, 2005, p. 110–118.

[CD09] I. Couso et D. Dubois, « On the Variability of the Concept of Variance for Fuzzy Random Variables », *dans* : *IEEE Transactions on Fuzzy Systems* 17.*5* (oct. 2009), p. 1070–1080, doi : 10.1109/TFUZZ.2009.2021617.

[Che+09] Wei-Lung Chen, Tung-Hu Tsai, Chien-Cheng Huang, Jiann-Hwa Chen et Cheng-Deng Kuo, « Heart rate variability predicts short-term outcome for successfully resuscitated patients with out-of-hospital cardiac arrest », *dans* : *Resuscitation* 80.*10* (2009), p. 1114–1118.

[Coe+08] C. A. S. Coelho, C. A. T. Ferro, D. B. Stephenson et D. J. Steinskog, « Methods for Exploring Spatial and Temporal Variability of Extreme Events in Climate Data », *dans* : *Journal of Climate* 21.*10* (mai 2008), p. 2072–2092, ISSN : 0894-8755, 1520-0442, DOI : 10.1175/2007JCLI1781.1, URL : http://journals.ametsoc.org/doi/abs/10.1175/2007JCLI1781.1 (visité le 02/10/2018).

[Cou+07] Inés Couso, Didier Dubois, Susana Montes et Luciano Sánchez, « On various definitions of the variance of a fuzzy random variable », *dans* : *Proceedings of the 5th International Symposium on Imprecise Probabilities and Their Applications, Prague, Czech Republic*, 2007.

[CSV18] Cinzia Cappiello, Walter Samá et Monica Vitali, « Quality Awareness for a Successful Big Data Exploitation », *dans* : *Proceedings of the 22Nd International Database Engineering & Applications Symposium*, IDEAS 2018, New York, NY, USA : ACM, 2018, p. 37–44, ISBN : 978-1-4503-6527-7, DOI : 10.1145/3216122.3216124, URL : http://doi.acm.org/10.1145/3216122.3216124 (visité le 01/10/2018).

[DDD19] Rodolphe Devillers, Eric Desjardin et Cyril De Runz, « Imperfection of Geographic Information : Concepts and Terminologies », *dans* : *Geographic Data Imperfection 1*, John Wiley & Sons, Ltd, 2019, chap. 2, p. 11–24, ISBN : 9781119507284, DOI : 10.1002/9781119507284.ch2, eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119507284.ch2, URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119507284.ch2.

[Ded14] Adenekan Dedeke, « Building quality into information supply chains : robust information supply chains », *dans* : *Information Quality*, Routledge, 2014, p. 99–110.

[Dem68] Arthur P. Dempster, « A generalization of Bayesian inference », *dans* : *Journal of the Royal Statistical Society* 30.*2* (1968), p. 205–247.

[DG08] Jeffrey Dean et Sanjay Ghemawat, « MapReduce : simplified data processing on large clusters », *dans* : *Communications of the ACM* 51.*1* (2008), p. 107–113.

[DGH10] Zhiyong Dong, Qingyang Gu et Xu Han, « Ambiguity aversion and rational herd behaviour », *dans* : *Applied Financial Economics* 20.*4* (2010), p. 331–343.

[DHP03] Didier Dubois, Allel HadjAli et Henri Prade, « Fuzziness and Uncertainty in Temporal Reasoning », *dans* : *J.UCS Journal of Universal Computer Science* 9.*9* (2003), p. 1168–1194, URL : http://www.jucs.org/jucs_9_9/fuzziness_and_uncertainty_in.

[DOP00] Didier Dubois, Walenty Ostasiewicz et Henri Prade, « Fuzzy sets : history and basic notions », *dans* : *Fundamentals of fuzzy sets*, Springer, 2000, p. 21–124.

[DPS96] Didier Dubois, Henri Prade et Philippe Smets, « Representing partial ignorance », en, *dans* : *IEEE Transactions on Systems, Man, and Cybernetics - Part A Systems and Humans* 26.*3* (mai 1996), p. 361–377, ISSN : 10834427, DOI : 10.1109/3468.487961, URL : http://ieeexplore.ieee.org/document/487961/ (visité le 03/10/2018).

[Dun73] J. C. Dunn, « A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters », *dans* : *Journal of Cybernetics* 3.*3* (1973), p. 32–57, DOI : 10.1080/01969727308546046, eprint : https://doi.org/10.1080/01969727308546046, URL : https://doi.org/10.1080/01969727308546046.

[FBY05] Klaus Fog, Christian Budtz et Baris Yakaboylu, *Storytelling*, Springer, 2005.

[Fel15] David Felber, « Order statistics and variability in data streams », thèse de doct., UCLA, 2015.

[Fig15] Ana Figueiras, « Towards the understanding of interaction in information visualization », *dans* : *2015 19th International Conference on Information Visualisation*, IEEE, 2015, p. 140–147.

[Fio07a] Céline Fiot, « Extraction de séquences fréquentes : des données numériques aux valeurs manquantes », thèse de doct., Université Montpellier II-Sciences et Techniques du Languedoc, 2007.

[Fio07b] Céline Fiot, « Frequent Sequence Discovery : from Numerical Data to Missing Values », Theses, Université Montpellier II - Sciences et Techniques du Languedoc, sept. 2007, URL : https://tel.archives-ouvertes.fr/tel-00179506.

[Fis+12a] Mikkel Fishman, Frank J. Jacono, Soojin Park, Reza Jamasebi, Anurak Thungtong, Kenneth A. Loparo et Thomas E. Dick, « A method for analyzing temporal patterns of variability of a time series from Poincaré plots », *dans* : *Journal of Applied Physiology* 113.*2* (2012), PMID : 22556398, p. 297–306, doi : 10.1152/japplphysiol.01377.2010, eprint : https://doi.org/10.1152/japplphysiol.01377.2010, url : https://doi.org/10.1152/japplphysiol.01377.2010.

[Fis+12b] Mikkel Fishman, Frank J. Jacono, Soojin Park, Reza Jamasebi, Anurak Thungtong, Kenneth A. Loparo et Thomas E. Dick, « A method for analyzing temporal patterns of variability of a time series from Poincaré plots », *dans* : *Journal of Applied Physiology* 113.*2* (mai 2012), p. 297–306, issn : 8750-7587, doi : 10.1152/japplphysiol.01377.2010, url : https://www.physiology.org/doi/full/10.1152/japplphysiol.01377.2010 (visité le 13/09/2018).

[Fis05] Peter F. Fisher, « Models of uncertainty in spatial data », *dans* : *Geographical Information Systems. Principles, Techniques, Management and Applications*, sous la dir. de Paul A. Longley, Michael F. Goodchild, David J. Maguire et David W. Rhind, t. 1, Wiley, 2005, chap. 13, p. 191–205.

[Fol+96] James D Foley, Foley Dan Van, Andries Van Dam, Steven K Feiner, John F Hughes, J Hughes et Edward Angel, *Computer graphics : principles and practice*, t. 12110, Addison-Wesley Professional, 1996.

[Fou+17] Hacène Fouchal, Emilien Bourdy, Geoffrey Wilhelm et Marwane Ayaida, « A validation tool for cooperative intelligent transport systems », *dans* : *J. Comput. Science* 22 (2017), p. 283–288, doi : 10.1016/j.jocs.2017.05.026, url : https://doi.org/10.1016/j.jocs.2017.05.026.

[FPS96] Usama Fayyad, Gregory Piatetsky-Shapiro et Padhraic Smyth, « From data mining to knowledge discovery in databases », *dans* : *AI magazine* 17.*3* (1996), p. 37–37.

[FR99] Paul D Feigin et Sidney I Resnick, « Pitfalls of fitting autoregressive models for heavy-tailed time series », *dans* : *Extremes* 1.*4* (1999), p. 391–422.

[Fra+04] Piero Fraternali, Pier Luca Lanzi, Maristella Matera et Andrea Maurino, « Model-driven Web usage analysis for the evaluation of Web application quality. », *dans* : *J. Web Eng.* 3.*2* (2004), p. 124–152.

[FS69] Ivan P. Fellegi et Alan B. Sunter, « A Theory for Record Linkage », *dans* : *Journal of the American Statistical Association* 64.*328* (1969), p. 1183–1210, doi : 10.1080/01621459.1969.10501049, url : https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049.

[FT99] B. J. Fogg et Hsiang Tseng, « The Elements of Computer Credibility », *dans* : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, Pittsburgh, Pennsylvania, USA : Association for Computing Machinery, 1999, p. 80–87, isbn : 0201485591, doi : 10.1145/302979.303001, url : https://doi.org/10.1145/302979.303001.

[GBS08] Saurabh Ganeriwal, Laura K. Balzano et Mani B. Srivastava, « Reputation-based Framework for High Integrity Sensor Networks », *dans* : *ACM Trans. Sen. Netw.* 4.*3* (juin 2008), 15 :1–15 :37, issn : 1550-4859, doi : 10.1145/1362542.1362546, url : http://doi.acm.org/10.1145/1362542.1362546.

[GH07] Fabrice Guillet et Howard J. Hamilton, *Quality Measures in Data Mining*, Springer, jan. 2007, isbn : 978-3-540-44918-8, url : https://www.springer.com/gp/book/9783540449119.

[GH11] Manish Gupta et Jiawei Han, « Heterogeneous network-based trust analysis : a survey », en, *dans* : *ACM SIGKDD Explorations Newsletter* 13.*1* (août 2011), p. 54, issn : 19310145, doi : 10.1145/2031331.2031341, url : http://dl.acm.org/citation.cfm?doid=2031331.2031341 (visité le 01/10/2018).

[Gro05] Robert Grossman, *KDD '05 : Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 618050, Chicago, Illinois, USA : ACM, 2005, isbn : 1-59593-135-X.

[GRR14] Venkat N. Gudivada, Dhana Rao et Vijay V. Raghavan, « NoSQL Systems for Big Data Management », *dans* : *Proceedings of the 2014 IEEE World Congress on Services*, SERVICES '14, USA : IEEE Computer Society, 2014, p. 190–197, isbn : 9781479950690, doi : 10.1109/SERVICES.2014.42, url : https://doi.org/10.1109/SERVICES.2014.42.

[Gua+09] Sean L Guarino, Jonathan D Pfautz, Zach Cox et Emilie Roth, « Modeling human reasoning about meta-information », *dans* : *International Journal of Approximate Reasoning* 50.*3* (2009), p. 437–449.

[He+07] Li-Ping He, Hong-Zhong Huang, Li Du, Xu-Dong Zhang et Qiang Miao, « Human-computer interaction : Pearson prentice hall », *dans* : (22 juil. 2007), (visité le 14/08/2018).

[Hei+97] Martin Theus Heike, Heike Hofmann, Bernd Siegl et Antony Unwin, « Manet extensions to interactive statistical graphics for missing values », *dans* : *In New Techniques and Technologies for Statistics II*, 1997.

[Hey04] Jonathan Hey, « The data, information, knowledge, wisdom chain : the metaphorical link », *dans* : *Intergovernmental Oceanographic Commission* 26 (2004), p. 1–18.

[HL12] Johannes Held et Richard Lenz, « Towards Measuring Test Data Quality », *dans* : *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, Berlin, Germany : ACM, 2012, p. 233–238, isbn : 978-1-4503-1143-4, doi : 10.1145/2320765.2320830, url : http://doi.acm.org/10.1145/2320765.2320830.

[HPK11] Jiawei Han, Jian Pei et Micheline Kamber, *Data mining : concepts and techniques*, Elsevier, 2011.

[HT05] Heike Hofmann et Martin Theus, *Interactive graphics for visualizing conditional distributions*, 2005.

[Hül14] Eyke Hüllermeier, « Learning from imprecise and fuzzy observations : Data disambiguation through generalized loss minimization », *dans* : *Int. J. Approx. Reasoning* 55.7 (2014), p. 1519–1534.

[ISO11] ISO, *Qualité des données – Partie 1, Aperçu*, FR, déc. 2011, url : https://www.iso.org/fr/standard/50798.html.

[J96] Rothenberg J., « Metadata to support data quality and longevity », *dans* : (1996).

[JV97] Matthias Jarke et Yannis Vassiliou, « Data Warehouse Quality : A Review of the DWQ Project. », *dans* : *International Conference on Information Quality – IQ'97*, 1997, p. 299–313.

[Kan17] Kyo Chul Kang, « Chapter 2 Variability Modeling », *dans* : 2017.

[Kei+08a] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer et Guy Melançon, « Visual Analytics : Definition, Process, and Challenges », *dans* : *Information Visualization : Human-Centered Issues and Perspectives*, sous la dir. d'Andreas Kerren, John T. Stasko, Jean-Daniel Fekete et Chris North, Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 154–175, isbn : 978-3-540-70956-5, doi : 10.1007/978-3-540-70956-5_7, url : https://doi.org/10.1007/978-3-540-70956-5_7.

[Kei+08b] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer et Guy Melançon, « Visual analytics : Definition, process, and challenges », *dans* : *Information visualization*, Springer, 2008, p. 154–175.

[Kei00] Daniel A. Keim, « Designing pixel-oriented visualization techniques : theory and applications », *dans* : *IEEE Transactions on Visualization and Computer Graphics* 6.*1* (jan. 2000), p. 59–78, DOI : 10.1109/2945.841121.

[KM82] Daniel G Krige et Eduardo J Magri, « Studies of the effects of outliers and data transformation on variogram estimates for a base metal and a gold ore body », *dans* : *Journal of the International Association for Mathematical Geology* 14.*6* (1982), p. 557–564.

[KM87] Rudolf Kruse et Klaus Dieter Meyer, *Statistics with vague data*, t. 6, Springer Science & Business Media, 1987.

[KN02] Ralf Körner et Wolfgang Näther, « On the variance of random fuzzy variables », *dans* : *Statistical modeling, analysis and management of fuzzy data*, Springer, 2002, p. 25–42.

[Koo+06] Sang Ok Koo, Hyuk Don Kwon, Chang Geol Yoon, Won Seok Seo et Soon Ki Jung, « Visualization for a multi-sensor data analysis », *dans* : *International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)*, IEEE, 2006, p. 57–63.

[LBT17] Ibrahim Louhi, Lydia Boudjeloud-Assala et Thomas Tamisier, « Subspace Clustering et Visualisation des Flux de Données », *dans* : *17ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2017, 24-27 Janvier 2017, Grenoble, France*, sous la dir. de Fabien L. Gandon et Gilles Bisson, t. E-33, RNTI, Éditions RNTI, 2017, p. 327–332, URL : http://editions-rnti.fr/?inprocid=1002293.

[Leb+19] Brice Leblanc, Emilien Bourdy, Hacène Fouchal, Cyril de Runz et Secil Ercan, « Unsupervised Driving Profile Detection Using Cooperative Vehicles' Data », *dans* : *Communication Technologies for Vehicles - 14th International Workshop, Nets4Cars/Nets4Trains/Nets4Aircraft 2019, Colmar, France, May 16-17, 2019, Proceedings*, sous la dir. de Benoît Hilt, Marion Berbineau, Alexey V. Vinel, Magnus Jonsson et Alain Pirovano, t. 11461, Lecture Notes in Computer Science, Springer, 2019, p. 27–37, DOI : 10.1007/978-3-030-25529-9_3, URL : https://doi.org/10.1007/978-3-030-25529-9%5C_3.

[LR14a] Tatiana Lukoianova et Victoria L. Rubin, « Veracity Roadmap : Is Big Data Objective, Truthful and Credible ? », en, *dans* : *Advances in Classification Research Online* 24.*1* (jan. 2014), p. 4–15, ISSN : 2324-9773, DOI : 10.7152/acro.v24i1.14671, URL : http://journals.lib.washington.edu/index.php/acro/article/view/14671 (visité le 01/10/2018).

[LR14b] Tatiana Lukoianova et Victoria L. Rubin, « Veracity Roadmap : Is Big Data Objective, Truthful and Credible ? », en, *dans* : *Advances in Classification Research Online* 24.*1* (jan. 2014), p. 4–15, issn : 2324-9773, (visité le 01/10/2018).

[Lub99] MA Lubiano, « Variation measures for imprecise random elements », thèse de doct., Ph. D. Thesis, Universidad de Oviedo, Spain, 1999 (in Spanish), 1999.

[Mac+12] Alan M MacEachren, Robert E Roth, James O'Brien, Bonan Li, Derek Swingley et Mark Gahegan, « Visual semiotics & uncertainty visualization : An empirical study », *dans* : *IEEE Transactions on Visualization and Computer Graphics* 18.*12* (2012), p. 2496–2505.

[Mac67] James MacQueen, « Some methods for classification and analysis of multivariate observations », *dans* : *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, t. 1, 14, Oakland, CA, USA, 1967, p. 281–297.

[Mac92] Alan M. MacEachren, « Visualizing Uncertain Information », *dans* : *Cartographic Perspectives 13* (juin 1992), p. 10–19, doi : 10.14714/CP13.1000, url : https://cartographicperspectives.org/index.php/journal/article/view/cp13-maceachren.

[Mar19] Arnaud Martin, « Conflict Management in Information Fusion with Belief Functions », *dans* : *Information Quality in Information Fusion and Decision Making*, sous la dir. d'Éloi Bossé et Galina L. Rogova, Springer, 2019, p. 79–97, doi : 10.1007/978-3-030-03643-0_4, url : https://doi.org/10.1007/978-3-030-03643-0%5C_4.

[MGD05] Denis McQuail, Peter Golding et Els De Bens, *Communication theory and research*, Sage, 2005.

[Mik+95] Thomas Mikosch, Tamar Gadrich, Claudia Kluppelberg et Robert J. Adler, « Parameter Estimation for ARMA Models with Infinite Variance Innovations », *dans* : *The Annals of Statistics* 23.*1* (1995), p. 305–326, issn : 00905364, url : http://www.jstor.org/stable/2242413.

[MWC99] Subramanian Muthu, Larry Whitman et S. Hossein Cheraghi, « Business process reengineering : a consolidated methodology », *dans* : *Proceedings of the 4 th Annual International Conference on Industrial Engineering Theory, Applications, and Practice*, San Antonio, Texas, USA, 1999.

[NLF18] Tawfiq Nebbou, Mohamed Lehsaini et Hacène Fouchal, « Advanced Measurement of Road Traffic Information in City Environments », *dans* : *14th International Wireless Communications & Mobile Computing Conference, IWCMC 2018, Limassol, Cyprus, June 25-29, 2018*, IEEE, 2018, p. 1255–1260, DOI : 10.1109/IWCMC.2018.8450322, URL : https://doi.org/10.1109/IWCMC.2018.8450322.

[Nor02] Donald A. Norman, *The Design of Everyday Things*, Reprint Paperback, New York : Basic Books, 2002, ISBN : 0-465-06710-7.

[Oca85] M. Ocagne, *Coordonnées parallèles et axiales : méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles*, Gauthier-Villars, 1885.

[OOK14] E. Olshannikova, A. Ometov et Y. Koucheryavy, « Towards Big Data Visualization for Augmented Reality », *dans* : *2014 IEEE 16th Conference on Business Informatics*, t. 2, juil. 2014, p. 33–37, DOI : 10.1109/CBI.2014.42.

[PA03] Missier P et Batini C. A, « A multidimensional model for information quality », *dans* : (2003).

[Pai+11] Ian Painter, Julie Eaton, Don Olson, Debra Revere et Lober Bill, « Generation of Prediction Intervals to Assess Data Quality in the Distribute System Using Quantile Regression », *dans* : *JSM proceedings, Section on Statistics in Defense and National Security*, 2011.

[Pao+06] Jean-Noël Paoli, Olivier Strauss, Bruno Tisseyre et Philippe Lagacherie, « Utilisation d'un variogramme flou dans une méthode d'agrégation sémantique », *dans* : *LFA Logique Floue et ses Applications*, Toulouse, France : Cépaduès Editions, 2006, p. 303–306.

[PBM04] Malay K Pakhira, Sanghamitra Bandyopadhyay et Ujjwal Maulik, « Validity index for crisp and fuzzy clusters », *dans* : *Pattern recognition* 37.*3* (2004), p. 487–501.

[Pea06] Ronald K. Pearson, « The Problem of Disguised Missing Data », *dans* : *SIGKDD Explor. Newsl.* 8.*1* (juin 2006), p. 83–92, ISSN : 1931-0145, DOI : 10.1145/1147234.1147247, URL : http://doi.acm.org/10.1145/1147234.1147247.

[Per06] Veronika Peralta, « Data Quality Evaluation in Data Integration Systems », Theses, Université de Versailles-Saint Quentin en Yvelines ; Université de la République d'Uruguay, nov. 2006, URL : https://tel.archives-ouvertes.fr/tel-00325139.

[Pok13] Jaroslav Pokorny, « NoSQL databases : a step to database scalability in web environment », *dans* : *International Journal of Web Information Systems* 9.*1* (2013), p. 69–82.

[PWL97] Alex T Pang, Craig M Wittenbrink et Suresh K Lodha, « Approaches to uncertainty visualization », *dans* : *The Visual Computer* 13.*8* (1997), p. 370–390.

[Qui69] Willard van O Quine, *Set theory and its logic*, t. 9, Harvard University Press, 1969.

[Rag97] Dave Raggett, *HTML 3.2 Reference Specification*, World Wide Web Consortium, Recommendation REC-html32, jan. 1997.

[Red97] Thomas C. Redman, *Data Quality for the Information Age*, 1st, Norwood, MA, USA : Artech House, Inc., 1997, isbn : 0890068836.

[RHJ04] Sarvapali D. Ramchurn, Dong Huynh et Nicholas R. Jennings, « Trust in Multi-agent Systems », *dans* : *Knowl. Eng. Rev.* 19.*1* (mar. 2004), p. 1–25, issn : 0269-8889, doi : 10.1017/S0269888904000116, url : http://dx.doi.org/10.1017/S0269888904000116.

[Rie10] Soo Young Rieh, « Credibility and cognitive authority of information », *dans* : (2010).

[Rin+13] Alexander Rind, Taowei David Wang, Wolfgang Aigner, Silvia Miksch, Krist Wongsuphasawat, Catherine Plaisant, Ben Shneiderman et al., « Interactive information visualization to explore and query electronic health records », *dans* : *Foundations and Trends® in Human–Computer Interaction* 5.*3* (2013), p. 207–298.

[RL06] Victoria L Rubin et Elizabeth D Liddy, « Assessing Credibility of Weblogs. », *dans* : *AAAI Spring Symposium : Computational Approaches to Analyzing Weblogs*, 2006, p. 187–190.

[Rou87] Peter J. Rousseeuw, « Silhouettes : A graphical aid to the interpretation and validation of cluster analysis », *dans* : *Journal of Computational and Applied Mathematics* 20 (1987), p. 53–65, issn : 0377-0427, doi : https://doi.org/10.1016/0377-0427(87)90125-7, url : http://www.sciencedirect.com/science/article/pii/0377042787901257.

[RR40] P. F. Russell et T. Ramachandra Rao, « On Habitat and Association of Species of Anopheline Larvae in South-eastern Madras. », *dans* : *Journal of the Malaria Institute of India* 3.*1* (1940), 153–178 pp.

[Run+08] Cyril de Runz, Frédéric Blanchard, Eric Desjardin et Michel Herbin, « Exploration d'un ensemble de quantités floues », *dans* : *Logique floue et ses applications - LFA'08*, 2008.

[Run+10a] Cyril de Runz, Frédéric Blanchard, Philippe Vautrot et Éric Desjardin, « Visualisation de données spatiotemporelles imprécises : application en archéologie », *dans* : *Atelier Fouille de Données Complexes, conférence Extraction et Gestion des Connaissances 2010*, Hammamet, Tunisie, 2010, p. 69–80.

[Run+10b] Cyril de Runz, Éric Desjardin, Frederic Piantoni et Michel Herbin, « Anteriority index for managing fuzzy dates in archaeological GIS », *dans* : *Soft Computing - A Fusion of Foundations, Methodologies and Applications* 14.*4* (2010), p. 339–344.

[Run08] Cyril de Runz, « Imperfection, temps et espace : modélisation, analyse et visualisation dans un SIG archéologique », Thèse de doctorat, Université de Reims Champagne-Ardenne, France, 2008, URL : http://crestic.univ-reims.fr/publication/1615/pdf.

[SBM18] Özge Sürer, Robin Burke et Edward C. Malthouse, « Multistakeholder Recommendation with Provider Constraints », *dans* : *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, Vancouver, British Columbia, Canada : ACM, 2018, p. 54–62, ISBN : 978-1-4503-5901-6, DOI : 10.1145/3240323.3240350, URL : http://doi.acm.org/10.1145/3240323.3240350.

[Sco15] David W Scott, *Multivariate density estimation : theory, practice, and visualization*, John Wiley et Sons, 2015.

[SE04] Pernici B Scannapieco M et Pierce E., « Advances in Management Information Systems », *dans* : (2004).

[SG02] Joseph L Schafer et John W Graham, « Missing data : our view of the state of the art. », *dans* : *Psychological methods* 7.*2* (2002), p. 147.

[SG09] John Sharko et Georges Grinstein, « Visualizing Fuzzy Clusters Using RadViz », *dans* : *Proceedings of the 2009 13th International Conference Information Visualisation*, IV '09, Washington, DC, USA : IEEE Computer Society, 2009, p. 307–316, ISBN : 978-0-7695-3733-7, DOI : 10.1109/IV.2009.74, URL : https://doi.org/10.1109/IV.2009.74.

[Sha76] Glenn Shafer, *A Mathematical Theory of Evidence*, Princeton, N.J., USA : Princeton University Press, 1976.

[Sid+12] Fatimah Sidi, Payam Hassany Shariat Panah, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim et Aida Mustapha, « Data quality : A survey of data quality dimensions », *dans* : *2012 International Conference on Information Retrieval & Knowledge Management, Kuala Lumpur, Malaysia, March 13-15, 2012*, sous la dir. de Ramlan Mahmod, Rusli Abdullah, Lili Nurliyana Abdullah, Tengku Mohd Tengku Sembok, Alan F. Smeaton, Fabio Crestani, Shyamala Doraisamy, Rabiah Abdul Kadir et Mahamod Ismail, IEEE, 2012, p. 300–304, doi : 10.1109/InfRKM.2012.6204995, url : https://doi.org/10.1109/InfRKM.2012.6204995.

[SLW97] Diane M. Strong, Yang W. Lee et Richard Y. Wang, « Data Quality in Context », *dans* : *Commun. ACM* 40.5 (mai 1997), p. 103–110, issn : 0001-0782, doi : 10.1145/253769.253804, url : http://doi.acm.org/10.1145/253769.253804.

[Sme91] Philippe Smets, « Varieties of ignorance and the need for well-founded theories. », *dans* : *Information Sciences* 57.*58* (1991), p. 135–144.

[Sme98] Philippe Smets, « Probability, Possibility, Belief : Which and Where ? », *dans* : *Handbook of Defeasible Reasoning and Uncertainty Management Systems.* Sous la dir. de D. Gabbay et Ph. Smets, t. 1, Doordrecht : Kluwer, 1998, p. 1–24.

[ST03] Scannapieco M Santis LD et Catarci T, « Trusting data quality in cooperative information systems », *dans* : (2003).

[ST17] Jonas Sjöbergh et Yuzuru Tanaka, « Visualizing Missing Values », *dans* : *21st International Conference Information Visualisation, IV 2017, London, United Kingdom, July 11-14, 2017*, IEEE Computer Society, 2017, p. 242–249, doi : 10.1109/iV.2017.12, url : https://doi.org/10.1109/iV.2017.12.

[SVS05] A. Simitsis, P. Vassiliadis et T. Sellis, « Optimizing ETL processes in data warehouses », *dans* : (avr. 2005), p. 564–575, doi : 10.1109/ICDE.2005.103.

[SWZ00] G. Shankaranarayanan, Richard Y. Wang et Mostapha Ziad, « Ziad : IP-MAP : Representing the Manufacture of an Information Product », *dans* : *In Proceedings of the International Conference on Information Quality (IQ*, 2000, p. 1–16.

[TAF12] Matthias Templ, Andreas Alfons et Peter Filzmoser, « Exploring Incomplete Data Using Visualization Techniques », *dans* : *Adv. Data Anal. Classif.* 6.*1* (avr. 2012), p. 29–47, issn : 1862-5347, doi : 10.1007/s11634-011-0102-y, url : https://doi.org/10.1007/s11634-011-0102-y.

[TC06] James J. Thomas et Kristin A. Cook, « A Visual Analytics Agenda », *dans* : *IEEE Comput. Graph. Appl.* 26.*1* (jan. 2006), p. 10–13, ISSN : 0272-1716, DOI : 10.1109/MCG.2006.5, URL : https://doi.org/10.1109/MCG.2006.5.

[Twe97] Lisa Tweedie, « Characterizing interactive externalizations », *dans* : *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, ACM, 1997, p. 375–382.

[Wan98] Richard Y. Wang, « A Product Perspective on Total Data Quality Management », *dans* : *Commun. ACM* 41.*2* (fév. 1998), p. 58–65, ISSN : 0001-0782, DOI : 10.1145/269012.269022, URL : http://doi.acm.org/10.1145/269012.269022.

[WB04] Jon Wiles et Joseph Bondi, *Supervision : A guide to practice*, Prentice Hall, 2004.

[WB97] Pak Chung Wong et R Daniel Bergeron, « Multivariate visualization using metric scaling », *dans* : *Proceedings. Visualization'97 (Cat. No. 97CB36155)*, IEEE, 1997, p. 111–118.

[WKP14] Colin Ware, John GW Kelley et David Pilar, « Improving the display of wind patterns and ocean currents », *dans* : *Bulletin of the American Meteorological Society* 95.*10* (2014), p. 1573–1581.

[WR12] Wouter Weerkamp et Maarten de Rijke, « Credibility-inspired ranking for blog post retrieval », *dans* : *Information Retrieval* 15.*3* (juin 2012), p. 243–277, ISSN : 1573-7659, DOI : 10.1007/s10791-011-9182-8, URL : https://doi.org/10.1007/s10791-011-9182-8.

[WS96] Richard Y. Wang et Diane M. Strong, « Beyond Accuracy : What Data Quality Means to Data Consumers », *dans* : *Journal of Management Information Systems* 12.*4* (1996), p. 5–33, DOI : 10.1080/07421222.1996.11518099, URL : https://doi.org/10.1080/07421222.1996.11518099.

[Yag91] Ronald R Yager, « Connectives and quantifiers in fuzzy sets », *dans* : *Fuzzy sets and systems* 40.*1* (1991), p. 39–75.

[YC07] Jae Keun Yoo et R Dennis Cook, « Optimal sufficient dimension reduction for the conditional mean in multivariate regression », *dans* : *Biometrika* 94.*1* (2007), p. 231–242.

[Yur18] Bilal Yurdakul, « Statistical Properties of Population Stability Index », *dans* : (2018).

[Zad65] Lotfi A. Zadeh, « Fuzzy Sets », *dans* : *Information Control* 8 (1965), p. 338–353.

[Zad78] Lotfi A. Zadeh, « Fuzzy sets as a basis for a theory of possibility », *dans* : *Fuzzy Sets and Systems* 1 (1978), p. 3–28.

[ZC07] Torre Zuk et Sheelagh Carpendale, « Visualization of Uncertainty and Reasoning », *dans* : *7th International Symposium on Smart Graphics (June 25-27, 2007, Kyoto, Japan), 4569. (Berlin, Heidelberg), (Andreas Butz and Brian Fisher and Antonio Krüger and Patrick Olivier and Shigeru Owada, Ed.) Springer-Verlag* (2007), p. 164–177.

[ZHC07] Yue Zhang, Jason I Hong et Lorrie F Cranor, « Cantina : a content-based approach to detecting phishing web sites », *dans* : *Proceedings of the 16th international conference on World Wide Web*, ACM, 2007, p. 639–648.

[Zhi+15] Shi Zhi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji et Jiawei Han, « Modeling Truth Existence in Truth Discovery », *dans* : *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, Sydney, NSW, Australia : ACM, 2015, p. 1543–1552, isbn : 978-1-4503-3664-2, doi : 10.1145/2783258.2783339, url : http://doi.acm.org/10.1145/2783258.2783339.

[ZMP18] Kuang Zhou, Arnaud Martin et Quan Pan, « A belief combination rule for a large number of sources », *dans* : *Journal of Advances in Information Fusion* 13.*2* (déc. 2018), url : https://hal.archives-ouvertes.fr/hal-01883239.

[Zuk08] Torre D. Zuk, « Visualizing Uncertainty », *dans* : *PhD thesis, Department of Computer Science, University of Calgary, Calgary, Alberta, Canada* (2008).

Analyse et visualisation pour l'étude de la qualité des séries temporelles de données imparfaites

Dans ce travail de thèse, nous nous intéressons à la qualité des informations récoltées par des capteurs sur le web. Ces données forment des séries de données temporelles qui sont incomplètes et imprécises, et sont sur des échelles quantitatives peu comparables. Dans ce contexte, nous nous intéressons plus particulièrement à la variabilité et la stabilité de ces séries temporelles. Nous proposons deux approches pour les quantifier. La première se base sur une représentation à l'aide des quantiles, la seconde est une approche floue. A l'aide de ces indicateurs, nous proposons un outil de visualisation interactive dédié à l'analyse de la qualité des récoltes effectuées par les capteurs. Ce travail s'inscrit dans une collaboration CIFRE avec la société Kantar.

Mots-clés : Qualité, Variabilité, Données imparfaites, Fouille de données, Visualisation, Séries temporelles

Analysis and visualization for studying the quality of imperfect data time series

This thesis focuses on the quality of the information collected by sensors on the web. These data form time series that are incomplete, imprecise, and are on quantitative scales that are not very comparable. In this context, we are particularly interested in the variability and stability of these time series. We propose two approaches to quantify them. The first is based on a representation using quantiles, the second is a fuzzy approach. Using these indicators, we propose an interactive visualization tool dedicated to the analysis of the quality of the harvest carried out by the sensors. This work is part of a CIFRE collaboration with Kantar.

Keywords : Quality, Variability, Imperfect data, Data mining, Visualization, Time series

Discipline : INFORMATIQUE

Spécialité : Intelligence Artificielle

Université de Reims Champagne-Ardenne

CRESTIC - EA 3804

UFR Sciences Exactes et Naturelles,
- Moulin de la Housse - BP 1039 -
51687 Reims CEDEX 2