

**POLYTECHNIQUE MONTRÉAL**

affiliée à l'Université de Montréal

**Développement d'un outil de segmentation des comportements d'achat des clients en se basant sur leurs données morphologiques**

**SAFA EL AYEB**

Département de mathématique et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Novembre 2019

# **POLYTECHNIQUE MONTRÉAL**

affiliée à l'Université de Montréal

Ce mémoire intitulé :

## **Développement d'un outil de segmentation des comportements d'achat des clients en se basant sur leurs données morphologiques**

présenté par **Safa EL AYEB**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

**Martin TRÉPANIÉ**, président

**Bruno AGARD**, membre et directeur de recherche

**Christophe DANJOU**, membre

## DÉDICACE

*A mes très chers parents*

*Auxquels je dois ma réussite et toute mon existence*

*C'est grâce à votre soutien que je suis où je suis*

*Même si vous étiez loin de moi pendant mes deux années, vous êtes toujours mon*

*Inspiration et ma guidance*

*A mes chers petits frères, pour leur amour, et leur attention,*

*A Achref, je te dois mon courage et ma détermination,*

*A M. Bruno Agard, pour ta guidance et ton soutien,*

*A tous mes amis au labo, je vous remercie du fond du cœur, et je n'oublierai jamais les*

*Bons moments que nous avons vécu ensemble,*

*Je vous dédie mon travail, que vous y trouviez l'expression de ma profonde gratitude et*

*Reconnaissance*

*Safa*

## REMERCIEMENTS

J'aimerais profiter de ces quelques lignes pour remercier les personnes qui m'ont aidé et soutenu durant mes deux années de maîtrise. J'aimerais commencer par remercier mon encadrant M. Bruno Agard. Durant mes travaux de maîtrise, vous étiez toujours source de soutien et d'inspiration. Je vous remercie de vos conseils, guidance, et votre compréhension et soutien durant les moments les plus difficiles. Votre présence a été indispensable pour ce projet. Vous avez marqué ce travail, aussi bien que ma personne.

Je remercie également mes amis du laboratoire en Intelligence de données. Vous n'étiez pas de simples collègues, mais bien des amis. Grâce à vous l'environnement de travail est toujours plaisant. Vous n'hésitez pas donner de l'aide lorsque vous le pouvez. J'ai beaucoup appris de vous.

Je remercie également toute l'équipe de Logistik Inc., de votre confiance, votre accueil, vos explications et votre enthousiasme. Ce fut un vrai plaisir de travailler avec votre entreprise.

Je remercie également toute ma famille qui m'a soutenu, qui m'inspirait pour rester forte même loin d'eux, et toujours de me pousser pour donner le meilleur de moi-même.

Je remercie aussi les membres du jury, qui m'ont fait l'honneur de juger ce travail. Je remercie tous mes enseignants à l'École polytechnique de Montréal, et les membres administratifs, votre aide est très valorisante pour nous.

## RÉSUMÉ

L'analyse des marchés commerciaux est actuellement un processus aussi bien scientifique qu'industriel. Il consiste à recueillir et explorer des informations reliées aux clients en vue de mieux comprendre leurs comportements, habitudes et intérêts. Cette analyse est fortement utilisée par les entreprises afin de les guider dans leurs décisions opérationnelles et stratégiques. Ce mémoire présente un des outils d'analyse des marchés les plus utilisés dans la littérature, qui est la segmentation. Cette approche vise à diviser un ensemble d'individus hétérogènes, en groupes plus homogènes, en se basant sur des critères prédéterminés. La segmentation des marchés est employée dans plusieurs domaines, et vise à diviser un ensemble de clients en plus petits groupes ayant des comportements similaires. Elle peut être employée pour l'amélioration de l'impact et des revenus des produits et services existants ou pour préparer l'introduction de nouveaux produits sur les marchés. Dans notre cas, la segmentation des clients est basée sur l'évolution de tailles des vêtements qu'ils commandent, ou ce qu'on appellera leurs données morphologiques. Les clients font leurs achats sur une plateforme d'achat en ligne fournie par notre partenaire industriel. Notre partenaire est une entreprise qui se charge de la sous-traitance des programmes d'uniformes pour ses clients. Dans ce travail, nous avons utilisé des méthodes de « data mining » pour analyser l'historique des données d'achat d'un des clients de notre partenaire pour une durée d'étude déterminée. Les analyses sont par la suite concentrées sur le type de vêtements le plus commandé, soit les chemises. Des séries chronologiques des tailles commandées pour chaque client sont construites. Ces séries chronologiques permettent d'étudier l'évolution des tailles dans le temps. La segmentation est par la suite appliquée sur ces séries chronologiques afin d'obtenir des groupes ayant des comportements similaires. La similarité entre les clients est basée sur une métrique appelée le Dynamic Time Warping. Cette distance a été choisie parce qu'elle est la plus adaptée pour comparer les séries temporelles en se basant sur leurs formes et en ne tenant pas compte du décalage sur l'axe temporel. Plusieurs tests basés sur différentes méthodes et algorithmes ont eu lieu. Aussi plusieurs structures et variantes des séries temporelles ont été testées. Les séries temporelles retenues sont des séries exprimées en fonction de la variation des mensurations dans le temps. Elles ont été normalisées avec la norme  $z$ . Deux segmentations avec des nombres de groupes égaux à six et à dix ont été réalisées. Finalement, une évaluation et une analyse des résultats de la segmentation ont été effectuées pour valider les résultats recueillis. En premier lieu, l'évaluation s'est basée sur la visualisation des groupes obtenus et sur le critère de silhouette. Ce

critère permet de mesurer la qualité d'une segmentation donnée, en se basant sur l'homogénéité intragroupe et l'hétérogénéité inter-groupe. Nous avons trouvé que la segmentation en dix groupes a permis d'avoir de meilleurs groupes. Dans un second temps, pour chacun des cas, l'analyse des groupes nous a permis de comprendre la structure et les caractéristiques de ces derniers. Nous avons trouvé que la majorité des clients ont des courbes d'évolution croissante. Nous avons aussi constaté que les groupes ayant des courbes d'évolution stables, ont des pourcentages de retour relativement élevés par rapport à leurs pourcentages d'achat. La répartition des hommes et des femmes dans les groupes est semblable à celle du groupe de clients initial. Cependant, les groupes ayant un pourcentage de femme plus important que les autres ont des moyennes de tailles plus petites.

## ABSTRACT

The analysis of commercial markets is currently a scientific as well as an industrial process. It consists of collecting and exploring information related to clients in order to better understand their behaviours, habits and interests. This analysis is widely used by companies to guide them in their operational and strategic decisions. This paper presents one of the most widely used market analysis tools in the literature, namely segmentation. This approach aims to divide a set of heterogeneous individuals into more homogeneous groups, based on predetermined criteria. Market segmentation is used in several areas and aims to divide a set of customers into smaller groups with similar behaviors. It can be used to improve the impact and revenues of existing products and services or to prepare for the introduction of new products to markets. In our case, customer segmentation is based on the size evolution of the clothes they order, or what we will call their morphological data. Customers make their purchases on an online shopping platform provided by our partner. Our partner is a company that subcontracts uniform programs for its customers. In this work, we used data mining methods to analyze the purchase data history of one of our partner's customers for a specific study period. The analyses are then focused on the most ordered type of clothing, namely shirts. Time series of the sizes ordered for each customer are built. These time series make it possible to study the evolution of sizes over time. Segmentation is then applied to these time series to obtain groups with similar behaviors. The similarity between customers is based on a metric called Dynamic Time Warping. This distance was chosen because it is the most suitable for comparing time series based on their shapes and ignoring the time axis offset. Several tests based on different methods and algorithms were carried out. Also, several structures and variants of the time series were tested. The time series used are series expressed as a function of the variation in measurements over time, and which were normalized with the z-norm. Two segmentations with numbers of groups equal to six and ten were performed. Finally, an evaluation and analysis of the segmentation results was carried out to validate the results collected. First, the evaluation was based on the visualization of the groups obtained and on the silhouette criterion. This criterion makes it possible to measure the quality of a given segmentation, based on intra-group homogeneity and inter-group heterogeneity. We found that the segmentation into ten groups allowed us to have better groups. This segmentation improved the silhouette index obtained with six groups. In a second

step, for each case, the analysis of the groups allowed us to understand their structure and characteristics. We found that most clients have increasing trend curves. We also found that groups with stable evolution curves have relatively high return rates compared to their purchase rates. The distribution of men and women in the groups is similar to that of the original client group. However, groups with a higher percentage of women were found to have smaller average sizes.



## TABLE DES MATIÈRES

DÉDICACE.....	III
REMERCIEMENTS .....	IV
RÉSUMÉ.....	V
ABSTRACT .....	VII
TABLE DES MATIÈRES .....	IX
LISTE DES TABLEAUX.....	XI
LISTE DES FIGURES .....	XII
LISTE DES SIGLES ET ABRÉVIATIONS .....	XIV
CHAPITRE 1 INTRODUCTION.....	1
CHAPITRE 2 REVUE DE LA LITTÉRATURE.....	4
2.1 L'industrie du vêtement et « les tailles ».....	4
2.2 Séries temporelles intermittentes .....	8
2.3 La segmentation .....	13
2.3.1 Algorithmes de segmentation.....	14
2.3.2 Métriques de distances .....	18
2.3.3 Évaluation de la segmentation: .....	20
2.4 Conclusion.....	24
CHAPITRE 3 PROBLÉMATIQUE ET MÉTHODOLOGIE .....	25
3.1 Contexte de l'étude.....	25
3.2 Problématique.....	26
3.3 Objectifs .....	26
3.3.1 Objectif général .....	26
3.3.2 Objectifs spécifiques .....	26

3.4	Méthodologie .....	26
3.4.1	Nettoyage de la base de données.....	27
3.4.2	Transformation en série temporelle.....	28
3.4.3	Segmentation des séries temporelles.....	29
3.4.4	Validation de la segmentation .....	29
3.4.5	Interprétation des résultats .....	30
3.5	Conclusion.....	30
CHAPITRE 4	CAS D'ÉTUDE.....	31
4.1	Présentation du partenaire industriel .....	31
4.2	Statistiques sur les données .....	32
4.3	Application de la méthodologie .....	35
4.3.1	Préparation des données .....	35
4.3.2	Construction des séries temporelles .....	36
4.3.3	Segmentation.....	41
4.3.4	Analyse des résultats .....	51
4.3.5	Synthèse .....	58
4.4	Conclusion.....	59
CHAPITRE 5	CONCLUSION, PERSPECTIVES ET RECOMMANDATIONS.....	60
RÉFÉRENCES	.....	63

## LISTE DES TABLEAUX

Tableau 2-1: Tableau de l'interprétation du critère de silhouette tiré de [Cardoso et Carvalho, 2009] .....	23
Tableau 4-1: Caractéristiques de la segmentation avec six groupes .....	54
Tableau 4-2: Caractéristiques de la segmentation avec dix groupes.....	57

## LISTE DES FIGURES

Figure 2.1: Revenu de l'industrie de vêtement pour le marché Canadien en 2018, tiré de [Trendexna, 2018].....	5
Figure 2.2: Évolution du chiffre d'affaire du commerce en ligne pour l'habillement, tiré de [Statista, 2019].....	5
Figure 2.3 : Les principales mensurations utilisées pour la description de la morphologie humaine, tiré de [Liu et al, 2014].....	7
Figure 2.4 : Exemples de séries temporelle tiré de [Kass et al., 2014]. .....	8
Figure 2.5 : Exemple de demande intermittente, tiré de [Petropoulos et al., 2014].....	10
Figure 2.6 : Effet de la correction de valeurs aberrantes par un filtre médian, tiré de [Ragot, 2019] .....	13
Figure 2.7 : Étape de la segmentation, tiré de [Xu et Wunsch, 2005].....	14
Figure 2.8 : Méthodes de segmentation. ....	15
Figure 2.9 : Graphe des différentes méthodes d'évaluation de la segmentation, tiré de [Aghabozorgi et al., 2015].....	20
Figure 3.1: Étapes de la méthodologie .....	27
Figure 4.1: Évolution nombre de clients ayant passé des commandes .....	32
Figure 4.2: Répartition des produits sur les commandes des clients.....	33
Figure 4.3: Répartition des retours par type de produits. ....	34
Figure 4.4: Extrait de la base des données traitée .....	36
Figure 4.5: Répartition du nombre de commandes des chemises par saison .....	37
Figure 4.6: Transformation des données en séries temporelles .....	38
Figure 4.7: Transformation en série temporelles continues .....	39
Figure 4.8 : Correction des valeurs aberrantes par le filtre médian .....	40

Figure 4.9 : Matrice de distance sur un échantillon de 10 clients .....	42
Figure 4.10 : Résultats de la segmentation du premier type de données en 10 groupes .....	44
Figure 4.11 : Résultats de la segmentation du deuxième type de données en 10 groupes .....	44
Figure 4.12 : Dendrogramme de la classification hiérarchique .....	46
Figure 4.13 : Courbe d'évolution de l'indice de silhouette en fonction du nombre de groupes ....	47
Figure 4.14 : Courbe de la perte d'inertie en fonction du nombre de groupes.....	47
Figure 4.15 : Résultats de la segmentation avec 6 groupes et une normalisation avec la norme Z .....	49
Figure 4.16 : Résultats de la segmentation avec 10 groupes et une normalisation avec la norme Z .....	49
Figure 4.17: Pourcentage des clients pour la segmentation en six groupes. ....	50
Figure 4.18: Pourcentage des clients pour la segmentation en dix groupes.....	51
Figure 4.19 : Courbes des centroïdes de la segmentation en six groupes .....	52
Figure 4.20 : Courbes des centroïdes de la segmentation en dix groupes.....	56

## LISTE DES SIGLES ET ABRÉVIATIONS

DTW Dynamic Time Warping

SSE Sum of Squared Error

CAH Classification ascendante hiérarchique

TADPOLE Time-series Anytime Density Peaks

## CHAPITRE 1 INTRODUCTION

Dans un environnement industriel où la compétitivité est un atout pour la survie des entreprises, la collecte et l'analyse des données deviennent plus une nécessité qu'un luxe [Rygielski et al., 2002]. Qu'il s'agisse des données de ses propres clients, des données externes ou de la corrélation entre les deux, le but est d'être capable d'offrir le meilleur service ou produit sur le marché, avoir des clients satisfaits, et avec les plus faibles coûts possibles pour l'entreprise. L'extraction de ces connaissances à partir des données collectées est maintenant devenue possible grâce à l'émergence de plusieurs techniques d'analyse de données parmi lesquelles on cite principalement la prédiction, la régression, la classification et la segmentation [Goebel et Gruenwald, 1999] [Gorunescu, 2011]. Ces techniques, bien qu'elles datent de plusieurs dizaines d'années, sont toujours en progression et des avancements se font en continu pour chacune d'elles.

L'achat en ligne, ou encore appelé cybercommerce, est un phénomène en pleine expansion. Il est défini par le fait de vendre ou d'acheter un bien ou un service et le payer par le biais d'un site Internet et dont la livraison peut se faire à domicile ou dans un magasin. L'achat en ligne admet de plus en plus de clients potentiels et les études estiment que le nombre de ces derniers devrait atteindre 1,2 million par l'année 2020. Aussi, une recherche du CEFRIO a estimé qu'environ 64% des adultes au Québec ont effectué un achat en ligne en 2018, en comparaison à 46,9% à l'échelle du Canada et 45% en à l'échelle internationale pour même années [CEFRIIO, 2018].

Dans le cadre de notre étude, nous travaillons avec un partenaire dont les clients effectuent des commandes de vêtement en ligne. Le domaine de l'habillement a évolué grandement durant la dernière décennie et de plus en plus de recherches s'en intéressent et essaient de résoudre les problèmes auxquels il fait face. Parmi ces problèmes, on note principalement le manque d'un système de taille unifié. En effet, l'industrie de l'habillement est une industrie globale qui couvre un large éventail de clients. L'absence d'un système de mesures international fait de l'achat des vêtements en ligne une expérience de plus en plus délicate. Le problème de taille induit des coûts cachés pour les entreprises. Le pourcentage des retours dans l'industrie de l'habillement peut atteindre jusqu'à 23% de la marchandise achetée et le pourcentage des retours liés à une taille inadéquate atteint 60 à 70%.

En conséquence, pour remédier à ces problèmes, de plus en plus de solutions ont été élaborées et adoptées par les plateformes d'achat en ligne. Alors que certaines de ces alternatives restent dans un contexte traditionnel, et reposent sur le principe de fournir les clients avec le plus de détails et de mensurations possibles concernant leurs produits, pour assurer qu'ils retrouvent leurs bonnes tailles. D'autres compagnies ont eu recours à de nouvelles technologies, dont la plus répandue est la numérisation 3D, où les clients pourront essayer les vêtements dans des salles d'essayage virtuelles (ebay, Fits.me, Metail, etc.). Une autre façon de résoudre ce problème est de suivre les achats des clients, comprendre leurs morphologies et tailles, pour être ensuite capables de leur recommander la bonne taille. C'est précisément ce qu'on cherche à faire dans ce projet.

Notre projet a donc pour but de comprendre les habitudes d'achat d'une clientèle qui achète des vêtements sur une plateforme en ligne. Nous allons baser nos recherches sur l'analyse des historiques de commandes des clients d'un partenaire industriel, et plus particulièrement les tailles des vêtements qu'ils achètent, pour en déduire leur morphologie. Nous allons par la suite effectuer une segmentation des clients en nous basant sur des séries chronologiques des variations de tailles commandées. Cette segmentation va nous permettre d'observer les différents comportements typiques des clients. Les groupes vont être analysés afin de comprendre leurs structures, et les spécificités de chacun. Les résultats de ce projet serviront comme appuie pour d'autres projets futurs. Ces projets peuvent se centrer sur l'analyse des retours et la corrélation de ceux-ci aux tailles commandées. D'autres projets possibles peuvent porter sur la prédiction de l'évolution des données morphologiques d'un nouveau client, en se basant sur le groupe auquel il appartient.

Le mémoire va comporter quatre chapitres supplémentaires. Au cours du deuxième chapitre, nous allons définir et expliquer les notions utilisées au cours de ce travail à travers une revue de littérature. Nous allons commencer par recenser la littérature pour l'industrie de l'habillement et les données morphologiques, par la suite nous allons définir les séries temporelles, leur caractère intermittent et les moyens de les lisser et de les agréger. Après ça, nous allons aborder la segmentation des marchés, ainsi que les algorithmes de segmentation et les principales métriques de distance.

Au cours du troisième chapitre, nous allons présenter nos objectifs spécifiques et généraux, et nous allons détailler la méthodologie que nous allons suivre. Cette méthodologie peut être décortiquée



en cinq étapes clés, qui sont (1) la préparation de la base de données, (2) la construction des séries temporelles, (3) la segmentation, (4) son évaluation, et finalement (5) l'interprétation des résultats.

Dans le quatrième chapitre, nous allons commencer par présenter le partenaire industriel, ainsi que des statistiques des données analysées. Par la suite, nous allons appliquer la méthodologie mise en place dans le cadre du projet. Les résultats de la segmentation sont par la suite analysés et interprétés. Finalement, le dernier chapitre comportera la conclusion, ainsi que les perspectives et les recommandations du projet.

## CHAPITRE 2 REVUE DE LA LITTÉRATURE

Dans ce chapitre, nous proposons de présenter les concepts de base en liaison avec notre étude. La première partie évoquera l'industrie du vêtement, et les notions de morphologie et de taille. Ensuite, les séries temporelles et en particulier les séries temporelles intermittentes seront présentées. Finalement, la troisième partie va aborder l'approche de la segmentation, ainsi que les algorithmes mis en œuvre et les métriques de distance présentes dans la littérature.

### 2.1 L'industrie du vêtement et « les tailles »

Depuis la nuit des temps, les vêtements ont toujours été « marqueurs » des cultures et des sociétés humaines [Cassagnes-Brouquet et Dousset-Seiden, 2012]. À travers les années, la manière de se vêtir a beaucoup évolué. Depuis l'invention de la machine à coudre, suivie par la révolution industrielle, l'habillement est passé du costume personnalisé conçu sur mesure, au prêt-à-porter (vêtements conçus selon des normes standardisées et vendus en tant que produit fini) fabriqués à la chaîne au sein des usines. Le concept d'industrie du vêtement n'a quant à lui commencé à se généraliser qu'au début du vingtième siècle [Wilson, 2005]. En effet, l'encyclopédie Canadienne définit l'industrie du vêtement comme étant le fait de « confectionner », produire et vendre « des vêtements de consommation courante, des vêtements industriels de même que des uniformes » [Encyclopedia, 2019].

Aujourd'hui, l'industrie du vêtement représente l'un des piliers les plus importants des économies locales et mondiales. En effet, en 2018 cette industrie a atteint une valeur globale de 2,5 billions de dollars américain [McKinsey & Company, 2018], et une valeur d'environ 31 milliards de dollars au Canada, comme on peut le voir sur la figure 2.1 [Trendex North America, 2018]. L'industrie de vêtement est aussi en expansion continue. Ceci est lié à l'intégration des technologies avancées dans les processus de production, mais aussi indéniablement grâce à la croissance du commerce en ligne. Dans le contexte de la mondialisation, le commerce en ligne a présenté une nouvelle opportunité pour les manufacturiers pour vendre leurs marchandises et d'étendre leurs chaînes logistiques au-delà des bordures géographiques.

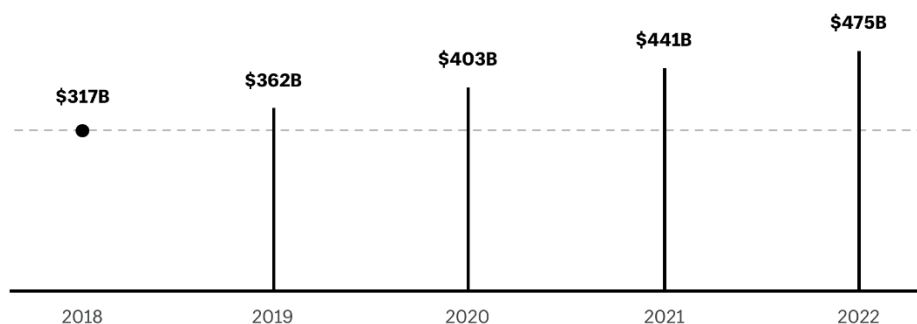


Figure 2.1: Revenu de l'industrie de vêtement pour le marché Canadien en 2018, tiré de [Trendexna, 2018].

C'est grâce au commerce électronique que le revenu global de l'industrie du vêtement a prévu d'atteindre 362 milliards de dollars en 2019, comme le montre la figure 2.2. Sur cette figure, on peut voir aussi que les revenus du commerce électronique pour l'habillement sont prévus augmenter d'environ 50% entre les années 2018 et 2022, atteignant une valeur globale de 475 milliards de dollars.

### Ecommerce clothing segment

Worldwide revenue by billions of USD



Data via Statista

Figure 2.2: Évolution du chiffre d'affaire du commerce en ligne pour l'habillement, tiré de [Statista, 2019]

Parce que le marché de l'industrie du vêtement est en expansion continue, les enjeux auxquels il fait face sont de plus en plus nombreux. Parmi les principaux défis de ce domaine [Sen, 2008] met l'accent sur le court cycle de vie des produits, la demande souvent imprévisible et la chaîne d'approvisionnement qui est assez sophistiquée. [Tokatli, 2007] de son côté insiste sur le fait que les goûts et tendances des clients sont de plus en plus imprévisibles et changent plus rapidement. Ce changement du comportement de la clientèle est expliqué par l'évolution de la société et son influence principalement par les réseaux sociaux émergents. Devant cette évolution du comportement des consommateurs, les manufacturiers sont aujourd'hui obligés de satisfaire un public plus exigeant, en offrant des expériences plus personnalisées.

Dans le contexte actuel de mondialisation, un autre problème auquel fait face l'industrie du vêtement est la différence des morphologies à travers le monde, ainsi que la différence des standards de tailles entre les diverses marques. Dans leur livre [Winks et al., 1997] abordent la problématique de la morphologie, en introduisant l'évolution qui a eu lieu dans le domaine du prêt-à-porter. D'après les auteurs, parmi les principaux problèmes rencontrés par les marques du prêt-à-porter, est le fait de mettre au point un ensemble fini de tailles pour une population non seulement ayant des formes de corps différentes, mais aussi avec des préférences et des goûts distincts. Pour pallier ce problème, les entreprises doivent créer des vêtements qui touchent un grand éventail de personnes d'un côté et les garder satisfaits d'un autre côté.

La raison pour laquelle on trouve plusieurs systèmes de tailles malgré la tentative de l'ISO à créer une table de tailles standard, surgit dans le fait que certaines marques ont une clientèle ciblée sur laquelle ils se basent en établissant leurs tailles.

Ainsi, une taille est caractérisée par un ensemble de mensurations qui sont corrélées l'une à l'autre. Parmi les plus importantes, on peut citer la longueur des jambes, le tour de taille, le tour de poitrine, la largeur des hanches et la largeur des épaules. Ces mensurations peuvent être observées sur la figure 2.3.









Factors	Height		Girth		Small joint girth	Waist hip	Upper torso	Shoulder
FPs								
	Neck height	Inseam	Breast girth	Waist girth	Wrist girth	Distance waist to hip	Back length	Shoulder width

Figure 2.3 : Les principales mensurations utilisées pour la description de la morphologie humaine, tiré de [Liu et al, 2014]

Beaucoup de travaux se sont intéressés à l'intégration des techniques d'analyse de données au domaine de l'habillement. [Zheng et al, 2007] ont cherché, à partir d'images 3D, d'analyses factorielles et de segmentation, à choisir des mensurations clés des tailles des soutiens-gorges pour les femmes en Chine. Aussi [Hui et Ng, 2009] pour leur part, se sont basés sur des réseaux de neurones et les algorithmes de régression pour prédire l'efficacité des coutures sur un tissu tricoté. [Han et Nam, 2011] ont aussi employé des techniques d'apprentissage statistique dans le cadre de détection du corps humain par des systèmes de vision 3D. Leurs algorithmes ont permis d'améliorer la performance de ces systèmes en enrichissant sa base d'entraînement avec des modèles diversifiés et différents l'un de l'autre. Et finalement, depuis 2003, [Cordier et al., 2003] ont mis en place une boutique virtuelle, avec des mannequins virtuels, hommes et femmes, adaptables pour plusieurs formes de corps différents et des vêtements variés. Leur application permettrait aux utilisateurs de changer l'arrière-plan de la chambre d'essayage virtuelle, ainsi que d'effectuer quelques mouvements, afin de permettre aux clients de voir le produit dans des états variés avant de l'acheter.

Ainsi, nous pouvons voir qu'il existe plusieurs travaux qui mettent en œuvre les techniques d'analyse de données dans des recherches abordant la morphologie humaine et les tailles. Et parmi les principaux problèmes mis en exergue, il y a le fait que les donneurs d'ordres dans cette industrie ne semblent pas suivre les normes de tailles nationales ou internationales. En plus, chaque fabricant semble adhérer à des normes qui satisfont une clientèle cible. Le troisième point, est la difficulté d'avoir des tailles qui iront à différentes formes de corps.

## 2.2 Séries temporelles intermittentes

Les séries temporelles sont un type de données définies comme étant une séquence de valeurs obtenues par des mesures répétitives au cours du temps d'un phénomène précis [Han et al, 2006]. Les séries temporelles illustrent l'évolution d'un type de données au cours du temps par des observations ordonnées et mesurées à des intervalles égaux [Kurbalija et al., 2012], [Rakthanmanon et al, 2012]. La recherche dans la littérature révèle une unanimité sur l'intérêt croissant autour des séries temporelles au cours des dernières décennies. [Yang et Wu, 2006] incluent l'analyse des série temporelles parmi les dix défis de l'analyse de données. Le nombre croissant d'études reposant sur des séries temporelles peut s'expliquer par la versatilité de celles-ci et qu'elles peuvent représenter des phénomènes dans différents domaines. Ainsi, nous pouvons trouver des analyses de séries temporelles dans les domaines de la finance [Fu et al, 2006], du multimédia [Ratanamahatana et Keogh, 2005], de la psychologie [Kurbalija et al., 2012], de la génétique [Bar-Joseph et al, 2002], etc. Des exemples de séries temporelles sont donnés par la figure 2.4.

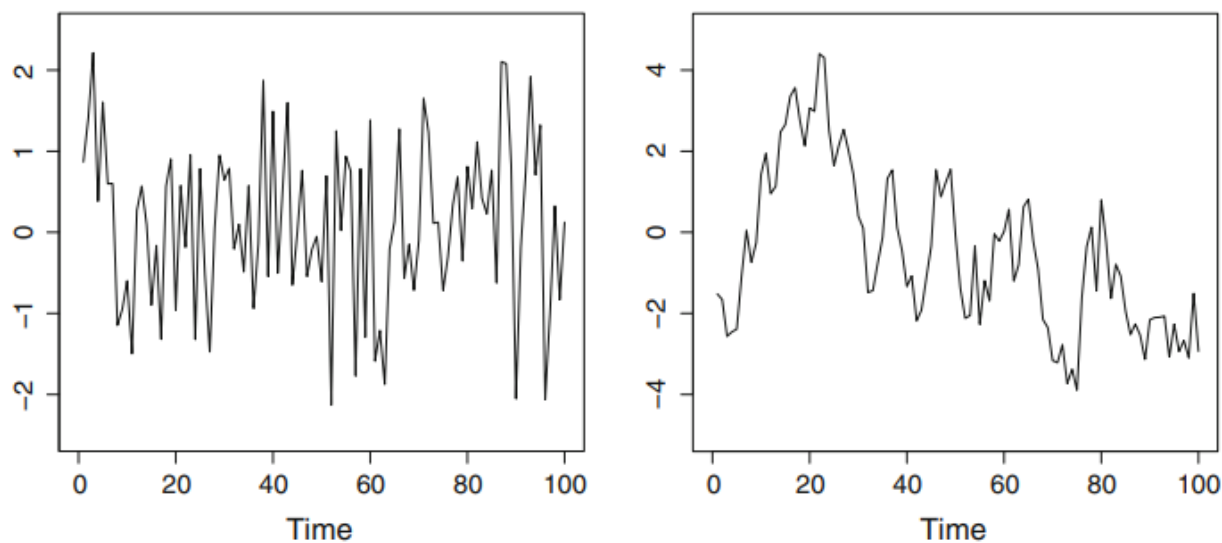


Figure 2.4 : Exemples de séries temporelle tiré de [Kass et al., 2014].

Bien que variées, les applications traitant les séries temporelles peuvent être divisées en trois importantes familles [Aghabozorgi et al, 2015] :

- La recherche des tendances dans les données;

- La prédiction des comportements futurs;
- L'évaluation des corrélations entre les séries temporelles.

Ainsi, plusieurs applications se basent sur l'analyse des séries temporelles telles que l'indexation, la classification, la segmentation, la prédiction, etc. D'après [Rakthanmanon et al, 2012], la majorité de ces applications se basent sur une recherche de similarité. La recherche de similarité peut s'appuyer sur une comparaison des séries temporelles en leur totalité, ou aussi appelé « whole matching », ou bien en comparant une partie de série temporelle à une autre complète, et on appelle ceci le « subsequence matching » [Sadahiro et Kobayashi, 2014].

Parmi les particularités des séries temporelles, on peut distinguer la saisonnalité, la tendance, le cycle, et l'aléatoire. Dans ce qui suit nous allons nous intéresser au caractère intermittent des séries temporelles.

Les séries temporelles intermittentes sont des séries, ayant peu d'observations, qui diffèrent grandement d'une période à l'autre, avec des intervalles d'occurrence irréguliers [Croston, 1972], [Kourentzes, 2013]. On retrouve des séries temporelles intermittentes souvent en étude de demande lorsque les produits sont acquis de manière aléatoire, non fréquente, et avec des volumes irréguliers [Syntetos et Boylan, 2010], [Persson et al., 2017]. Une série de demande intermittente présente plusieurs intervalles de demande nulles. [Johnston et al., 2003] observent que les produits ayant une demande intermittente peuvent représenter jusqu'à 60% du total de ventes pour un fabricant ou un détaillant. [Bartezzaghi et al., 1999] expliquent l'intermittence dans le secteur industriel par cinq facteurs qui sont : le nombre des clients, leur hétérogénéité, la fréquence, la variation et la corrélation entre leurs demandes. Une demande intermittente peut être observée par exemple dans le domaine de vente de maintenance en aéronautique [Ghobbar et Friend, 2003], les pièces de rechange de voitures [Syntetos et Boylan, 2005] ou des pièces de rechange pour les machines lourdes [Willemain et al., 2004]. Le caractère intermittent de la demande et le manque d'historique de données rendent la prédiction de la demande et la gestion des stocks de tels produits plus complexes [Jiantong et Biyu, 2009], [Petropoulos et al., 2014].

La figure 2.5 affiche un exemple d'une série temporelle intermittente. On peut observer sur la figure la présence de plusieurs intervalles où la demande est nulle. La quantité de la demande est également variable et irrégulière.

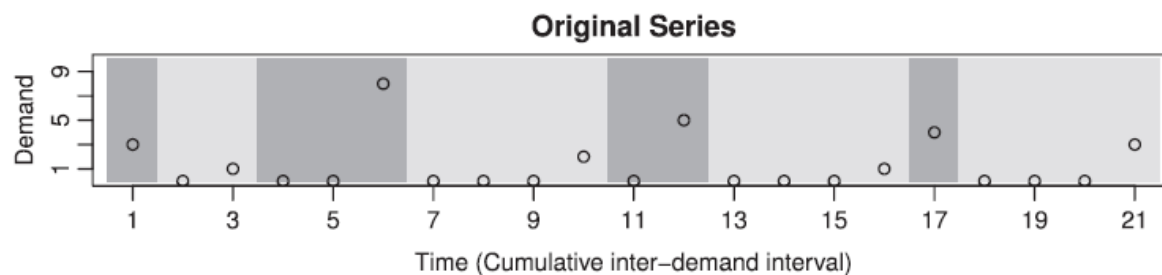


Figure 2.5 : Exemple de demande intermittente, tiré de [Petropoulos et al., 2014]

Pour pallier les problèmes de la prédiction avec des demandes intermittentes, une approche classique consiste à lisser les données ou les agréger. Dans la littérature, il existe plusieurs méthodes pour lisser les demandes intermittentes afin de rendre possibles leur exploitation. Parmi ces méthodes, on trouve le calcul des moyennes mobiles simple, le lissage exponentiel simple (SES) [Brown, 1963] et la méthode présentée par [Croston, 1972]. La méthode de Croston n'est autre qu'une alternative du lissage exponentiel. Toutes les deux permettent de corriger l'irrégularité dans les données. Néanmoins, parfois leur utilisation risque aussi de biaiser les résultats à cause de la perte d'information induite [Syntetos and Boylan, 2001]. Pour ceci, plusieurs modèles ont été développés afin de corriger ces défauts.

On peut citer par exemple le travail de [Murray et al., 2018]. Ce dernier a mis en place un modèle appelé « ASACT ». Il se base sur une suite d'agrégation au niveau temporel le plus fin, un lissage avec la méthode de Croston, et ensuite une réagrégation au niveau temporel désiré. En outre, l'agrégation des données est aussi utilisée comme outil de lissage [Petropoulos et al., 2016]. Le principe de l'agrégation est simple. Il s'agit de synthétiser plusieurs valeurs en une seule variable représentative [Grabisch et al., 2011]. Les supports d'agrégation peuvent être des intervalles temporels aussi bien que des groupes de clients [Syntetos et al., 2016]. Le but est d'alléger les données, et enlever l'intermittence.

Cependant, l'agrégation n'a pas été exempte de critiques. [Rehm et Gmel, 2001] supportent le fait que travailler avec des données agrégées risque de faire perdre de l'information, et affaiblit la consistance des modèles obtenus. [Tiao, 1972] a établi que l'agrégation ne permet pas d'améliorer les résultats des prédictions, en comparaison aux résultats obtenus avec des données non agrégées. [Vliegenthart, 2014] affirme que le processus d'agrégation est fortement avantageux dans le cadre



des études ayant des fins de causalité, surtout au niveau global et non individuel. Finalement, [Jin et al., 2015] ont prouvé que l'agrégation permettait de réduire la variabilité et de donner des prédictions plus exactes. Pour l'étude des tendances d'une population où on cherche à discerner des tendances globales plutôt que des comportements individuels, il est donc avantageux de considérer des données agrégées.

Une autre problématique rencontrée avec les séries temporelles, est la présence de valeurs aberrantes. Les travaux dans la littérature sont unanimes à définir les valeurs aberrantes comme étant des observations isolées, anormales et « distantes » de manière remarquable par rapport au reste des valeurs d'une série [Grubbs, 1969], [Barnett et Lewis, 1984]. Une valeur aberrante est aussi considérée comme une valeur « erronée » [Munoz-Garcia et al., 1990]. Pour ceci, il est nécessaire de les considérer individuellement afin de décider de les garder ou les supprimer [Shen et al, 2004].

Il existe dans la littérature plusieurs méthodes de détection de valeurs aberrantes, réparties en méthodes directes et indirectes. Un recensement de toute ces méthodes peut être trouvé dans l'article de [Billor et Kiral, 2008], on peut en citer le test de Dixon [Dixon, 1950], la matrice d'influence [Peña and Yohai, 1995] ou l'estimateur M [Huber, 1973]. Cependant, il est intéressant d'ajouter que dans certains cas, la qualification d'une valeur aberrante repose sur un jugement subjectif, et dépendant du contexte ou de la nature même des données étudiées [Planchon, 2005]. En ce qui concerne le traitement des valeurs aberrantes, il existe deux manières de les traiter. La première consiste à les supprimer et la deuxième à les substituer [Ragot, 2019]. La suppression des valeurs jugées anormales a pour conséquence de générer des valeurs manquantes dans les séries temporelles, et donc la perte d'information. Dans ce qui suit, nous nous intéressons à la substitution des valeurs aberrantes dans le cas des séries temporelles par le biais de filtres.

Parmi les filtres les plus rencontrés dans un tel contexte, nous trouvons le filtre par moyenne mobile et le filtre médian. Le filtre par moyenne mobile est utilisé depuis des décennies pour le lissage des données [Savitzky et Golay, 1964]. Il est également l'une des méthodes les plus utilisées pour le lissage des données [Azami et al., 2012]. Le filtre médian quant à lui a été introduit par [Tukey, 1971] dans le but de lissage séries temporelles dans un contexte économique. Il a été par la suite largement étudié dans divers domaines tels que l'analyse de signaux [Frieden, 1976], le traitement de la voix [Rabiner, 1975] et surtout en analyse d'images [Wang et Lin, 1997], [Astola et al, 1988].

Les deux algorithmes reposent sur un même principe. En parcourant les données, la k-ème valeur de la série est remplacée soit par la valeur moyenne ou par la médiane, dans la fenêtre de valeurs  $(n-k)$  et  $(n+k)$ , où  $n$  est la largeur de fenêtre fixée [Stone, 1995]. La formule de ce filtre est donnée par l'équation (1).

$$y(n) = med [x(n - k), x(n - k + 1), \dots, x(n), \dots, x(n + k - 1), x(n + k)] \quad (1)$$

D'une part, l'inconvénient du filtre par moyenne mobile est sa sensibilité aux valeurs extrêmes. Une valeur trop extrême dans un intervalle influencera la moyenne de manière significative. La médiane, d'une autre part, est moins sensible à ces valeurs écartées. Elle permet de les enlever, sans trop déformer la série initiale. Afin de ne pas induire une grande altération à la série, il suffit de choisir une largeur de fenêtre assez petite, ce qui permettra d'enlever seulement les pics trop éloignés [Moore et Jorgenson, 1993]. L'effet de déformation induit par le filtre peut être mesuré par la distorsion, présentée par [Stone, 1995]. Dans leur article, la distorsion présentée comme la racine de la moyenne des carrées des différences entre la série initiale et la série filtrée, point par point.

$$distorsion = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2} \quad (2)$$

Un exemple de l'effet d'un filtre médian sur une série temporelle est affiché sur la figure 2.6. Sur la figure 2.6, il est possible de voir que la série temporelle initiale présente deux valeurs aberrantes. L'application du filtre médian a permis d'enlever ces deux valeurs, tout en conservant l'allure de la courbe.

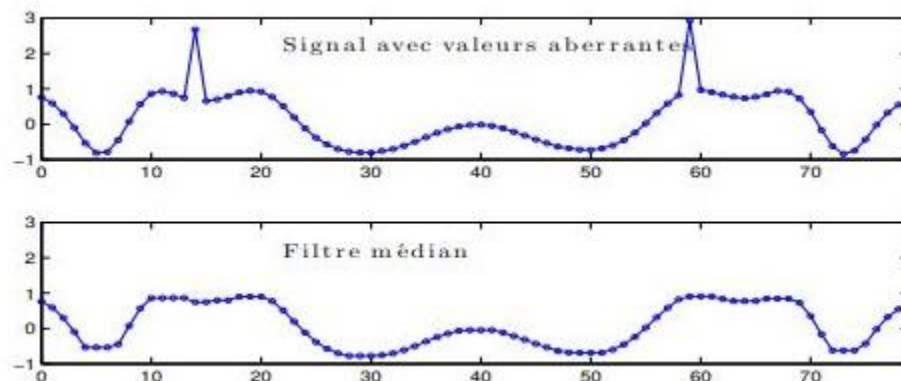


Figure 2.6 : Effet de la correction de valeurs aberrantes par un filtre médian, tiré de [Ragot, 2019]

## 2.3 La segmentation

[Berkhin, 2006] et [Han et al., 2011] définissent la segmentation comme étant le processus de diviser un groupe d'entités hétérogènes en sous-groupes homogènes. [Miller et Han, 2001] et [Backer et Jain, 1981] ajoutent que le but de la segmentation est de découvrir la structure des données et de créer des groupes, les plus semblables entre eux et les plus distincts des autres groupes, tout en se basant sur une mesure de similarité. De ce fait, beaucoup définissent les segments en termes d'homogénéité intra classe et hétérogénéité interclasse [Hansen et Jaumard, 1997], [Jain et Dubes, 1988]. La segmentation est appliquée dans un large éventail de domaines, notamment en biologie et l'analyse de l'ADN [Ben-Dor et al, 1999], en analyse des textes [Dhillon et al., 2001], en transport [Agard et al., 2013], en reconnaissance de formes et segmentation d'images [Bowyer et Ahuja, 1996], [Duda et Hart, 1973], et la liste s'étend de plus en plus.

En l'occurrence, la segmentation des marchés consiste à diviser les clients dans un marché spécifique afin de cerner les groupes ayant les mêmes comportements, besoins, et attentes [Kotler et al., 1991]. Ceci permet aux entreprises de discerner quelle portion de la population consomme ses produits, et développer des stratégies marketing plus ciblées. La segmentation vise en fait à expliquer un phénomène en se basant sur des variables descriptives représentant la réalité [Aurier, 1989]. La segmentation peut donc se baser sur des critères comportementaux, tels que les achats précédents, géographiques, tels que l'appartenance à des régions chaudes ou froides, psychographiques, tels que les intérêts et démographiques, tels que le genre ou la catégorie d'âge.

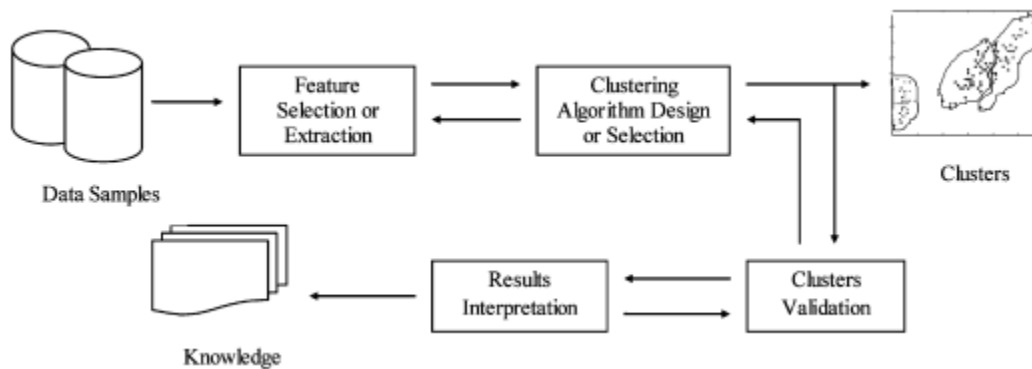


Figure 2.7 : Étape de la segmentation, tiré de [Xu et Wunsch, 2005]

La segmentation peut être décomposée en quatre étapes principale, comme le montre la figure 2.7. La première étape, consiste à analyser les données et extraire les attributs importants à la segmentation. Cette étape vise à avoir des données plus ciblées pour le cas d'études particulier.

La deuxième étape, repose sur le choix de la méthode et du critère de similarité à employer. Tel qu'expliqué plus haut, la segmentation est fondée sur le fait de diviser des entités en groupes similaire. Pour cette raison, le choix d'un critère pour évaluer leur similarité est très important. Dans la littérature, il existe d'innombrables algorithmes de segmentation, chacun plus performant avec un type particulier de données, et donc choisir lequel sera utilisé pour chaque cas d'études est indispensable.

Par la suite, la segmentation passe par l'étape de validation des groupes créés. [Xu et Wunsch, 2005] mettent l'accent sur l'importance de cette étape, donner confiance aux chercheurs quant à la qualité de la segmentation, et pour pouvoir comparer les performances des différentes méthodes.

Finalement, la dernière étape de la segmentation est l'évaluation et l'interprétation des résultats. Cette phase sert à comprendre la composition des groupes créés et d'en extraire l'information recherchée.

### 2.3.1 Algorithmes de segmentation

La majorité des études dans la littérature divisent les algorithmes de segmentation en deux grandes familles. Sur la figure 2.8, on distingue les méthodes hiérarchiques et les méthodes de partitionnement. Il existe aussi d'autres méthodes que nous allons aborder de manière générale.

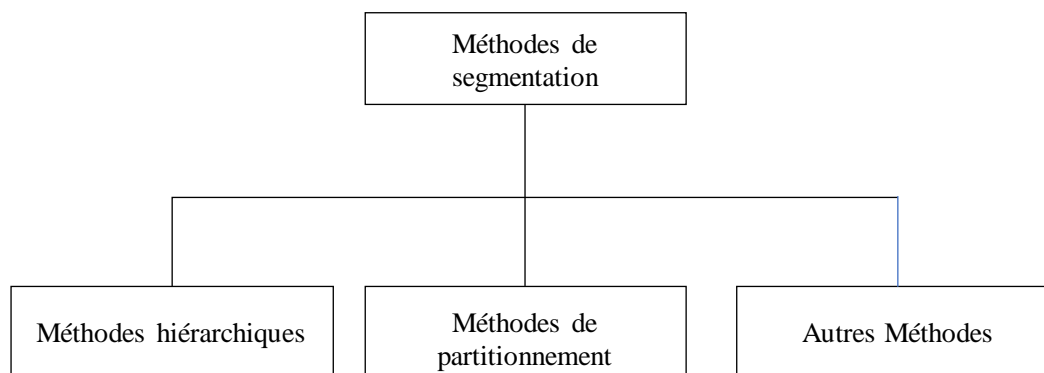


Figure 2.8 : Méthodes de segmentation.

### 2.3.1.1 Les méthodes hiérarchiques

La segmentation hiérarchique consiste à arranger les entités dans des groupes en suivant une structure hiérarchique en se basant sur une matrice de distance ou matrice de similarité. En effet, nous parlons de distance lorsqu'on travaille avec des données quantitatives et de mesure de similarité avec des données qualitatives. Le résultat d'une telle segmentation est présenté sous la forme d'un arbre aussi appelé dendrogramme, sur lequel on peut aussi lire le nombre de groupes optimal pour la segmentation [Berry and Linoff, 2004].

Les méthodes hiérarchiques peuvent être divisées en méthodes par agglomération et par division [Jain et Dubes, 1988], [Kaufman et Rousseeuw, 2009]. Les méthodes par agglomération commencent par  $N$  groupes, où chaque groupe contient une seule entité. Ces derniers sont groupés ensemble en fonction de leurs similarités, jusqu'à avoir un seul groupe qui contient toutes les unités. L'algorithme par division est l'inverse exact, il commence par un groupe contenant les  $N$  individus, et par division on finit par avoir des singletons individuels. Toutefois, comme l'illustre [Everitt et al., 2001], chaque division admet  $2^{(N-1)}$  possibilités, ce qui fait que les méthodes par division sont trop coûteuses, et c'est pourquoi les méthodes par agglomération sont généralement les plus utilisées.

Parmi les principales méthodes utilisées lors de l'implémentation d'une segmentation hiérarchique agglomérative, il y a les méthodes « single-linkage », « complete-linkage » et « WARD » [Ward, 1963]. Le principe de ces méthodes est de regrouper les points en s'assurant à chaque pas de minimiser les distances intra-classe et de maximiser les distances interclasse [Gonzalez, 2008]. La méthode single linkage mesure la similarité en calculant la plus petite distance entre un objet d'un

groupe et tous les éléments dans les autres groupes. Cette méthode est versatile et s'applique avec différents types de données. Son inconvénient est qu'elle ne priorise pas l'homogénéité intragroupe lors des agglomérations. La méthode complete-linkage mesure le maximum de distance entre les éléments des groupes. Cette méthode produit des groupes plus compact, et elle est adaptée pour plusieurs types d'applications [Jain et Dubes, 1988]. Finalement, la méthode WARD se base sur une mesure de l'inertie, qui représente la moyenne des carrés des distances entre les éléments d'un groupe. À chaque étape, les deux groupes minimisant l'augmentation la somme des distances quadratiques entre tous les groupes sont regroupés.

Les méthodes de segmentation hiérarchique sont connues pour leur flexibilité, leur facilité de mise en œuvre et la possibilité de les appliquer à différents types de données [Berkhin, 2006]. On leur reproche principalement le fait que l'assignation des points n'est pas itérative. À chaque fois qu'un objet est classé, l'algorithme n'évalue pas sa position et la possibilité de le classer dans un autre groupe. Pour cette raison, des améliorations aux algorithmes hiérarchiques ont été conçues, et on peut citer les travaux de [Fisher, 1996], [Karypis et al., 1999].

### **2.3.1.2 Les méthodes de partitionnement**

À l'inverse des méthodes hiérarchiques, les méthodes de partitionnement se basent sur le groupement de toutes les entités en un nombre fixé de groupes, mais sans structure hiérarchique. La particularité de ces algorithmes, c'est que l'assignation des points aux groupes y est itérative. En d'autres termes, l'assignation des points est évaluée à chaque itération, pour reconsidérer s'ils sont classés correctement, et le réassigner si c'est ce n'est pas le cas, et ceci est l'un des avantages de cette méthode [Berkhin, 2006].

Les deux algorithmes les plus connus parmi les méthodes par partitionnement sont k-means [Hartigan et Wong, 1979], [MacQueen et al., 1967] et k-medoids [Kaufman et Rousseeuw, 1990], [Ng et Han, 2002]. Les principes des deux méthodes sont similaires. Ils s'appuient sur une fonction objectif qui va renseigner sur la similarité entre les points, et sur le choix d'un représentant pour chaque groupe. En ce qui concerne k-means, comme son nom l'indique, chaque cluster est représenté par son centroïde ou son point moyen, alors que pour k-medoid c'est les medoids qui sont représentatifs. Les deux représentants sont calculés de manière à avoir une distance cumulée minimale par rapport à tous les points du groupe [Struy et al., 1997]. La différence entre les deux c'est que les medoids sont des points existants du groupe, alors que le centroid, est simplement un

point moyen dans l'espace des points [Aggarwal et al., 1999]. À chaque itération, la distance de chaque point par rapport à chacun des représentants des groupes est calculée, le point est assigné, et de nouveaux représentants sont calculés, jusqu'à obtenir le nombre voulu de groupes.

Malgré les avantages des méthodes par partitionnement, leur plus grand défaut c'est le fait que la construction des groupes dépend fortement du choix des points de départ. Et par conséquent appliquer ces algorithmes deux fois sur les mêmes données ne va pas aboutir aux mêmes résultats [Pena et al., 1999]. Aussi, avec ces algorithmes, le choix du nombre de classes reste à la décision de l'utilisateur, du moment où ils n'offrent pas un moyen pour décider, et ils sont sensibles aux points aberrants. [Velmurugan et Santhanam, 2010] et [Arora et al., 2016] ont effectué une comparaison entre les deux méthodes, pour trouver que face à un grand nombre de points, ce sont les k-medoids qui permettent d'avoir un temps de calcul plus intéressant. Pour toutes ces raisons, plusieurs améliorations ont été apportées aux deux algorithmes, dont on peut citer les travaux de [Bradley et Fayyad, 1998], [Pelleg et Moore, 2000], [Kanungo et al., 2002], [Van der Laan et al., 2003].

### **2.3.1.3 Autres méthodes de segmentation**

Dans cette partie nous avons présenté en détail les méthodes hiérarchiques et de partitionnement vu que celles-ci sont les plus présentes dans la littérature. Hormis ces deux grandes familles, il existe beaucoup d'autres méthodes de segmentation. Les plus importantes sont les méthodes basées sur les réseaux de neurones artificiels [Bishop, 1995], [Pal et al., 1993]. Pour ces algorithmes, la segmentation est traitée comme un problème d'optimisation, avec une fonction de perte à minimiser [Xu, 2005]. Une autre famille de segmentation est celle des Méthodes de segmentation floue ou le « fuzzy clusterin ». À la différence des autres méthodes présentées plus tôt où chaque élément n'est assigné qu'à un seul groupe, avec la segmentation floue les éléments peuvent appartenir à plusieurs groupes avec un degré d'appartenance [Zadeh, 1965]. On peut citer aussi les méthodes de segmentation basées sur la densité tel que l'algorithme TADPole élaborée par [Begum et al., 2015], la segmentation à base de probabilité [Ng et Han, 2002], [Kaufman et Rousseeuw, 2009], et la segmentation basée sur la forme tel que k-Shape de [Paparrizos et Gravano, 2015]. Un recensement bien détaillé peut être trouvé dans les travaux de [Berkhin, 2006].

### 2.3.2 Métriques de distances

Afin de pouvoir regrouper les clients en segments, il est nécessaire de calculer une distance entre les clients, deux à deux [Price et al., 2009]. Il existe dans la littérature plusieurs métriques de calcul d'écart. Parmi les distances les plus utilisées, on trouve la distance euclidienne [Berkhin, 2006] et la distance de Manhattan [Bakar et al., 2006]. Les distances euclidiennes et de Manhattan se calculent respectivement par les formules des équations (3) et (4).

$$d_{euclidienne}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (3)$$

$$d_{manhattan} = \sum_i |x_i - y_i| \quad (4)$$

Bien qu'elles soient très efficaces, le problème avec ces distances est qu'elles ne sont pas adaptées pour des séries où l'on travaille avec des données qualitatives ou lorsqu'on cherche à analyser un comportement temporel [He et al., 2018]. En effet, quand appliquées à des séries temporelles, ces métriques calculent la distance point par point, ne prenant pas en considération, la ressemblance de forme entre des séries lorsque celles-ci sont décalées dans le temps.

Pour remédier à ce problème [Skoe et Chiba, 1978] ont développé une technique nommée « Dynamic Time Wrapping » (DTW). Cette méthode a été conçue en premier lieu pour des systèmes de reconnaissance vocale. Cet algorithme se base sur le calcul de similarité en forme au lieu de calculer la concordance sur l'axe temporel [Aghabozorgi et al., 2015]. Pour comprendre le principe du Dynamic Time Warping, considérons deux séries temporelles A et B, de longueurs respectives m et n, telles que :

$$A = a_1, a_2, \dots, a_m$$

$$B = b_1, b_2, \dots, b_n$$

Pour aligner la similarité entre ces deux séries temporelles, une matrice est calculée. Chaque élément de la matrice contient la distance quadratique entre les éléments des deux séries temporelles  $d(a_i, b_j) = (a_i - b_j)^2$ . Afin de trouver le meilleur alignement, le Dynamic Time Warping recherche le chemin optimal qui minimise la distance cumulée entre tous les éléments des deux séries, comme on peut le voir sur la figure 2.9. On peut le voir sur la figure, qu'à chaque pas, la



recherche du chemin optimal est accomplie sur une fenêtre de distorsion, dont la largeur est définie au préalable. La formule de la distance DTW est donnée par l'équation (5). Une fenêtre de largeur égale à 1, ramène à la distance euclidienne.

$$DTW(A, B) = \min \sqrt{\sum_{i,j} d(ai, bj)} \quad (5)$$

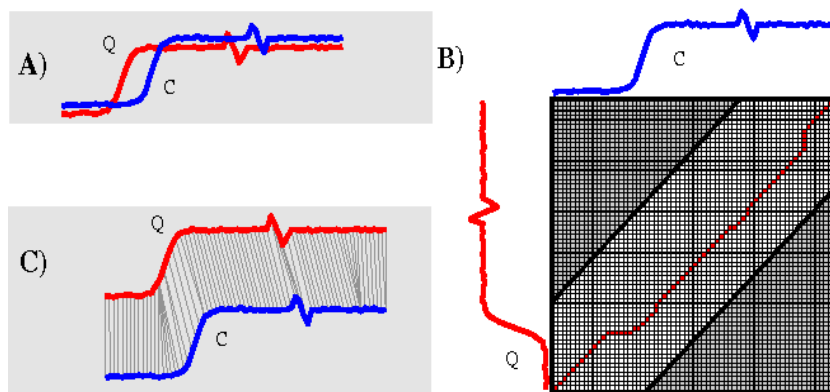


Figure 1.9 : Principe du Dynamic Time Warping [Ratanamahatan et Keogh, 2013]

Néanmoins, beaucoup de critiques ont été adressées à cette méthode, concernant surtout son temps de calcul coûteux [Zhang et al., 2006], [Chadwick et al., 2011]. Mais les progrès technologiques et les ordinateurs de plus en plus performants de nos jours, ainsi que les améliorations appliquées à l'algorithme ont permis de réfuter ces critiques [Wang et al., 2013]. [Rakthanmanon, et al., 2013] ont en fait, par leurs travaux, prouvé qu'il est possible de diminuer notablement le temps de calcul du DTW appliqué sur des bases de données massives.

Dans cette section nous avons présenté trois métriques de distances utilisées dans la littérature, notamment la distance euclidienne, la distance de manhattan et le Dynamic Time Warping. Le DTW étant la métrique la plus adaptée au traitement des séries temporelles, nous allons l'utiliser pour notre cas d'étude.

### 2.3.3 Évaluation de la segmentation:

Souvent, la répartition obtenue par un processus de segmentation n'est pas connue à priori. C'est pourquoi une évaluation des résultats est nécessaire. L'évaluation de la segmentation consiste à valider les résultats obtenus suite à la création des groupes [Kim et al, 2017], et s'assurer que ces résultats sont représentatifs. Elle vise à mesurer la qualité des partitions obtenues [Cardoso et Carvalho, 2009]. L'évaluation de la segmentation peut se baser sur deux critères principaux: la visualisation, et la mesure de la qualité de la segmentation par des indicateurs scalaires [Aghabozorgi et al., 2015]. La figure 2.9 représente ces différentes méthodes d'évaluation présentes dans la littérature. Comme on peut le voir sur la figure 2.9, les méthodes d'évaluation sont divisées principalement en deux familles, soient l'évaluation par une visualisation ou par une mesure scalaire. Les mesures scalaires se basent sur des critères externes (l'indice de pureté, l'indice de Jaccard et l'entropie.), sur des critères internes (le C-index, la somme des erreurs quadratiques et l'indice de silhouette.) ou sur des critères relatifs.

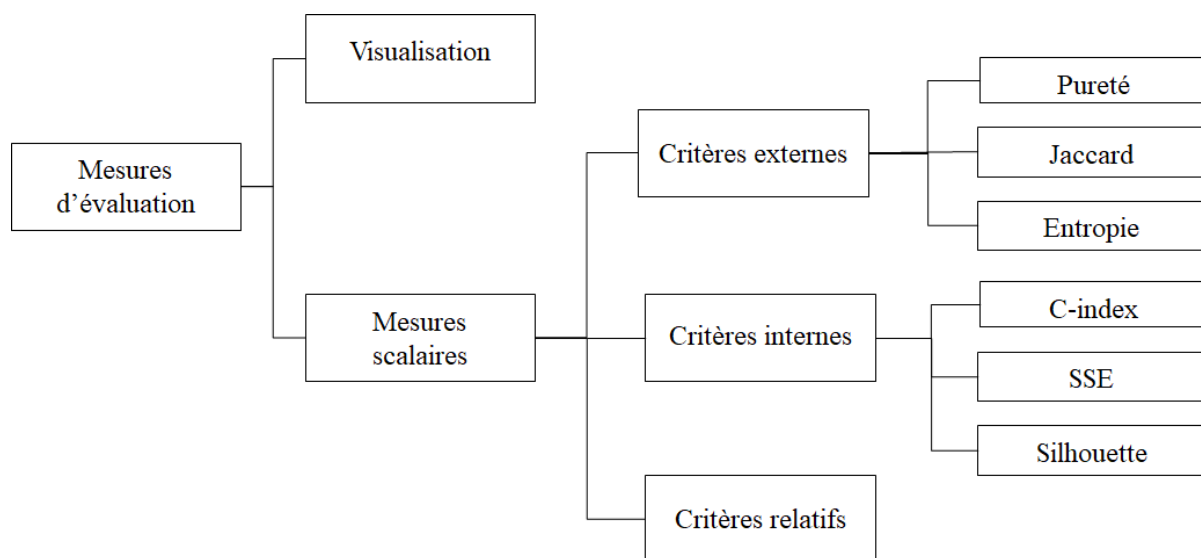


Figure 2.9 : Graphe des différentes méthodes d'évaluation de la segmentation, tiré de [Aghabozorgi et al., 2015]

La visualisation repose sur une description de la partition obtenue. Les groupes sont examinés individuellement, et en comparaison avec l'ensemble. Pour valider une segmentation par la visualisation, les groupes doivent être distincts, et présenter des profils diversifiés. La visualisation est un processus subjectif [Hair et al., 1998], et nécessite une connaissance profonde du domaine, et des résultats attendus. Pour ceci, des indices scalaires plus fiables sont souvent utilisés. Comme

le montre la figure 2.10, les indices scalaires se divisent en critères externes, internes ou relatifs [Wedel et Kamakura, 2000]. Les critères externes sont obtenus en comparant la partition obtenue avec une partition réelle du groupe d'individus. Ils reposent sur des informations exogènes et nécessitent d'avoir une structure proposée de l'échantillon étudié en vue de le comparer aux résultats obtenus [Halkidi et al., 2001]. Parmi les critères externes les plus utilisés, on cite la pureté, l'entropie et l'indice de Jaccard.

La pureté d'une segmentation est un indice qui varie entre 0 et 1. Considérons un ensemble d'éléments ayant différentes classes  $t_j$ , segmentés sur un nombre  $k$  de groupes. La première étape du calcul de l'indice de pureté est de caractériser chaque groupe par sa classe dominante. L'indice de pureté est obtenu en calculant le rapport des nombres d'éléments bien classés dans un groupe, par le nombre total d'éléments. Une segmentation parfaite a une pureté égale à 1, tandis qu'une mauvaise aura 0 [Wang et al., 2006], [Rendón et al., 2011], [Kim et Ramakrishna, 2005]. L'équation de l'indice de pureté est présentée par l'équation (6) ci-dessous [Deepa et Revathy, 2012].

$$Pureté = \frac{1}{N} \sum_{i=1}^k \max_j |c_i \cap t_j| \quad (6)$$

Avec  $N$  le nombre total d'éléments,  $k$  le nombre de groupes,  $c_i$  présente le groupe et  $t_j$  la classe avec le plus d'éléments dans  $c_i$ .

L'entropie repose sur une mesure de la variabilité de la distribution des éléments à l'intérieur de chaque classe [Kou et al., 2014], [Zhao et Karypis, 2001]. L'entropie et la pureté sont inversement liées. Lorsque la pureté est élevée et l'entropie est faible, nous pouvons déduire que la qualité de la segmentation est bonne [Deepa et Revathy, 2012]. L'entropie est donnée par l'équation (7).

$$Entropy = \sum_{r=1}^N \frac{n_r}{n} \left( -\frac{1}{\log q} \sum_{i=1}^q \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \right) \quad (7)$$

L'indice de Jaccard est employé pour comparer la similarité entre une partition à priori des données et le résultat de segmentation de ces mêmes données. L'indice de Jaccard détermine le nombre de

paires d'éléments assigné à la même classe, dans les deux partitions. Cet indice varie entre 0 et 1, où 1 indique que les deux partitions sont identiques [Ansari et al., 2011]. Il est calculé par l'équation (8) ci-dessous:

$$J = \frac{a}{a + b + c} \quad (8)$$

Où  $a$  est le nombre de paires classées de manière similaire,  $b$  le nombre de paires ayant la même classe mais dans des groupes différents, et  $c$  le nombre de paires appartenant au même groupe et ayant différentes classes.

Bien que ce type d'indices soit simple à calculer, une partition réelle des éléments d'une segmentation n'est pas toujours accessible, en particulier dans un contexte industriel. Dans ce cas, on a recours aux critères internes. Les critères internes s'appuient sur la structure des segments résultants, et évalue la structure des segments avec les données [Aghabozorgi et al., 2015]. Parmi les critères internes les plus utilisés on trouve la somme des erreurs quadratiques (SSE), le C-index, et l'indice de silhouette [Arbelaitz et al., 2013]. La somme des erreurs quadratique est obtenue par le calcul de la somme du carré de la distance entre chaque élément et la moyenne du groupe auquel il appartient. Il renseigne sur la dispersion au sein du groupe [Kim et al., 2017]. Sa mesure est donnée par l'équation (9) ci-dessous [Thinsungnoen et al., 2015].

$$SSE(X) = \sum_{l=1}^K \sum_{x_j \in C_j} \|x_j - m_l\|^2 \quad (9)$$

Où  $x_j$  représente l'élément du groupe  $i$ , et  $m_i$  la moyenne de ce groupe.

Le C-index est un critère mis en place par [Hubert et Levin, 1976]. Il est employé pour mesurer la similarité des groupes, choisir le nombre de groupes à considérer ou comme étant un critère pour l'évaluation de la segmentation [Dimitriadou et al., 2002]. La valeur du C-Index varie entre 0 et 1. La segmentation est meilleure lorsque ce critère est minimal. Le C-Index est donné par la formule de l'équation (10) ci-dessous [Charrad et al., 2014]. Il est obtenu en calculant le rapport entre la somme des distances entre tous les éléments du groupe  $i$ ,  $S_{\min}$  la sommes des distances minimales entre tous les éléments quelques soient leurs groupes, et  $S_{\max}$  la somme des distances maximales entre tous les éléments.

$$C - index = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (10)$$

Dans ce qui suit, nous allons expliquer le principe de calcul de l'indice de silhouette, étant le critère utilisé plus tard dans le mémoire.

L'indice de Silhouette est un indice qui varie entre 0 et 1. Cet indice présente une distance normalisée, calculée par la division de la différence entre la distance moyenne entre tous les points appartenant au même groupe  $b(i)$ , et la distance moyenne du plus proche voisin  $a(i)$ , par le maximum des deux. En d'autres termes, nous calculons le rapport entre l'homogénéité à l'intérieur d'une classe, par rapport à l'hétérogénéité entre les classes. La formule de l'indice de silhouette est donnée par l'équation (11).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (11)$$

Nous avons choisi cet indice pour sa capacité à donner les meilleurs résultats surtout avec la classification hiérarchique [Arbelaitz et al., 2013]. On peut visualiser dans le tableau 2-1, une interprétation de l'indice de Silhouette. Sur ce tableau, on peut lire que lorsque la valeur de cet indice est inférieure à 0,25, ceci signifie qu'aucune structure n'est discernable dans les données. Lorsque la valeur de l'indice de silhouette est comprise entre 0,26 e 0,50, ceci informe qu'il existe une structure, mais que celle-ci est relativement faible. Lorsque l'indice de silhouette obtenu est compris entre 0,51 et 0,70, nous pouvons en conclure qu'il existe une structure raisonnable dans les données. Finalement, lorsque l'indice de silhouette est supérieur à 0,71, ceci indique que les données disposent d'une forte structure.

Tableau 2-1: Tableau de l'interprétation du critère de silhouette tiré de [Cardoso et Carvalho, 2009]

Silhouette	Proposed interpretation
$\leq 0.25$	No substantial structure has been found
0.26–0.50	The structure is weak and could be artificial; please try additional methods on this data set
0.51–0.70	A reasonable structure has been found
0.71–1	A strong structure has been found

Dans leur travail, [Halkidi et al., 2001] critiquent les temps de calcul coûteux des critères externes et internes, et proposent les indices relatifs. Les indices relatifs se basent sur une comparaison de différentes structures des segments.

## **2.4 Conclusion**

Finalement, au cours de ce chapitre nous avons présenté les notions de bases pour notre recherche. Nous avons commencé par présenter l'industrie du vêtement et les enjeux des systèmes de tailles dans le contexte actuel. Ensuite, nous avons présenté les séries temporelles et plus précisément les séries temporelles intermittentes. Finalement, nous avons abordés le concept de segmentation de marché, et les principaux algorithmes rencontrés dans la littérature, les mesures de similarité et les méthodes d'évaluation de la segmentation. La recherche dans la littérature nous a permis de connaître les méthodes de segmentation les plus utilisées, en particulier dans des contextes de recherches similaires à notre projet. Lors de l'application de la méthodologie, nous allons tester la méthode la plus adéquate avec nos données. Nous avons sélectionné le Dynamic Time Warping en tant que métrique de distance pour mesurer la similarité entre les clients. Cette distance est connue par sa flexibilité, et sa pertinence pour la comparaison des séries temporelles. Nous avons également sélectionné l'indice de silhouette pour évaluer les résultats obtenus. Cet indice est caractérisé par sa facilité de mise en œuvre, sa capacité de donner de bons résultats et la disponibilité d'une grille d'interprétation bien détaillée.

## CHAPITRE 3 PROBLÉMATIQUE ET MÉTHODOLOGIE

Au cours de ce chapitre, nous allons présenter le contexte de l'étude, ainsi que la problématique et les objectifs de notre projet. Dans un second lieu, la méthodologie suivie est explicitée.

### 3.1 Contexte de l'étude

Les évolutions faites dans le domaine industriel ont incité les manufacturiers à intégrer les nouvelles technologies dans leurs processus de production, depuis la conception du produit jusqu'à sa livraison chez le client. Dans notre recherche nous nous intéressons plus particulièrement à l'industrie du vêtement. Cette industrie a connu une grande progression avec l'expansion de l'e-commerce ou l'achat en ligne. L'achat en ligne a permis aux fabricants de mieux satisfaire les besoins de leurs clients, et aux clients de bénéficier de meilleurs choix de produits en moins de temps et avec moins d'efforts. L'achat en ligne a aussi permis aux manufacturiers de collecter et stocker des informations relatives aux clients et à recueillir leurs historiques d'achat.

Ainsi, pour pouvoir exploiter les données relatives aux clients, les industriels ont de plus en plus recours à la fouille de données. Il existe en fait plusieurs techniques d'analyse de données permettant l'analyse des marchés et même la gestion des relations clients. L'avantage de ces techniques c'est qu'elles permettent de traiter un grand volume de données et peuvent être appliquées dans divers domaines. Parmi les techniques d'analyse de données les plus utilisées par la communauté scientifique on note la régression, la classification, la segmentation, les réseaux de neurones, etc.

C'est dans ce contexte que se situe notre travail. Notre recherche va se baser sur l'analyse des historiques de commandes effectuées par des clients à travers une plateforme d'achat en ligne. Nous travaillons avec un partenaire de l'industrie du vêtement. Nous nous intéressons particulièrement aux tailles des vêtements commandées par les clients.

Les données collectées sont transformées sous forme de séries chronologiques afin de pouvoir suivre les évolutions des tailles des clients dans le temps. Cette analyse va par la suite permettre de grouper la clientèle dans des ensembles homogènes ayant des comportements similaires.

## **3.2 Problématique**

On cherche dans ce mémoire à utiliser des techniques de fouille des données industrielles afin de comprendre les habitudes d'achat des clients qui achètent des vêtements à travers la plateforme d'achat en ligne qu'offre notre partenaire.

## **3.3 Objectifs**

### **3.3.1 Objectif général**

Notre objectif principal est de développer un algorithme qui permettra de segmenter l'ensemble des clients de notre partenaire afin de comprendre leurs comportements d'achats et discerner la présence de comportements typiques ou atypiques parmi eux. Pour la segmentation, nous allons nous baser sur les tailles des vêtements que ces clients ont commandés à travers un site d'achat en ligne. Nous cherchons par la suite à analyser les groupes obtenus en les corrélant à d'autres facteurs tels que le sexe, les moyennes des tailles.

### **3.3.2 Objectifs spécifiques**

Les objectifs spécifiques reposent sur trois axes principaux qui vont fonder notre méthodologie. Le premier objectif est d'analyser et de formater les données collectées afin qu'elles soient adaptées au contexte de l'étude. Le deuxième objectif est de choisir la distance pour mesurer la similarité des individus, la méthode de segmentation adéquate pour le type de données et le nombre de groupes à retenir et valider le choix à travers des expérimentations. Le dernier objectif est de tester l'applicabilité de la méthode de segmentation retenue sur les données et d'interpréter les résultats de la segmentation en analysant la composition des ensembles créés.

## **3.4 Méthodologie**

Nous disposons des données de l'historique d'achat d'un ensemble de clients. Ces données comportent des informations propres aux clients tels que leurs sexes, leur codes postaux. Nous disposons également des achats qu'ils ont effectués durant une période de six ans. Pour chaque client, une série chronologique de ses achats et plus particulièrement des tailles qu'il a commandées sont ensuite construites. Ces séries nous permettent de construire des profils de variation des tailles



dans le temps et serviront d'entrée pour l'algorithme de segmentation. Sur la figure 3.1 ci-dessous, nous présentons la méthodologie élaborée dans le cadre de ce mémoire.

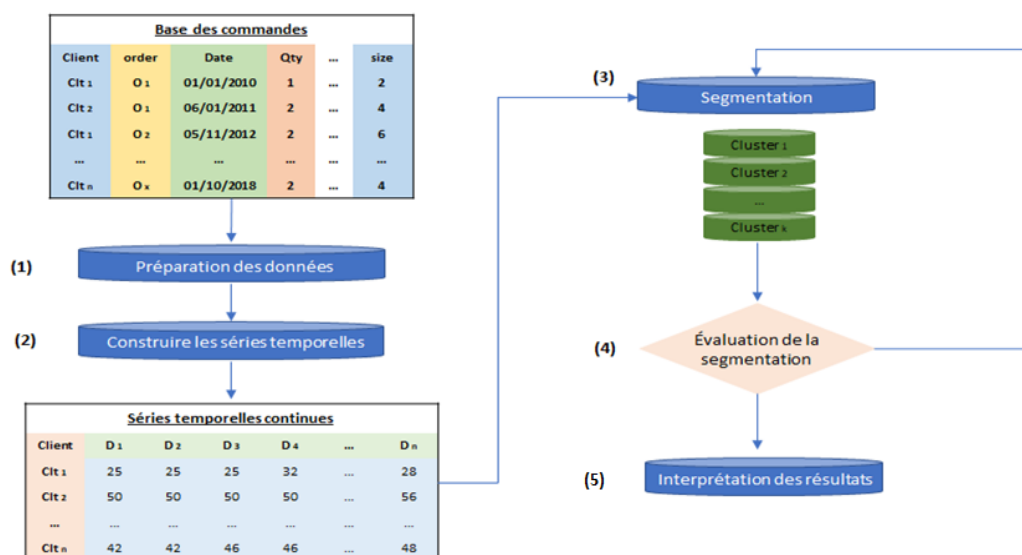


Figure 3.1: Étapes de la méthodologie

Comme on peut le voir sur la figure 3.1, notre travail peut être divisé en cinq étapes principales. (1) La première étape consiste à nettoyer et filtrer la base de données (section 3.3.1). Ensuite, pour pouvoir suivre l'évolution des mensurations (2) on doit construire des séries temporelles, et corriger leur intermittence (section 3.3.2). (3) Les séries temporelles sont ensuite segmentées et les groupes de clients ayant le même comportement sont créés (section 3.3.3). Par la suite (4) nous allons effectuer une évaluation de notre résultat de segmentation, et en s'y basant on peut changer nos paramètres de calcul, jusqu'à obtenir un résultat concluant (sections 3.3.3.2, 3.3.3.3 et 3.3.3.4). (5) La dernière étape consistera à analyser les résultats obtenus et les corrélérer aux autres données propres aux clients (section 3.3.4). Chaque étape de la démarche sera détaillée dans les éléments suivants.

### 3.4.1 Nettoyage de la base de données

Les données collectées dans le contexte industriel ne sont pas toujours utilisables dans leur état brut [Liu et al., 1992]. Souvent une préparation préalable est nécessaire. [Rahm et Do, 2000] consacrent leur article à étudier le nettoyage des données. D'après leur analyse, parmi les problèmes les plus rencontrés avec les données réelles, c'est le bruit qui peut exister dans ces données, la

redondance, les informations manquantes, et les inconsistances, notamment sur l'échelle temporelle.

Les données recueillies par notre partenaire industriel englobent l'historique d'achat de plus de 100 000 clients pour plus d'un million de commandes, et plus de 10 000 types de vêtements différents. Chaque ligne de la base de données est spécifique à une commande individuelle. En effet, chaque commande passée par le client est caractérisée par un numéro, une date, un ensemble d'articles commandés, le nombre de ces articles, des informations sur le client tels que son sexe ou son code postal, et des informations sur les produits achetés. Les informations sur les produits peuvent être de nature qualitative telles que le type du vêtement, ou de nature quantitative telle la taille de celui-ci. En effet, dans notre contexte d'étude, la taille n'est autre que les mensurations relatives de chaque type de vêtement, exprimées en pouces. Ainsi, cette première étape consiste à étudier la base de données dont nous disposons et recenser les vêtements avec lesquelles nous allons conduire notre étude, ainsi que notre population cible. Il s'agit de comprendre chaque attribut, sa signification, et ses différentes valeurs, pour pouvoir dans une seconde étape filtrer la base de données. Ceci revient à éliminer les variables non nécessaires pour l'étude et ne retenir que les individus ciblés par notre analyse. Cette étape a été accompagnée d'un échange avec les experts de l'entreprise afin de valider les décisions prises.

### **3.4.2 Transformation en série temporelle**

La base de données sous sa forme initiale est arrangée de manière que chacune de ses lignes présente un produit acheté, appartenant à une commande et un client déterminé, avec une quantité commandée ou retournée et une date à laquelle l'achat était effectué. Notre objectif, à cette phase, est de partir de cette base et construire un tableau de séries chronologiques qui nous permettra d'observer l'évolution des tailles des clients dans le temps. Cependant, la manière d'observer une taille diffère d'un type de vêtement à un autre, par exemple les tailles des chemises et des pantalons ne se calculent pas de la même manière. C'est pourquoi chaque tableau de série temporelle sera spécifique pour un type de vêtement particulier, incluant toutes les commandes des clients qui l'ont acheté pendant la période d'étude. Par ailleurs, chaque série chronologique est propre à un client et retrace l'ensemble des tailles qu'il a commandé, avec les dates auxquelles il a effectué ses achats.

D'un côté, afin de construire des séries temporelles, il est indispensable de commencer par choisir le niveau de granularité temporelle à laquelle nous désirons observer les données. Définir la

granularité revient à structurer de manière hiérarchique les données tout en choisissant le niveau de détails désiré [Euzenat, 1994]. La granularité dépend de la densité des données et de la continuité de celles-ci [Del Mondo, 2011]. Le choix de la granularité dépend fortement du contexte de l'étude. Elle peut être défini à l'année, au mois, au jour, ou même à la minute dans d'autres cas. Pour notre cas d'étude, nous avons choisi une granularité mensuelle pour les données.

D'un autre côté, étant donné que les séries temporelles construites peuvent être intermittentes, il est essentiel également de rechercher le moyen adéquat pour les agréger. À la fin de cette étape, on vise à avoir des séries temporelles continues et agrégées à la même granularité temporelle.

### **3.4.3 Segmentation des séries temporelles**

Cette étape représente l'axe principal de notre étude. Son but est de tester s'il est possible de diviser l'ensemble des clients retenus dans des groupes homogènes, en se basant sur les séries temporelles traçant l'évolution de leurs tailles. En effet, la segmentation passe par plusieurs étapes. En premier lieu, une métrique de distance adéquate pour notre type de données doit être sélectionnée. C'est avec cette métrique que nous allons évaluer la similarité des individus. Ainsi, en se basant sur cette métrique, une matrice de distance est calculée. Elle englobe les distances entre tous les individus deux à deux. Par la suite, la méthode de segmentation est choisie. Nous avons testé trois des méthodes présentées dans la section précédente (section 1.3.1). La mise en essai de ces différentes alternatives sur un échantillon, à la cinquième étape de la méthodologie (section 3.4.4) nous permettra de choisir la plus performante et la plus adéquate pour les données dont nous disposons. Finalement, la segmentation est réalisée, ayant en entrée la matrice des distances et en sortie des groupes voulus homogènes.

### **3.4.4 Validation de la segmentation**

Cette étape est cruciale et se déroule sur plusieurs échelons. En effet, la première série de tests a lieu au niveau du choix de la méthode de segmentation, tel qu'expliqué dans le paragraphe précédent. Un ensemble de trois méthodes est testé sur un échantillon, choisi de manière que ses individus puissent être groupés intuitivement, par la visualisation. La visualisation désigne le fait de classer les individus en nous basant sur la forme générale de la courbe d'évolution de sa taille. Les performances de ces algorithmes, en termes d'exactitude du regroupement, de rapidité, et de

variabilité entre les groupes créés sont comparées. En conséquence, la méthode la plus efficace est retenue pour la suite des segmentations.

La deuxième validation est effectuée après l'application de la segmentation sur l'ensemble des données. Les performances de l'algorithme retenu et la qualité des groupes générés sont évaluées. Pour ceci, un critère d'évaluation adéquat est adopté. Il existe dans la littérature plusieurs critères d'évaluation. L'homogénéité entre les membres d'une classe et l'hétérogénéité entre les classes sont parmi les critères les plus utilisés.

### **3.4.5 Interprétation des résultats**

Pour donner suite à la validation des résultats de la segmentation, nous passons à l'interprétation des résultats. Le but de cette étape est de comprendre les structures des groupes créés. La composition de chacun des groupes est analysée, en se basant sur des informations propres aux clients telles que le sexe, le code postal, les produits les plus achetés, l'ancienneté chez le partenaire, etc. Cette analyse nous permettra d'appréhender les caractéristiques communes des clients groupés ensemble et de comparer les différents groupes. Aussi, nous allons observer de quelle manière les tailles évoluent pour chaque segment de clients. Ceci peut nous guider à élaborer des règles visant la prédiction de la demande future.

## **3.5 Conclusion**

Ce chapitre nous a permis de présenter le contexte d'étude, la problématique et les objectifs de notre travail. Nous avons par la suite introduit la méthodologie suivie. L'implémentation de ces étapes va être illustrée au cours du chapitre suivant. Les résultats obtenus au cours de ce projet permettront à l'entreprise partenaire de comprendre les habitudes d'achat de ses clients en termes de tailles de vêtements commandés et aussi de voir l'évolution des tailles commandés. Les résultats de la segmentation permettront également à notre partenaire d'appréhender si ses clients ont des comportements similaires et peuvent former des groupes homogènes.

## CHAPITRE 4 CAS D'ÉTUDE

Au cours de ce chapitre, nous allons commencer par présenter notre partenaire industriel. Ensuite, nous allons introduire quelques statistiques descriptives relatives aux données utilisées à travers ce projet. Finalement, nous allons illustrer l'application de la méthodologie exposée au chapitre précédent dans le contexte pratique de l'étude et exhiber les résultats obtenus, ainsi que leur analyse.

Pour des raisons de confidentialité, les données réelles sont masquées dans ce mémoire.

### 4.1 Présentation du partenaire industriel

Notre partenaire assure la prestation de solutions pour la personnalisation, la conception et la production de tous les aspects de la sous-traitance d'un programme d'uniformes. L'entreprise habille des centaines de clients dans le monde, majoritairement au Canada, mais aussi en Europe, Australie, Asie, Afrique du Nord et au Moyen-Orient.

Notre partenaire opère de manière contractuelle. Chaque contrat ne concerne qu'un seul métier ou secteur d'activité. Les employés des entreprises pris en charge ne peuvent avoir leur uniforme qu'à travers notre partenaire qui met à leur disposition une plateforme en ligne à partir de laquelle chaque individu peut commander des parties d'équipement de son uniforme, durant toute l'année, grâce à un identifiant unique et un budget basé sur un système de points. Les points sont attribués selon le rôle de l'employé dans l'entreprise. Notre partenaire se charge également de la livraison des pièces d'uniformes jusqu'aux employés des entreprises clientes, par la sous-traitance d'un service de transport.

Dans le cadre de ce projet, nous limitons notre étude à un programme d'uniforme existant, avec un groupe de clients spécifique. Le groupe de clients en question est constitué de plus de cent mille individus. Le partenaire dispose d'archives importantes sur les historiques d'achats effectués par chaque individu sur une longue période de temps. Chaque client est obligé de s'équiper auprès de notre partenaire pour chaque pièce de son uniforme, nous disposons donc des données complètes de chacun.

La section suivante montrera quelques analyses statistiques sur les clients et les produits considérés. Toutes les analyses ont été effectuées avec le logiciel de programmation R.

## 4.2 Statistiques sur les données

La base de données fournie par l'entreprise inclut les commandes de tous ses clients appartenant à un contrat spécifique. Les données représentent leurs historiques d'achat entre les années 2012 et 2019. À chaque achat dans le système en ligne, les transactions sont archivées et stockées. Le suivi des achats effectués par chaque individu est donc possible. Nous disposons de plusieurs millions de transactions au total pour plusieurs milliers d'articles différents et plus de milles familles de produits. Nous allons étudier les commandes de plus que cent milles clients, dont environ 20% de femmes et 80% d'hommes.

Le nombre de clients effectuant des commandes est relativement stable durant la période de l'étude. Comme on peut le voir sur la figure 4.1, le nombre total de clients ne varie que faiblement d'une année à l'autre. Sur la figure, l'axe des ordonnées a été masqué pour des raisons de confidentialité.

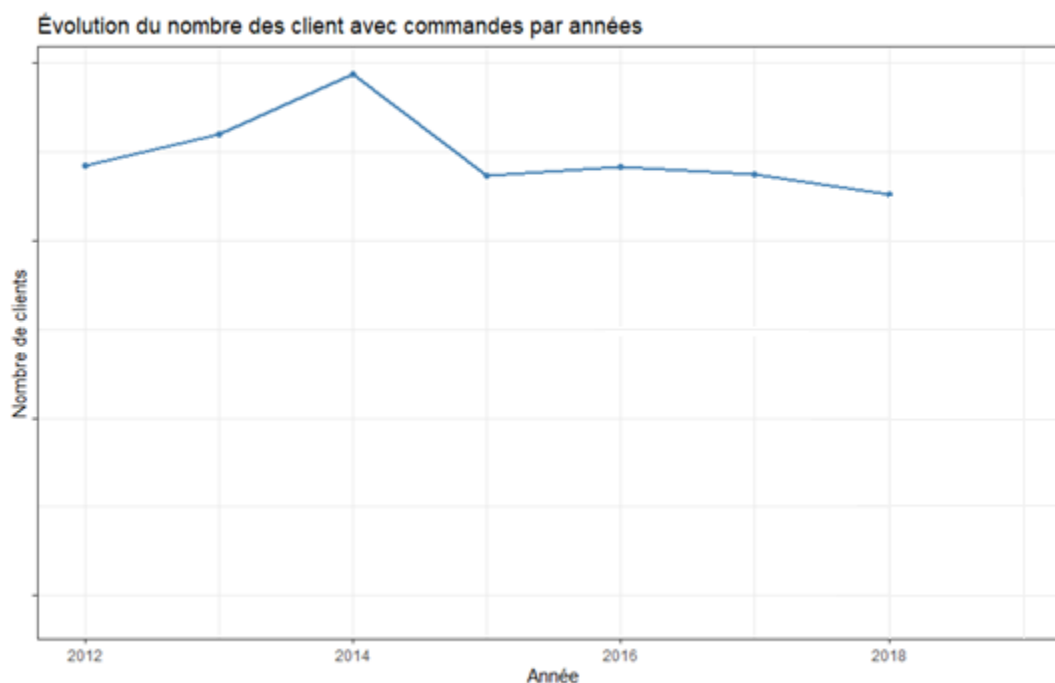


Figure 4.1: Évolution nombre de clients ayant passé des commandes

La base de données inclut au total plus de mille articles différents appartenant à vingt familles de produits, formant les uniformes des clients. Ces articles peuvent être des vêtements comme des chemises, des jupes et des pantalons, des chaussures, et même des chapeaux, des badges et des insignes. Sur la figure 4.2, nous pouvons voir tous les articles disponibles et les quantités totales achetées durant la période d'étude. Les trois articles les plus achetés sont les bas, les insignes métalliques et finalement les chemises. Sur la figure 4.2, on peut observer qu'il existe aussi des vêtements d'extérieurs, des cravates, et des ceintures. Les produits commandés peuvent être des vêtements d'hiver (11,6% produits), d'été (9,7% produits) ou pouvant être portés aux deux saisons (78,7% produits). Sur cette figure également, l'axe des ordonnées a été masqué pour des raisons de confidentialité.

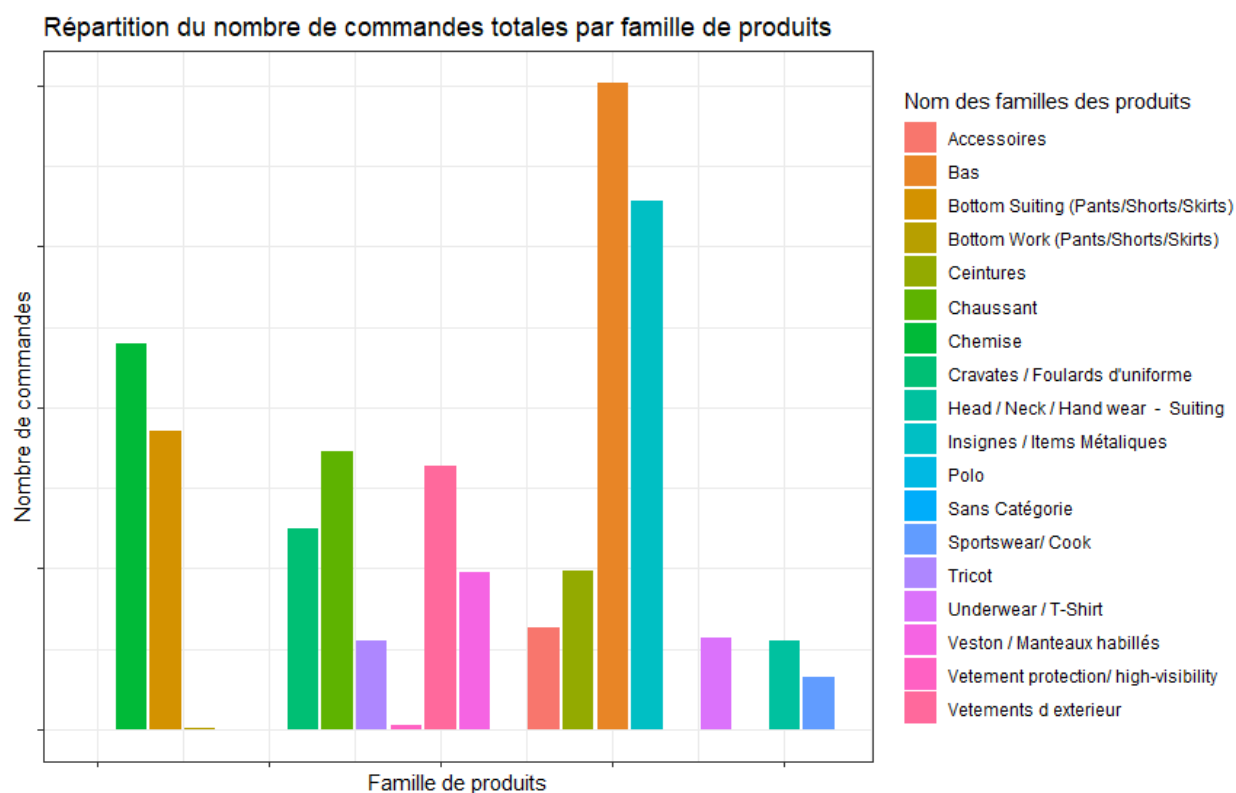


Figure 4.2: Répartition des produits sur les commandes des clients

De plus, la base de données contient deux catégories d'articles, ceux qui ont une taille et ceux qui n'en possède pas. En ce qui concerne la première catégorie, la taille peut être composée d'une à quatre mesures. Les mesures sont exprimées en pouces. Chaque mesure varie en moyenne sur un

intervalle de 40 pouces entre la valeur minimale et la maximale. Notre partenaire livre aussi des vêtements avec des tailles spécifiques pour les clients qui ont des besoins particuliers. Au total les données contiennent 28 commandes avec des tailles spécifiques, dont principalement des vêtements d'extérieur.

L'étude des retours a révélé que les articles achetés sont retournés avec un pourcentage de 1,05%. Le nombre de retours reste stable durant les années, à l'exception de l'année 2016 où les retours ont augmenté. Les experts de l'entreprise expliquent ceci par le lancement d'une nouvelle gamme de produits, ce qui pourrait expliquer la croissance du taux des retours durant cette période.

L'analyse a montré que les femmes ont retourné 4,3% des produits qu'elles ont acheté, alors que les hommes ont retourné 2,05% de leurs produits achetés. Sur la figure 4.3, nous pouvons contempler le nombre des retours totaux pour chaque type de vêtement. Les trois articles les plus retournés sont les pantalons/shorts/jupes, les manteaux habillés et les chemises.

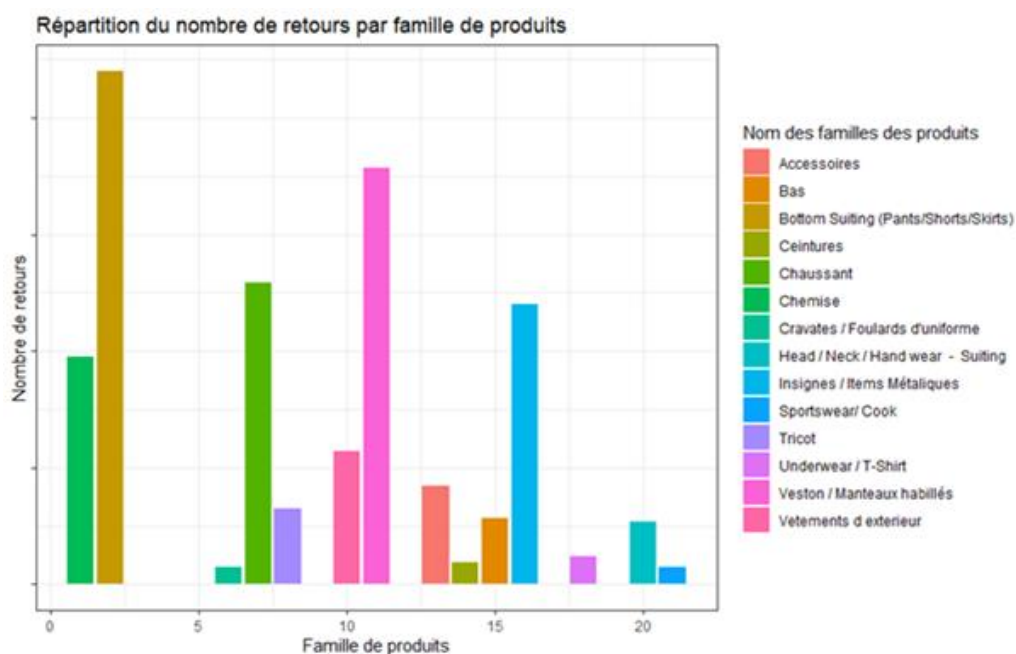


Figure 4.3: Répartition des retours par type de produits.



## 4.3 Application de la méthodologie

### 4.3.1 Préparation des données

La première étape de notre méthodologie repose sur le nettoyage et la préparation des données. Nous disposons de l'historique de commandes des clients pendant une durée d'étude de six années. Chaque ligne de cette base de données est spécifique à un produit acheté par un client donné. Chaque commande est caractérisée par un numéro, une date, et le nombre d'articles qu'elle inclue. Elle inclut en plus, des informations sur le client tels que son sexe, son code postal et son rôle au sein de l'entreprise. Finalement, l'historique des commandes contient des informations sur chaque produit acheté, tels que son type et sa description.

En plus des données de l'historique des commandes, nous disposons de la base de données incluant la description des produits disponibles pour ce contrat et une troisième base de données contenant les descriptions des tailles de ces produits. Comme expliqué plus tôt, au cours de ce projet, on désigne par taille l'ensemble des mensurations caractéristiques de chaque type de vêtement, exprimées en pouces.

Le premier palier est donc de joindre ces bases de données dans un seul fichier, pour avoir une meilleure visibilité sur les commandes et les produits achetés. Par la suite, nous avons analysé chacune des caractéristiques, pour être en mesure de sélectionner et ne garder que celles qui seront utiles pour nos analyses. Aussi, les commandes ayant des informations incomplètes ou manquantes ont été éliminées de la base de données.

Vu le fait que nous étudions des données relatives à l'achat d'uniformes, les commandes dont nous disposons contiennent tous les éléments nécessaires pour former un uniforme. Par conséquent, on peut trouver des commandes de pantalons, de chemises et des manteaux, mais aussi des chaussettes, des cravates et des badges. Nous cherchons à analyser les données morphologiques de nos clients, donc seuls les produits sur lesquels on peut suivre des mensurations ont été retenus. Tous les autres produits ont été écartés de la base de données.

L'analyse des données montre la présence de transactions avec de grandes quantités d'articles, pouvant atteindre 20 000 articles pour un seul achat. Ces transactions concernent des commandes de groupes, pour subvenir aux demandes de plusieurs personnes à la fois. Elles sont généralement basées sur une anticipation et une prévision des besoins futurs plutôt que de couvrir un besoin réel

d'un client. Pour cette étude nous avons besoin d'étudier les achats individuels de chacun des clients pour pouvoir suivre l'évolution de leurs mensurations dans le temps. Ce type de commandes de groupe a donc été exclu de la base.

En analysant le nombre de commandes moyen par client pour un produit donné, on trouve qu'en moyenne 75% des clients achètent sept fois ou plus au cours de la période d'étude. Cependant, quelques clients ont effectué moins de 3 commandes. En cherchant à travailler avec des données les plus précises possible, un seuil minimal de commandes a été fixé pour chaque type de produits. Pour chacun des produits considérés, les clients qui ont un nombre de commandes inférieur à ce seuil seront écartés de l'analyse. Ces individus ne nous permettent pas d'étudier une évolution réelle ou une tendance à travers le temps, par manque d'information.

La figure 4.4 ci-dessous présente un exemple de la base de données obtenue. Nous pouvons constater qu'elle renferme les informations sur les commandes, sur les clients ainsi que sur les produits commandés.

ORDER_NUM	SPECIFIC_NSN	LINE_NUMBER	QTY_ORDERED	QTY_SHIPPED	QTY_RETURNED	ENTRY_DATE	POSTAL_CODE	CLIENT_CODE	CLIENT_SE X	DESCRIPTION
1400	8405A	1	2	2	0	10/07/2012	B3Z ...	111	M	Chemise
1401	8405B	2	2	2	0	10/07/2012	B3Z ...	222	M	Bottom Suiting
1402	8405C	4	2	2	0	10/07/2014	B3Z ...	333	M	Bas
1403	8405D	6	1	1	0	10/07/2015	B3Z ...	444	M	Ceintures
1404	8405E	1	1	1	0	20/08/2017	B3M ...	555	F	Chaussant

Figure 4.4: Extrait de la base des données traitée

À la fin de cette première étape, les données obtenues jusqu'ici sont filtrées, préparées et centrées sur le besoin que nous avons. Seuls les attributs nécessaires sont retenus. La prochaine étape serait de partir de cette base de données pour construire les séries temporelles.

### 4.3.2 Construction des séries temporelles

Cette étape vise à créer des séries temporelles à partir des données de commandes. À cette fin, des tableaux de séries temporelles pour chacun des vêtements retenus vont être préparés, afin de pouvoir comparer l'évolution des mensurations des clients.

Chaque transaction faite par un client est caractérisée principalement par une date, et des mensurations dont le nombre peut varier entre un et quatre. L'un des problèmes rencontrés est le fait que les produits appartenant à la même famille n'ont pas tous le même nombre de mensurations

représentatives. Par exemple, un type de chemise peut avoir comme taille des mesures de tour de poitrine et la longueur des manches alors qu'un deuxième type de chemise peut avoir en plus la mesure de tour de col. Par conséquent, deux stratégies sont à considérer. La première consiste à étudier chaque type spécifique de vêtement seul, alors que la deuxième se base sur le choix d'une ou de deux mensurations significatives pour chaque famille de produits.

Aussi, les chemises commandées par les clients peuvent être des chemises d'hiver, d'été ou pour les deux saisons (bi-saison). Sur la figure 4.5, nous pouvons voir la répartition du nombre de commandes des chemises, en fonction de la saison. Nous pouvons remarquer que les chemises d'été et d'hiver ont un nombre égal de commandes, et plus important que les chemises bi-saison. Pour ce qui suit, nous allons effectuer la segmentation en considérant toutes les chemises ensemble.

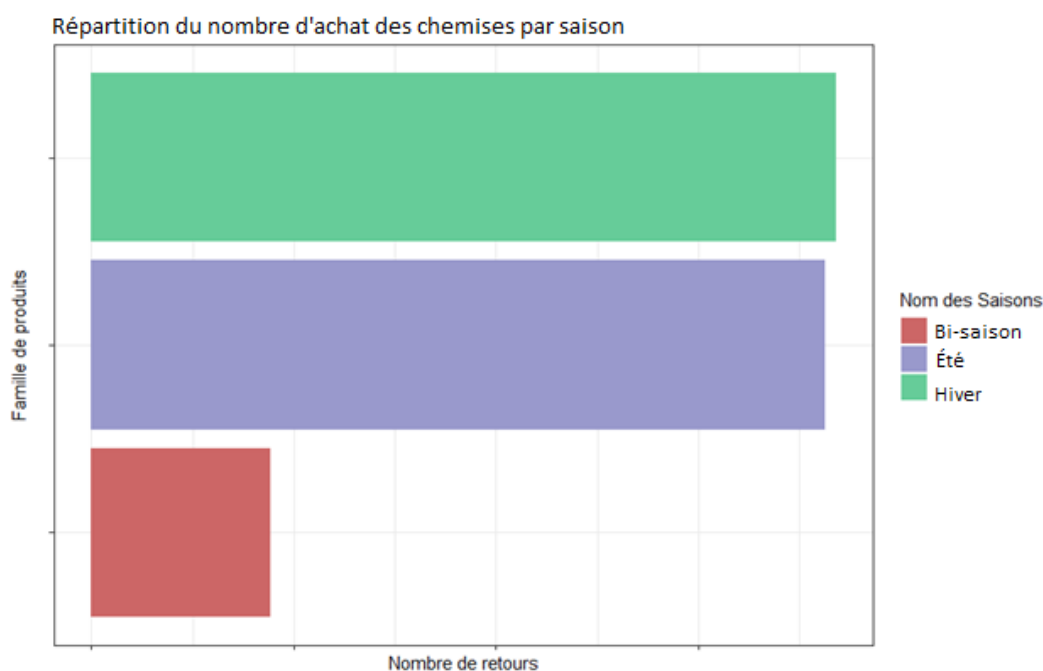


Figure 4.5: Répartition du nombre de commandes des chemises par saison

Après la mise en essai de la première stratégie, nous nous sommes retrouvés avec un grand nombre de sous-groupes à considérer, et un petit nombre des clients dans chacun de ces ensembles, ce qui ne nous permettra pas d'avoir des résultats représentatifs. Par voie de conséquence, nous avons décidé d'élire pour chaque produit une mensuration caractéristique. Par exemple, pour les chemises

on peut considérer le tour de buste, alors que pour les pantalons on choisit le tour de taille ou le tour des hanches. De la sorte, pour chaque client, et pour chacune des mensurations on élabore une série temporelle basée sur l'ensemble des produits compatibles avec cette mesure, durant les six années d'études. Sur la figure 4.6, on peut voir la transformation des données en séries temporelles. Chaque client est représenté par une ligne dans la base des séries temporelles.

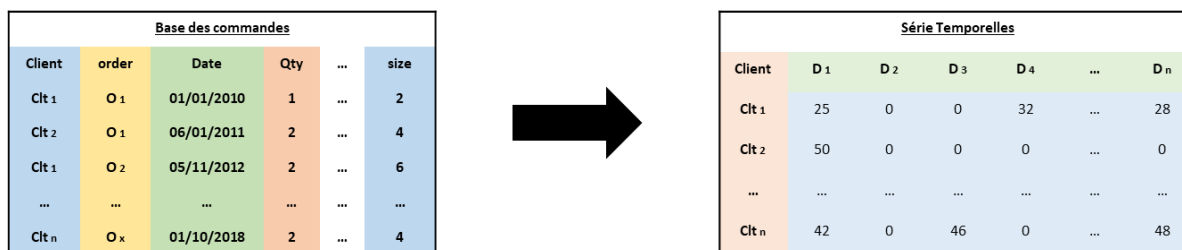


Figure 4.6: Transformation des données en séries temporelles

Étant considéré que chaque commande est effectuée à une date précise, les séries chronologiques ont une granularité quotidienne. Une telle temporalité quotidienne nous laisse avec des données fortement intermittentes. Nous rappelons que des données intermittentes sont caractérisées par plusieurs périodes de demandes nulles. Pour pallier ce problème, il est convenu de considérer une granularité temporelle moins fine, telle qu'au niveau mensuel. Les séries temporelles sont donc agrégées au niveau mensuel. Si le client achète le même article plusieurs fois le même mois, la mensuration retenue pour ce mois sera la moyenne de l'ensemble des mensurations. Bien que l'occurrence de cette situation soit rare, on peut l'expliquer par le fait qu'un client non satisfait de la taille du vêtement qu'il a acheté va retourner ce dernier, et en acheter un autre de taille plus convenable.

Néanmoins, même après cette première agrégation, les séries temporelles contiennent toujours plusieurs zéros, et il est nécessaire de les lisser. Le défi est que travailler avec des mensurations de tailles est un peu différent de la manipulation des données quantitatives, comme des quantités commandées par exemple. Avec ce type de données, un lissage par calcul de moyenne ou par la méthode de Croston [Croston, 1972] n'est pas possible. En effet, afin de corriger l'intermittence des données, nous avons mis en place un principe simple et particulier à notre cas d'étude : en commençant par la commande initiale de chaque usager, et tant que celui-ci n'effectue pas une

nouvelle commande, on considère que sa taille reste constante, et égale à la dernière taille qu'il a choisie. Sur la figure 4.7, on peut voir que les données du client de code 3 avait commandé une chemise avec une mesure de 45 pouces, ensuite avec une mesure de 46 pouces, et avec une mesure de 49 pouces à la fin de la période de l'étude. Les reste des mois où il n'a pas effectué d'achat, ses mesures sont nulles. Après le lissage, sur la figure 4.7, les données des clients deviennent continues, et ne représentent plus de valeurs nulles.

Code client	01/01/2012	01/02/2012	01/03/2012	01/04/2012	...	01/05/2018	01/06/2018	01/07/2018	01/08/2018
1	0	42	0	0	...	44.5	0	0	0
3	45	0	0	46	...	0	49	0	0
5	45	0	45	0	...	45	0	0	0
7	0	0	47	0	...	0	0	51	
9	46.5	0	44	0	...	0	0	0	46.5

↓

Code client	01/01/2012	01/02/2012	01/03/2012	01/04/2012	...	01/05/2018	01/06/2018	01/07/2018	01/08/2018
1	42	42	42	42	...	44.5	44.5	44.5	44.5
3	45	45	45	46	...	46	49	49	49
5	45	45	45	45	...	45	45	45	45
7	47	47	47	47	...	47	47	51	51
9	46.5	46.5	44	44	...	46.5	46.5	46.5	46.5

Figure 4.7: Transformation en série temporelles continues

Un autre phénomène observé dans les données est le fait que quelques clients présentent parfois des valeurs aberrantes. Ce qu'on désigne par valeur aberrante est une augmentation ou diminution de la mesure de taille de manière significative par rapport aux valeurs qui précèdent et qui suivent. Un exemple de ces valeurs est présenté sur la figure 4.8 (A). On peut observer sur la figure 4.8 (A) que le client présente une courbe d'évolution continue et qui augmente par échelon. La seule valeur aberrante est une mesure appartenant à une commande au cours de l'année 2013 pour laquelle la mesure du client a augmenté en pic pour une commande individuelle.

Pour remédier à un tel problème, plusieurs méthodes de filtrage existent, notamment le filtrage par médiane. Sur la figure 4.8 (B), on peut voir qu'initialement le signe présenté avait deux valeurs aberrantes. Par l'application du filtre par médiane ces deux valeurs ont été corrigées, permettant d'obtenir un signal plus lisse.

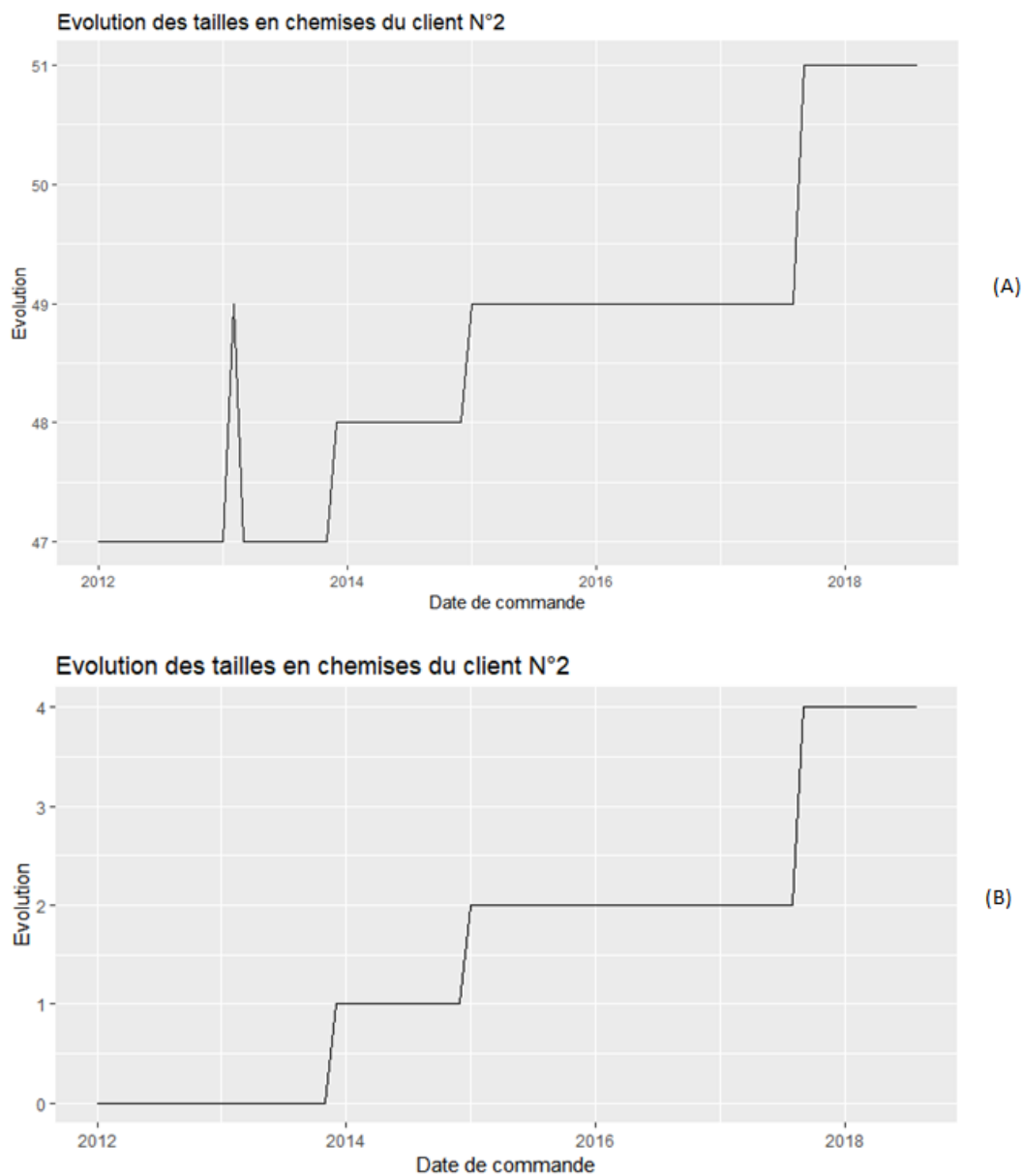


Figure 4.8 : Correction des valeurs aberrantes par le filtre médian

Pour notre cas d'études, nous allons réaliser nos recherches avec le produit le plus commandé parmi les clients, qui est la chemise. Dans ce qui suit nous considérons toute la famille des chemises, et nous ne faisons pas de différence entre les chemises à manches longues ou courtes, ou ayant différentes couleurs. Quant au choix de la mensuration caractéristique, nous avons trouvé que le

tour du buste est la mesure la plus représentative pour ce type de vêtement et en se basant sur cette mesure, les séries temporelles ont été construites.

### **4.3.3 Segmentation**

Au fil des étapes précédentes, notre objectif était de préparer les données et les séries temporelles pour la phase clé de notre étude, qui est la segmentation. Dans notre cas, segmenter revient à diviser les clients en plus petits groupes ayant des comportements similaires, en se basant sur l'évolution de leurs données morphologiques.

Dans cette partie, nous allons suivre les étapes d'une segmentation présentées dans la figure 2.7 dans le chapitre 2. Le premier pas étant de cerner les variables explicatives, ceci a eu lieu au cours des deux précédentes phases de la méthodologie. Nous allons considérer les chemises, et principalement la mensuration du tour de buste. La deuxième étape consiste à choisir la métrique de distance avec laquelle nous allons mesurer la similarité entre les individus et les paramètres de l'algorithme de segmentation. Plusieurs essais de segmentation vont avoir lieu. Nous allons effectuer des tests en considérant respectivement le nombre de groupes adéquat et la nature des séries temporelles. La qualité des groupes obtenus est validée à chaque série de tests. Finalement, les résultats sont interprétés.

#### **4.3.3.1 Choix de la métrique de distance**

Durant le processus du choix de la distance à utiliser pour comparer les séries temporelles, nous avons trouvé que le Dynamic Time Warping est la métrique la plus adaptée à notre cas d'étude. À l'encontre de la distance Euclidienne, cette distance permet de comparer les allures générales des séries temporelles et ne prend pas en considération les décalages ou la différence d'amplitudes. Donc pour comparer les allures des courbes d'évolution des données morphologiques, le DTW nous permet d'avoir les meilleurs résultats. Pour cette métrique, la fenêtre temporelle appropriée doit être définie. En effet, en considérant deux séries temporelles A et B, la fenêtre temporelle détermine le nombre de points de la série B qui vont être comparés à chaque point de la série A, afin de trouver un chemin optimal entre les deux séries [Gaudin et Nicoloyannis, 2005]. Pour le choix de la fenêtre temporelle, nous avons testé plusieurs valeurs. Nous avons constaté qu'à partir de  $n=2$ , il n'y avait plus d'amélioration pour le calcul de la matrice de distance. Nous avons donc fixé notre fenêtre temporelle à  $n=2$ . Par la suite, nous avons pu construire la matrice de distance,

qui contient les distances entre tous les clients deux à deux. Sur la figure 4.9, on peut voir un exemple de cette matrice sur un échantillon de dix clients. Comme l'illustre la figure 4.9, il s'agit d'une matrice carrée symétrique, à diagonale nulle. Pour chaque couple de clients, la distance entre leurs deux séries temporelles est mesurée. Ces matrices de similarité vont servir d'entrée pour l'algorithme de segmentation.

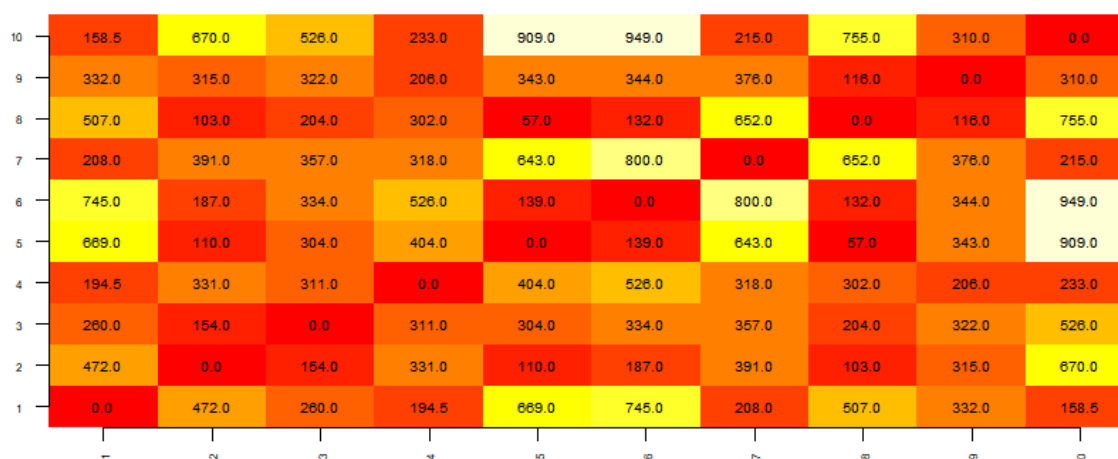


Figure 4.9 : Matrice de distance sur un échantillon de 10 clients

#### 4.3.3.2 Choix de l'algorithme de segmentation

Comme expliqué dans le chapitre précédent, nous avons sélectionné cinq alternatives à tester pour la segmentation de nos séries temporelles. Ceci inclut la méthode de segmentation partitionnelle «Shape Based», la méthode TADPOLE, et la méthode de segmentation hiérarchique agglomérative avec trois différentes distances, notamment la distance WARD, la distance «individuel» et «complete». Pour comparer les performances de ces méthodes, nous allons effectuer une évaluation de ces méthodes, en nous basant sur un critère externe, soit la pureté. Nous avons donc choisi un petit échantillon aléatoire de clients que nous pouvons classer manuellement. La segmentation par chacune de ces méthodes a été appliquée. Les résultats des cinq méthodes ont été comparés sur la base de l'exactitude de la segmentation et le temps de calcul. Les résultats de la comparaison ont montré que les trois alternatives de la méthode hiérarchique agglomérative ont abouti au même résultat semblable au regroupement de référence. Les autres méthodes ont également donné des groupes similaires, hormis quelques séries qui étaient classées différemment.



Par conséquent, nous avons décidé de poursuivre nos travaux avec une segmentation hiérarchique, avec la distance WARD, puisque c'est la distance la plus utilisée dans la littérature.

#### **4.3.3.3 Essais sur la nature des séries temporelles**

Dans cette partie, nous allons appliquer l'algorithme de segmentation par une classification hiérarchique agglomérative, et en se basant sur la distance DTW, sur deux différentes variantes de nos séries temporelles. En premier lieu, nous allons comparer les séries temporelles telles qu'elles étaient construites aux étapes précédentes. Ces séries temporelles représentent l'évolution des mensurations du tour de buste des clients dans le temps. Par la suite, nous allons transformer les séries temporelles de façon qu'elles représentent les variations dans le temps par rapport à la mesure initiale du client. Pour obtenir ce deuxième type de données, nous considérons la première taille commandée comme la référence. En avançant dans le temps, les tableaux sont remplis en calculant la différence par rapport à cette référence. Nous obtenons donc des séries temporelles de l'évolution des tailles. Si le client ne change pas de tailles, sa courbe d'évolution sera constante. Dans le cas où sa taille augmente, la courbe sera croissante, ayant en amplitude la différence entre les deux tailles commandées.

La segmentation par la méthode hiérarchique agglomérative a été appliquée sur les deux tableaux de séries temporelles. À priori, nous avons fixé le nombre de groupes recherché à  $k=10$ . Les résultats des segmentations sont donnés par les figures 4.10 et 4.11. La figure 4.10 correspond à la segmentation avec les données initiales, et la figure 4.11 avec les données transformées. Les deux figures présentent l'évolution des mesures des clients (l'axe des Y) dans le temps (l'axe des X).

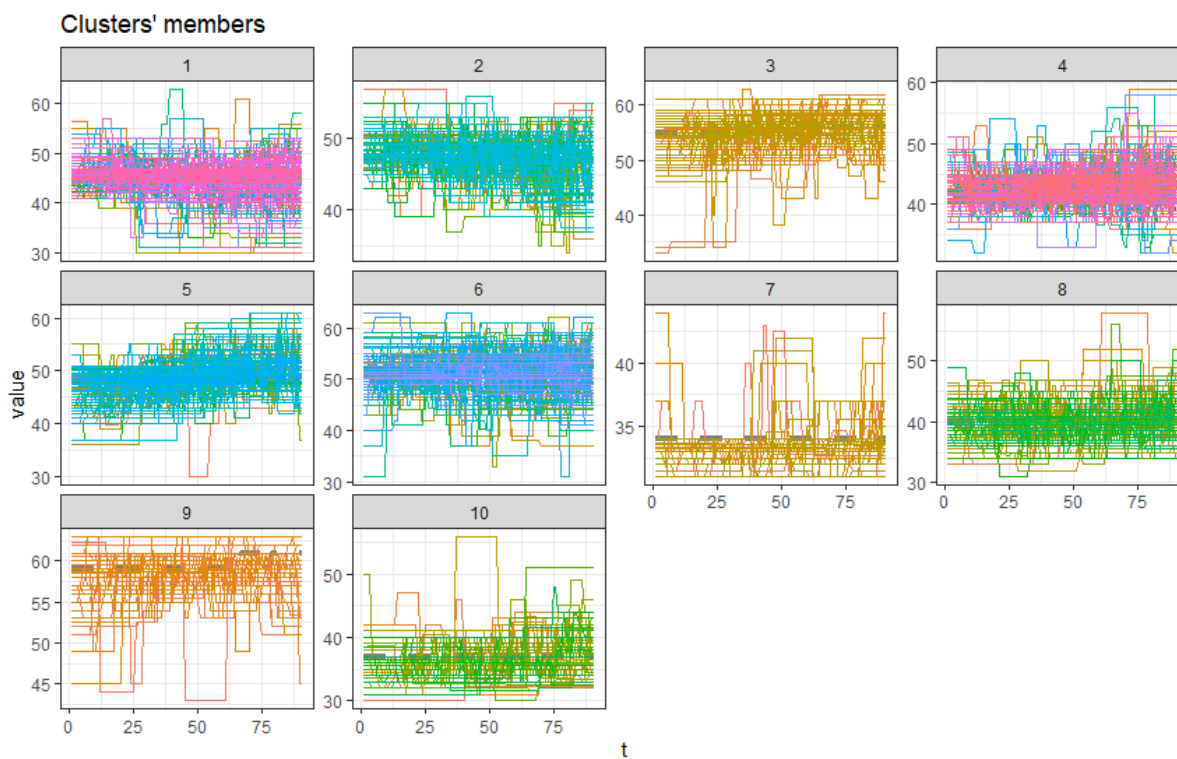


Figure 4.10 : Résultats de la segmentation du premier type de données en 10 groupes

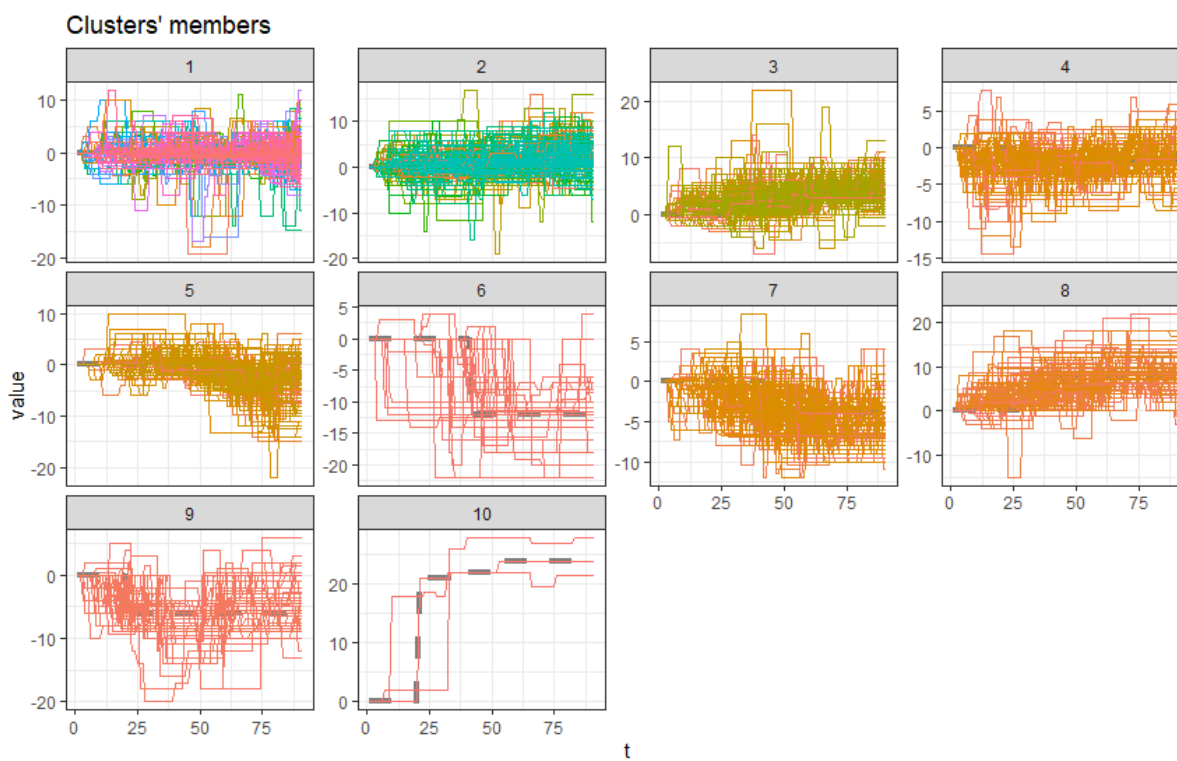


Figure 4.11 : Résultats de la segmentation du deuxième type de données en 10 groupes

Sur la figure 4.10, les courbes de mensurations varient, de façon générale, entre un minimum de 30 et un maximum de 60. Dans la figure 4.11, les courbes varient entre une valeur minimum de -20 et une valeur maximale de 20. Les valeurs négatives dans ce cas reflètent une diminution de taille.

Ce que nous remarquons à partir de ces deux figures, c'est le fait que la transformation des séries temporelles nous a permis d'améliorer le résultat observé. Aussi, les amplitudes des intervalles de variation sont comparables, mais nous remarquons que les tendances en augmentation et en diminution sont plus observables sur la figure 4.11. En raison de l'amélioration des résultats par la transformation des séries temporelles, nous décidons de conserver les séries transformées pour le reste de l'étude.

Pour les essais de l'étape précédente, nous avons fixé un nombre de groupe  $k=10$  afin de comparer les deux types de données. À l'étape suivante, nous effectuons des tests afin de sélectionner le nombre optimal de groupes.

#### **4.3.3.4 Essais sur le nombre de groupes**

Il existe plusieurs tests servant à fixer le nombre de groupes à considérer pour une segmentation. Pour notre étude, nous allons effectuer trois tests. Nous allons explorer le dendrogramme qui montre la disposition des groupes dans un arrangement hiérarchique. Nous allons aussi examiner les courbes de perte d'inertie et la courbe de variation de l'indice de silhouette en fonction du nombre de groupes.

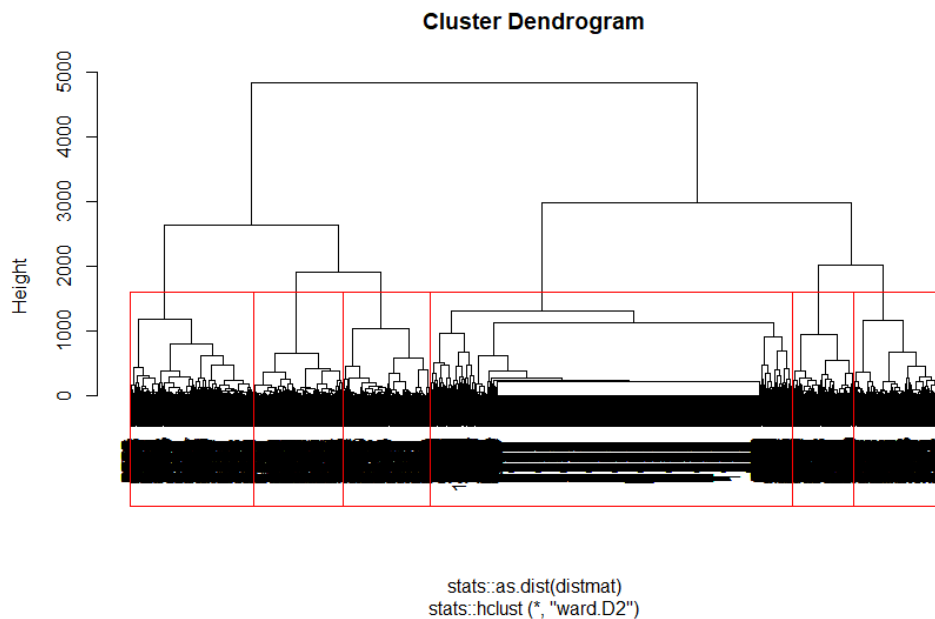


Figure 4.12 : Dendrogramme de la classification hiérarchique

La figure 4.12 représente le dendrogramme obtenu pour la segmentation des séries temporelles. L'arbre présenté commence par un nombre de classes égal au nombre total des clients. Avec chaque itération, les individus les plus proches sont rassemblés, jusqu'à finir avec un seul groupe contenant la totalité des individus.

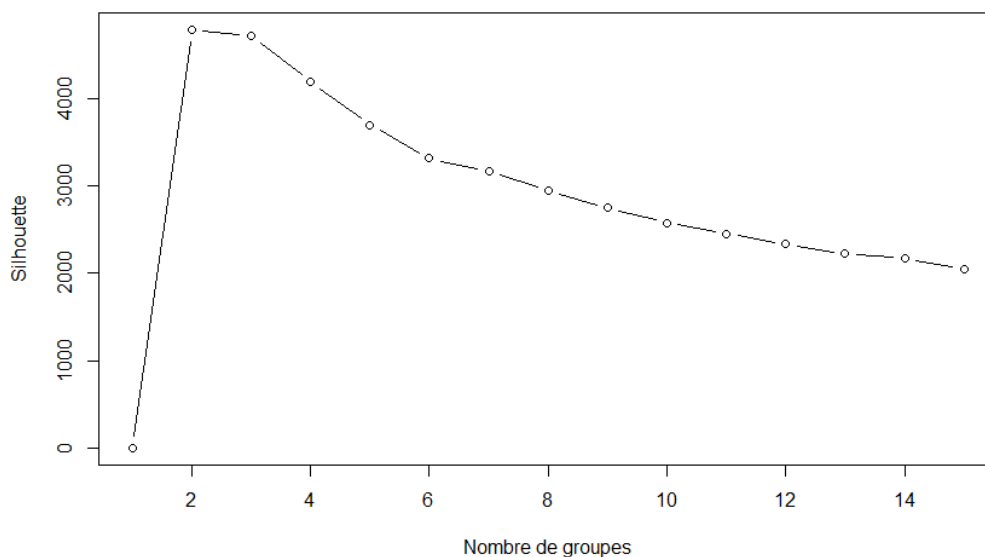


Figure 4.13 : Courbe d'évolution de l'indice de silhouette en fonction du nombre de groupes

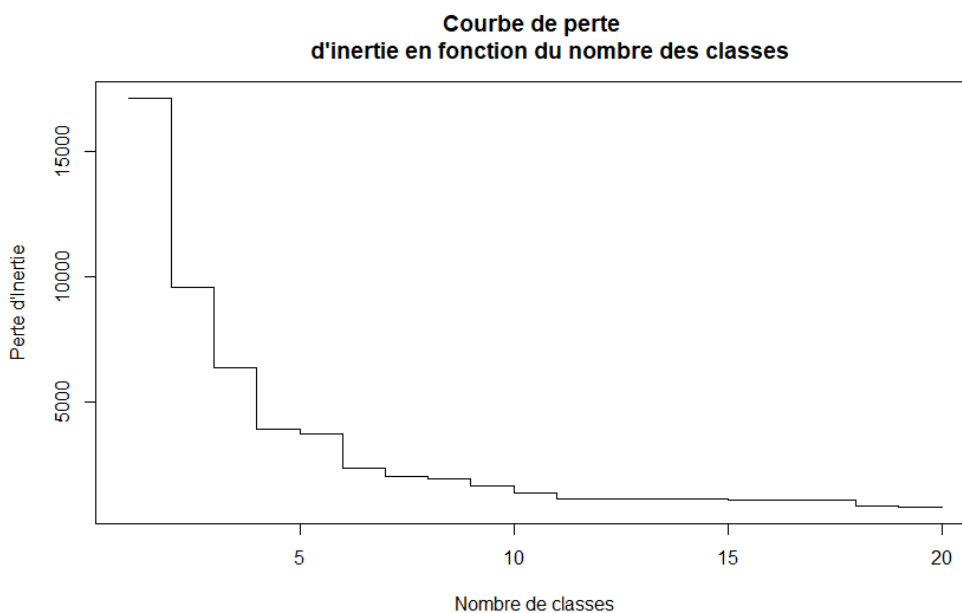


Figure 4.14 : Courbe de la perte d'inertie en fonction du nombre de groupes

Les deux schémas des figures 4.13 et 4.14 nous permettent également de lire le nombre optimal de groupes à choisir. D'une manière semblable, les courbes de décroissance d'inertie et l'indice de silhouette nous permettent d'élire le nombre de segments en choisissant le point avec une grande

penne, et à partir duquel la courbe ne varie que légèrement. D'après les figures 4.12, 4.13 et 4.14, nous observons que le nombre de groupe à établir est égal à six.

Parmi les caractéristiques d'une classification hiérarchique, le fait que le nombre de groupes n'est pas défini à l'avance. Les courbes de variations d'inertie, de variations de l'indice de silhouette et le dendrogramme suggèrent que le nombre de groupes à choisir est  $k=6$ . En nous basant sur des résultats préliminaires, nous décidons de considérer en plus un nombre plus grand de groupes pour pouvoir comparer les deux résultats. Après quelques tests, et des discussions avec les spécialistes de notre partenaire, nous avons décidé de tester la segmentation avec un nombre de segments égal à  $k=10$ .

Nous avons testé de normaliser les données avec une norme Z. La normalisation par la norme Z est une transformation linéaire qui consiste à transformer les données de façon à avoir une moyenne nulle et un écart type égal 1. Pour l'ensemble des données, la norme Z est calculée par la soustraction de la moyenne et la division par l'écart type comme le montre l'équation (12) ci-dessous. Cette transformation ne change pas la distribution des données. Elle sert à faciliter leur traitement et leur interprétation.

$$Z(x_i) = \frac{x_i - \bar{x}}{s_x} \quad (12)$$

Les résultats des deux segmentations des données normalisées en considérant  $k=6$  et  $k=10$  sont effectués. Les résultats sont présentés par les figures 4.15 et 4.16.

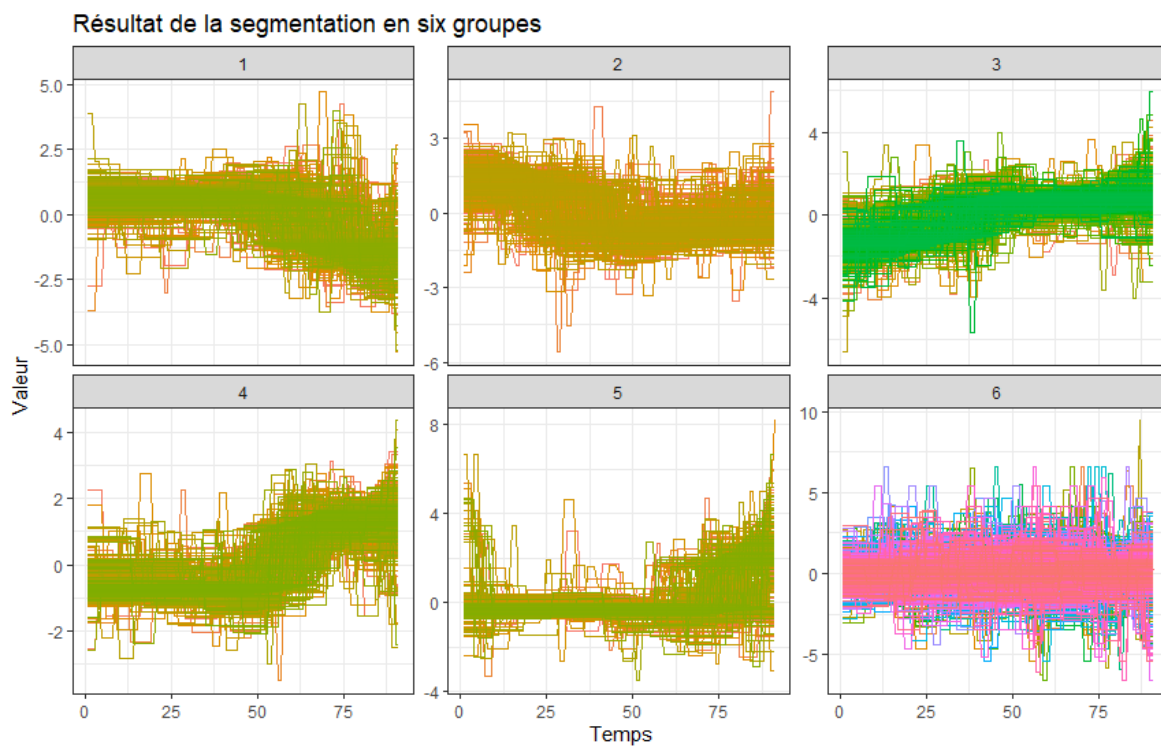


Figure 4.15 : Résultats de la segmentation avec 6 groupes et une normalisation avec la norme Z

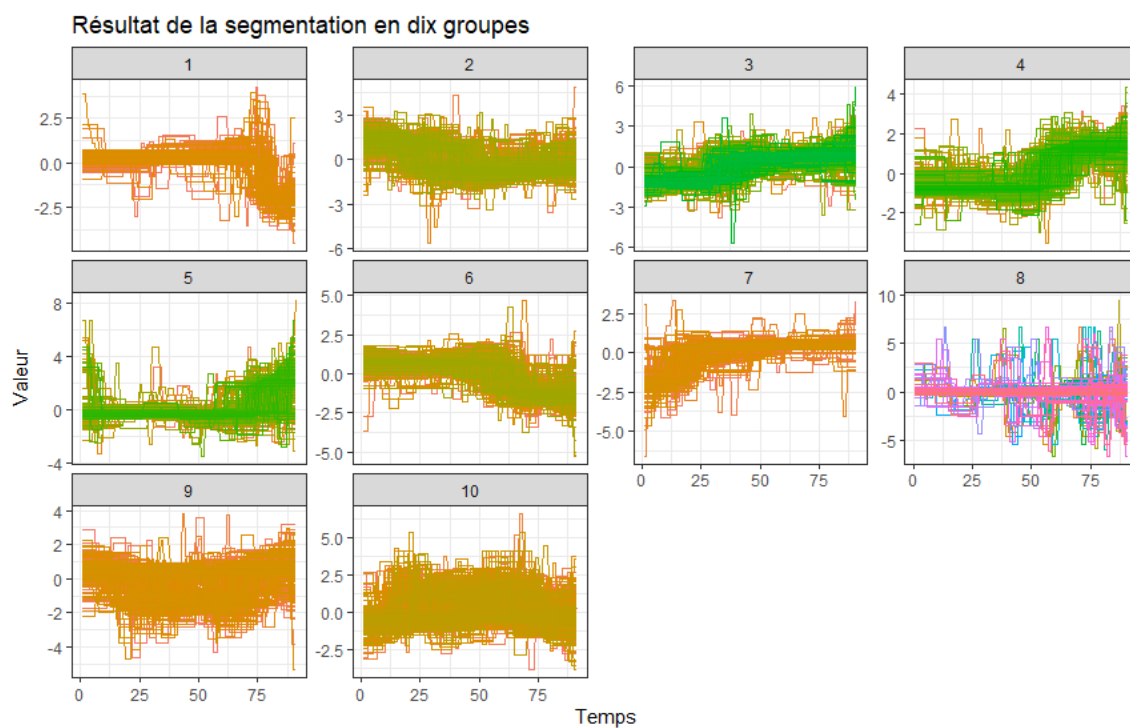


Figure 4.16 : Résultats de la segmentation avec 10 groupes et une normalisation avec la norme Z

La figure 4.15 présente le résultat de la segmentation par une classification hiérarchique agglomérative, en fixant le nombre de groupes à  $k=6$ . Les données ont été normalisées avec une norme Z. Nous remarquons aussi que les groupes sont plus définis, c'est-à-dire que les variations sont plus uniformes, et les tendances sont plus visibles. Sur la figure 4.15, on remarque que les groupes 3, 4 et 5 ont des tendances croissantes. Le groupe 1 a une tendance décroissante. Le groupe 2 est en fluctuation, et le groupe 6 ne présente pas une tendance visible.

La figure 4.16 présente le résultat de la segmentation par une classification hiérarchique agglomérative, en fixant le nombre de groupes à  $k=10$ . Les données ont été normalisées avec une norme Z. Sur la figure 4.16, nous remarquons que l'intervalle de variation a diminué pour avoir un minimum de -10 et un maximum de 10. Nous remarquons aussi que les groupes sont plus définis. Les variations sont plus uniformes, et les tendances sont plus visibles. Les courbes des groupes 3, 4 et 7 sont croissantes. Les courbes des groupes 1 et 5 sont constantes sur la grande partie de l'intervalle temporel, avec respectivement une diminution et une augmentation des tailles aux dernières années. Seul le groupe 6 présente une tendance fluctuante. Les groupes 2, 9 et 10 sont en fluctuation, et le groupe 8 ne présente aucune tendance.

Les figures 4.17 et 4.18 présentent les pourcentages de clients dans chaque groupe, respectivement pour la segmentation en six groupes et en dix groupes.

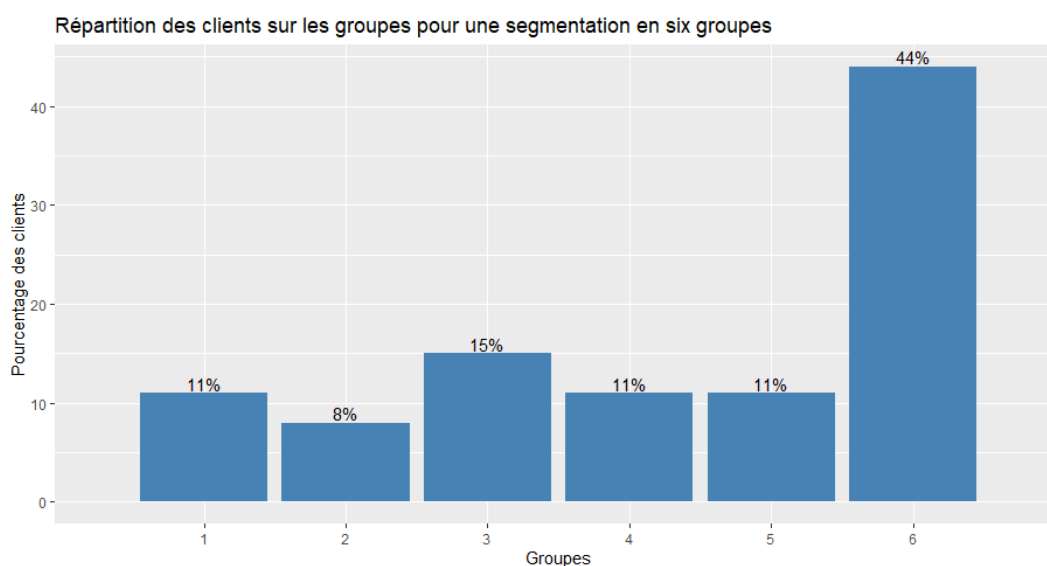


Figure 4.17: Pourcentage des clients pour la segmentation en six groupes.



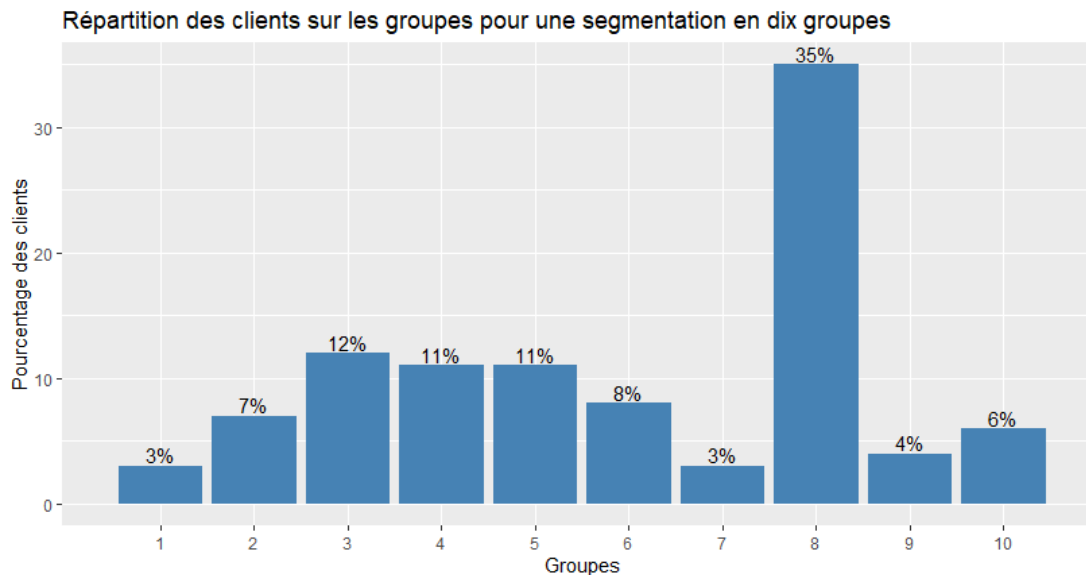


Figure 4.18: Pourcentage des clients pour la segmentation en dix groupes.

Les figures 4.17 et 4.18 présentent les tailles de chacun des groupes par rapport à l'ensemble des clients. Pour la première segmentation, c'est le groupe 6 qui regroupe le plus de clients (44 %). Pour la deuxième segmentation, c'est le groupe 8 qui est majoritaire (35 %).

Au sein de cette partie, nous avons mis en test différentes alternatives pour effectuer la segmentation. Ces tests nous ont permis de choisir la DTW comme mesure de similarité, et la méthode hiérarchique agglomérative comme méthode de segmentation. Le nombre de groupe optimal défini est 10. Les séries temporelles normalisées avec la norme Z ont permis d'avoir le meilleur résultat. Les résultats de la segmentation vont être analysés dans la section suivante.

#### 4.3.4 Analyse des résultats

À la suite des tests et des groupes générés à la section précédente, cette phase consiste à analyser la composition des segments obtenus et la lier avec les autres caractéristiques des clients. On rappelle que les résultats de la segmentation retenus sont ceux des séries d'évolution des mesures du tour de buste pour des chemises. Nous avons considéré un échantillon qui représente le quart de

l'ensemble des clients. Les séries temporelles ont été normalisées en utilisant une norme Z. Dans la partie précédente, nous avons expliqué que théoriquement le nombre de groupes optimal à choisir est égal à six. Suites aux essais pratiques et les discussions avec les expertises de notre partenaire, il était convenu de segmenter la population en dix groupes.

### Cas de six groupes :

La segmentation de l'ensemble de clients en six groupes a permis de générer la figure 4.15. Sur cette figure, nous pouvons observer les différentes tendances de chaque segment. L'indice de silhouette de cette segmentation est égal à 0,39. L'indice de silhouette est un indice qui varie entre 0 et 1. Il s'agit d'un indice à maximiser. La valeur obtenue prouve dans ce cas que la segmentation n'est pas optimale.

Pour comprendre davantage les comportements de chaque groupe, nous observons sur la figure 4.19, les centroïdes de chaque groupe.

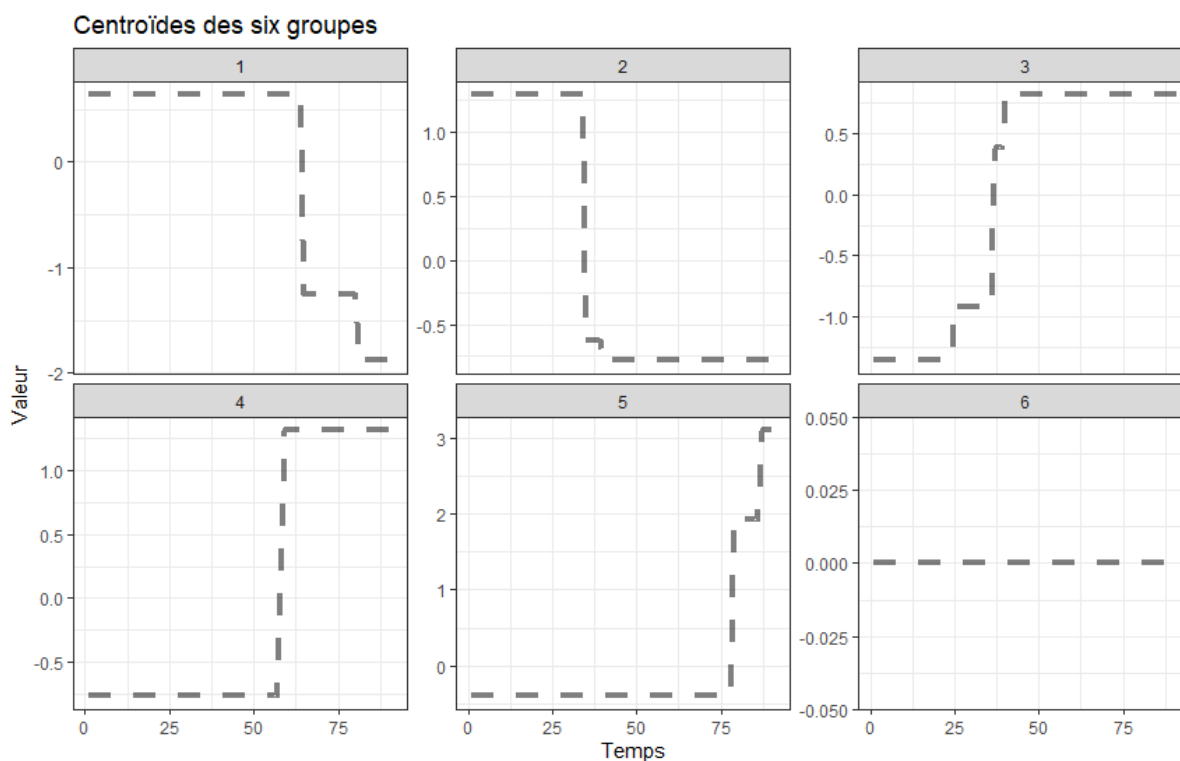






Figure 4.19 : Courbes des centroïdes de la segmentation en six groupes

Sur la figure 4.19, nous remarquons que les groupes 1 et 2 ont des courbes décroissantes, alors que les courbes 3, 4 et 5 ont des tendances croissantes. Seule la courbe 6 est constante. Sur le tableau 4-1, nous présentons une étude globale de chacun des segments. Dans ce tableau, nous présentons les tendances des courbes de chaque groupe, c'est-à-dire si elles sont stables, en croissance, en décroissance, en fluctuation, ou n'admettent aucune tendance visible. Le tableau inclut également le pourcentage des clients dans chaque segment par rapport à tout l'échantillon, et le pourcentage des clients pour chaque type de tendance, les pourcentages d'hommes et des femmes dans chaque groupe, le pourcentage des commandes en chemises effectuées par les clients de chaque segment ainsi que pour chaque type de tendance, et le pourcentage de retour. Pour chaque groupe, le pourcentage de retour présente le pourcentage du nombre des retours effectués par les clients dans ce groupe par rapport aux nombres de commandes. Il nous permet d'avoir une idée sur les habitudes de retour relativement à celle d'achat. Le tableau 4-1 contient également la moyenne des indices de silhouette observés pour les individus de chaque groupe et la moyenne des tailles.

Le tableau 4-1 présente une description globale des groupes. On peut remarquer que le groupe 6 contient le plus de clients (39 %). Il représente le groupe avec le plus de commandes. Le plus grand nombre de retours a été observé au groupe 3. Nous observons également que le groupe 2 possède une courbe fluctuante, nous désignons par cela le fait que la courbe est croissante par période, et décroissante par d'autres. Le groupe 6 en revanche est sans forme détectable. Les séries temporelles de ce groupe ne suivent pas une tendance particulière, bien que le centroïde de ce groupe affiche une évolution constante. Dans ce qui suit, nous allons analyser plus en profondeur les caractéristiques du groupe 6 et le groupe 3.

Tableau 4-1: Caractéristiques de la segmentation avec six groupes

Tendance	Groupes	Pourcentage clients	Pourcentage clients	Pourcentage Femmes	Pourcentage Hommes	Pourcentage achat	Pourcentage achat	Pourcentage retours	Moyenne de l'indice de silhouette	Moyenne des tailles
	3	15 %	37 %	18 %	82 %	11 %	38 %	8 %	0,18	47,23
	4	11 %		20 %	80 %	10 %		6 %	0,43	46,63
	5	11 %		20 %	80 %	17 %		5 %	0,25	45,57
	1	11 %	11 %	16 %	84 %	12 %	12 %	7 %	0,3	45,12
	2	8 %	8 %	15 %	85 %	11 %	11 %	7 %	0,3	45,43
	6	44 %	44 %	33 %	73 %	39 %	39 %	6 %	0,51	44,59

Groupe 6 :

Le groupe 6 est le groupe majoritaire en nombres de clients. Ce groupe retient 44 % des individus, avec 30 % de femmes et 70 % d'hommes. Les individus de ce groupe ont effectué environ 20 000 commandes, ce qui représente 38 % de l'ensemble des commandes effectuées en chemises. Les individus de ce groupe ont retourné 6 % de leurs achats. Les tailles dans ce groupe varient entre 30 et 63 pouces, avec une moyenne de 44 pouces, et un écart type de 5,2. L'indice de silhouette de ce groupe est 0,51. Ce groupe admet le plus haut indice de silhouette, ce qui indique que les individus de ce groupe sont les mieux classés.

### Groupe 3 :

Le groupe 3 est le deuxième groupe en nombre de clients. Il inclut 15 % de l'ensemble de la population. Ces individus sont répartis en 18 % de femmes et 82 % d'hommes. Les individus de ce groupe ont effectué 18 % de l'ensemble des commandes effectuées avec un taux de retour de 8 %. Les tailles dans ce groupe varient entre 30 et 63 pouces, avec une moyenne de 47 pouces, et un écart type de 5,6. L'indice de silhouette de ce groupe est 0,18, représentant le plus bas indice de tous les groupes.

La segmentation en six groupes ne permet pas donc d'avoir des résultats concluants. En prenant en considération le tableau d'interprétation dans la partie (2.3.3), on peut conclure que l'indice de silhouette associé à cette segmentation reflète l'existence d'une faible structure dans les données. Dans ce qui suit, nous allons aborder les résultats de la segmentation en dix groupes.

### Cas de dix groupes :

De façon analogue à la partie précédente, la segmentation en dix groupes a permis de générer la figure 4.15. Sur cette figure, nous pouvons observer les différentes tendances de chaque segment. L'indice de silhouette de cette segmentation est égal à 0,44. L'indice de silhouette obtenu par cette segmentation est supérieur à celui de la segmentation précédente. Cette valeur se rapproche du seuil confirmant avoir une bonne structure. Pour discerner mieux ces tendances, les centroïdes des différents groupes sont présentés sur la figure 4.20. Sur la figure 4.20, nous observons que les courbes centroïdes des groupes 1, 2 et 6 sont décroissantes. Celles des groupes 3, 4, 6 et 7 sont croissantes. Le centroïde du groupe 8 est constant, alors que ceux des groupes 9 et 10 sont fluctuants.

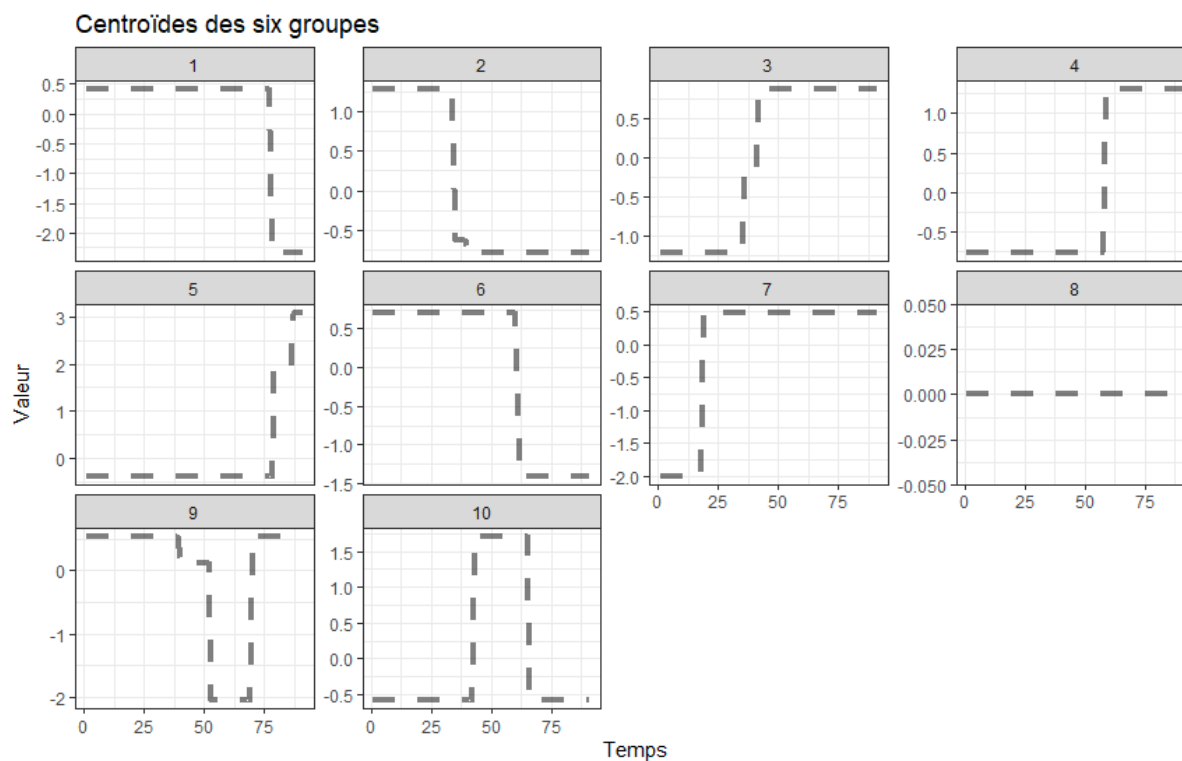







Figure 4.20 : Courbes des centroides de la segmentation en dix groupes.

Nous présentons dans le tableau 4-2, une analyse préliminaire de chacun des segments et de leurs tendances globales. Nous présentons également le pourcentage des clients dans chaque groupe par rapport au nombre total d'individus étudié, le pourcentage des hommes et des femmes, le pourcentage des achats, les pourcentages des retours effectués par rapport aux nombres de commandes (pourcentage retour), la moyenne des indices de silhouette dans chaque groupe, et la moyenne des tailles.

Tableau 4-2: Caractéristiques de la segmentation avec dix groupes.

Tendance	Groupes	Pourcentage clients	Pourcentage Femmes	Pourcentage Hommes	Pourcentage clients	Pourcentage achat	Pourcentage achat	Pourcentage retours	Moyenne de l'indice de silhouette	Moyenne des tailles
	1	3 %	18,5 %	81,5 %	14 %	4 %	14 %	9 %	0,46	45
	5	11 %	20 %	80 %		7 %		0,19	45,57	
	3	12 %	17,5 %	82,5 %	25 %	13 %	29 %	5 %	0,17	47,2
	4	11 %	20,1 %	79,9 %		12 %		7 %	0,39	46,6
	7	3 %	16 %	84 %		4 %		6 %	0,33	47,38
	6	8 %	14,7 %	81,5 %	8 %	8 %	8 %	8 %	0,28	45,14
	2	7 %	14,7 %	85,3 %	17 %	9 %	23 %	6 %	0,21	45,43
	9	4 %	15 %	85 %		5 %		6 %	-0,21	46,22
	10	6 %	12,5 %	87,5 %		9 %		6 %	-0,21	45,9
	8	35 %	33,5 %	66,5 %	35 %	26 %	26 %	6 %	0,88	43

Le tableau 4-2 présente une description générale des groupes obtenus à la suite de la segmentation. Sur le tableau, on remarque que les groupes 1 et 5 présentent des courbes stables. Les groupes 3, 4 et 7 sont en croissance. Le groupe 6 est le seul groupe en décroissance. Les groupes 2, 9 et 10 sont en fluctuation. Finalement, le groupe 8 ne présente aucune tendance particulière. Bien que beaucoup des séries temporelles du groupe 8 soient stables, l'ensemble ne semble suivre aucun modèle d'évolution particulier. On note aussi que le groupe 8 regroupe la majorité des clients étudiés (35 %). Dans ce qui suit, nous allons analyser de façon approfondie le groupe 8, étant le

groupe majoritaire, le groupe 6, étant le seul en décroissance, ainsi que le groupe 1 puisqu'il présente le plus grand taux de retours.

#### Groupe 8 :

Le groupe 8 est le groupe qui renferme le plus grand nombre de clients. Ce groupe retient 35 % de l'ensemble de la population étudiée, dont 34 % de femmes et 66 % d'hommes. Les individus de ce groupe ont effectué environ 16 000 commandes, soit 25 % de l'ensemble des commandes observées. 6 % de ces achats ont été retournés. Les tailles dans ce groupe varient entre 30 et 63 pouces, avec une moyenne de 43 pouces, et un écart type de 5,62. L'indice de silhouette de ce groupe est 0,88. Ce groupe admet le plus haut indice de silhouette, ce qui indique que les individus de ce groupe sont les mieux classés.

#### Groupe 6 :

Le groupe 6 est le seul groupe ayant une tendance décroissante. Il inclut 8 % de l'ensemble de la population, dont 15 % de femmes et 85 % d'hommes. Ce groupe a effectué 8 % du total des commandes, avec un taux de retours de 7 %. Les tailles varient entre 31 et 63 pouces, avec une moyenne de 45 pouces et un écart type de 5,2. Ce groupe admet un indice de silhouette d'une valeur de 0,28. L'indice de silhouette est supérieur à 0,25, ce qui indique la présence d'une structure. Il est cependant inférieur au seuil pour affirmer que les résultats sont assez bons.

#### Groupe 1 :

Le groupe 1 est le plus petit groupe donné par la segmentation. Ce groupe contient 3 % des clients étudiés. Les femmes dans ce groupe représentent 19 %, et les hommes 81 %. Ce groupe est responsable de 4 % des commandes effectuées, avec un pourcentage de retours d'environ 9 %. Les tailles dans ce groupe varient entre 30 et 63, avec une moyenne de 45, et un écart type de 5,6. L'indice de silhouette de ce groupe a une valeur de 0,46. La classification des individus de ce groupe est parmi les meilleurs.

### **4.3.5 Synthèse**

L'analyse des résultats de la segmentation a permis de comprendre la structure des groupes formés. Comme discuté plutôt, l'analyse prouve que la segmentation en dix groupes est meilleure que la



segmentation en six groupes. La segmentation en dix groupes a permis d'améliorer l'indice de silhouette. Cet indice a augmenté de 0,38 à 0,43.

Pour les deux segmentations, les groupes majoritaires sont des groupes qui, par la visualisation, ne révèlent aucun comportement détectable. Concernant les autres groupes, les tendances sont croissantes, décroissantes ou fluctuantes.

Les hommes sont majoritaires dans tous les groupes. Les taux de retours des groupes varient entre 5 % et 9 %. Les moyennes des mesures varient entre 43 et 47 pouces. Le groupe ayant la moyenne des tailles la moins importante est le groupe contenant le plus grand pourcentage de femmes.

Les analyses prouvent que l'ensemble des clients peut être divisé en 4 grands groupes, en se basant sur leurs comportements. Ces comportements peuvent être une croissance, une décroissance, une fluctuation, ou aucune tendance. La tentative d'expliquer ces comportements distincts, par l'examen des informations des clients n'a pas abouti à des explications exploitables.

Néanmoins, on peut constater que le plus grand nombre de commandes (29 %) a été effectué par des clients dont les tailles augmentent dans le temps. En considérant les comportements de ce groupe, l'entreprise peut améliorer ses prévisions de tailles pour les individus qui y appartiennent, et diminuer leurs taux de retours. D'un autre côté, les courbes 1 et 5 présentent un taux élevé de retour en comparaison aux autres groupes. Des mesures pour diminuer le taux des retours de ce groupe permettront de réduire le taux de retours global des chemises.

## **4.4 Conclusion**

Dans ce chapitre, nous avons commencé par présenter notre partenaire industriel. Par la suite, nous avons introduit les données analysées avec quelques statistiques descriptives. La méthodologie a été appliquée sur notre cas d'étude. Finalement, les résultats de la segmentation sont illustrés et analysés.

## CHAPITRE 5 CONCLUSION, PERSPECTIVES ET RECOMMANDATIONS

Le recours à l'analyse des données, dans la perspective d'accroître la productivité et la compétitivité, est un tournant crucial surtout pour des entreprises qui stockent leurs données depuis des années et qui peuvent enfin les exploiter grâce à l'émergence croissante de techniques de « data mining » et de « machine learning ». Parmi les outils les plus populaires dans le domaine d'analyse des données, on trouve notamment la segmentation. Cette méthode consiste à créer des groupes homogènes de produits ou d'individus en se basant sur des critères de similarité sélectionnés afin de discerner les comportements des consommateurs ou des clients.

Dans ce travail, nous présentons le déploiement du projet de recherche qui s'intitule « Développement d'un outil de segmentation des comportements d'achat des clients sur la base des données morphologiques ». Nous avons commencé par mener le recensement de l'état de l'art sur trois axes fondamentaux. Premièrement, nous avons présenté les données morphologiques dans le cadre de l'industrie du vêtement, et les enjeux auxquels cette industrie fait face. En second lieu, nous avons abordé la segmentation du marché et les travaux qui étaient faits sur ce sujet. Et finalement, nous nous sommes penchés sur les séries temporelles et quelques-unes de leurs caractéristiques, ainsi que les principales méthodes et métriques pour les analyser.

Dans un deuxième temps, nous avons consacré le troisième chapitre à l'explication plus profonde du cadre du projet essentiellement le marché d'achat en ligne, les objectifs de ce projet, et la méthodologie. Notre méthodologie se base en fait sur cinq étapes primordiales qui sont d'abord la préparation des données, ce qui va nous permettre de manipuler une base complète, bien structurée et ne contenant que les attributs nécessaires à nos analyses. Deuxièmement, nous nous sommes basés sur ces données pour créer des séries temporelles que nous avons lissées par la suite afin de corriger l'effet d'intermittence. La troisième étape est la segmentation, qui est au cœur de ce travail. Le but de la segmentation est de diviser les clients en groupes homogènes, en se prenant en considération l'évolution de leurs tailles dans le temps. Plusieurs tests ont été effectués afin de valider la méthode et les différents paramètres à utiliser. Les résultats de cette segmentation sont évalués et en se basant sur cette évaluation on juge la nécessité d'ajuster nos paramètres et réitérer

la segmentation. Finalement, l'étape cinq consiste à analyser les segments obtenus et comprendre leurs compositions.

Dans le quatrième chapitre, nous avons commencé par présenter notre partenaire industriel. Par la suite, nous avons appliqué la méthodologie présentée plus haut dans le cadre du contexte du projet. Et finalement, les résultats obtenus sont présentés et analysés au cours de la dernière partie. Nous avons effectué deux découpages en avec des nombres de groupes égaux à  $k=6$  et  $k=10$ . Ceci nous a permis de comparer les résultats de deux segmentations, et d'analyser leurs résultats. L'analyse des groupes, et l'indice de silhouette ont affirmé que la segmentation en dix groupes est meilleure que la segmentation en seulement six groupes. Nous avons remarqué que les structures des groupes sont relativement similaires. Cependant, la majorité des commandes effectuées sont associées aux clients dont la taille augmente dans le temps. Ces clients présentent une piste d'amélioration pour l'entreprise du côté des prévisions. Les groupes stables présentent également un taux élevé de retours. Prendre en considération ces individus permettra également de diminuer les taux de retours.

Ce projet nous a donc permis à manipuler les données des commandes des clients de notre partenaire industriel en vue de les segmenter en groupes de comportements similaires, et ceci en se basant sur l'évolution de leurs mensurations. Lorsqu'on parle de mensurations, on désigne celles des vêtements achetés, ce qui va naturellement aider l'entreprise à comprendre le besoin de ces clients en matière de taille de produits commandés et livrés.

Cependant, même si la segmentation est un jalon indispensable dans beaucoup d'études tel que la nôtre, elle reste une étape préparatoire pour d'autres recherches plus poussées [Zhang et al., 2010]. En ce qui concerne notre cas, la segmentation du marché va nous permettre d'orienter et structurer les analyses menées au niveau des retours qui sont des dépenses cachées pour l'entreprise. Être capable d'explorer les retours de chaque segment de clients en corrélation avec leurs tailles, et pouvoir prédire leurs retours futurs par exemple présentent de bonnes pistes de continuité.

Par ailleurs, une autre piste de recherche pour l'analyse des données morphologiques, serait d'étudier aussi les vraies mensurations corporelles de ces individus. Parce qu'en fait, lorsqu'on considère des mensurations de vêtements, le goût personnel intervient à un certain niveau : il y a des personnes qui préfèrent leurs vêtements serrés alors que d'autres les préfèrent plus amples. Par conséquent, deux personnes ayant la même morphologie pourront acheter des tailles différentes.

Lors de ce travail, nous avons été menés à faire des choix lors de plusieurs paliers du travail. Nous avons choisi de considérer la moyenne des tailles lors de plusieurs commandes effectuées le même mois. Nous avons aussi décidé d'éliminer les commandes uniques de tailles beaucoup plus grandes ou plus petites que ce que le client achète en les considérant en tant que valeurs aberrantes. Un autre choix effectué était d'écarter les clients avec un nombre de commandes inférieur au minimum fixé (4). Alors que ces choix nous ont permis d'avancer notre projet, il serait judicieux, dans une prochaine étude, de vérifier l'influence de ces choix sur la qualité des groupes obtenus.

Aussi, nous avons construit des séries temporelles de mêmes longueurs. Ces séries temporelles admettent la même date de début, et la même date de fin. Ceci suppose que tous les clients ont commencé à effectuer leurs commandes à la même date, ce qui n'est pas toujours le cas. Une piste d'amélioration serait donc de trouver les moyens de considérer les séries avec des longueurs, et échelles temporelles différentes.

## RÉFÉRENCES

- Agard B., Partovi Nia V. et Trépanier M., Assessing public transport travel behavior from smart card data with advanced data mining techniques, *World Conference on Transport Research*, vol. 13, p. 15-18, 2013.
- Agrawal R. et al., Automatic subspace clustering of high dimensional data for data mining applications, *Special Interest Group on Management of Data Record*, vol. 27, n° 2, p. 94-105, juin 1998.
- Aggarwal C. C. et al., Fast algorithms for projected clustering, *Special Interest Group on Management of Data Record*, vol. 28, n° 2, p. 61-72, 1999.
- Aghabozorgi S., Shirkhorshidi A. S. et Wah T. Y., Time-series clustering - a decade review, *Information Systems*, vol. 53, p. 16-38, 2015.
- Amed I. et al., *The state of fashion 2018: Renewed optimism for the fashion industry*, [En ligne]. Disponible : <https://www.mckinsey.com/industries/retail/our-insights/renewed-optimism-for-the-fashion-industry>, 2017.
- Ankerst M. et al., Optics: ordering points to identify the clustering structure, *Association for Computing Machinery Special Interest Group on Management of Data Record*, vol. 28, n° 2, p. 49-60, 1999.
- Ansari Z., Azeem M. F., Ahmed W. et Babu A. V., Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions, *World of Computer Science and Information Technology Journal*, vol 1., 2015.
- Arbelaitz O. et al., An extensive comparative study of cluster validity indices, *Pattern Recognition*, vol. 46, p. 243-256, 2013.
- Arora P., Varshney S. et al., Analysis of k-means and k-medoids algorithm for big data, *Procedia Computer Science*, vol. 78, p. 507-512, 2016.
- Astola J. et al., Median type filters for color signals, *1988 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol.2, p. 1753-1756, 1988.

- Aurier P., Segmentation : Une approche méthodologique, *Recherche et Applications en Marketing*, vol. 4, n° 3, p. 53-75, 1989.
- Azami H., Mohammadi K. et Bozorgtabar B., An improved signal segmentation using moving average and savitzky-golay filter, *Journal of Signal and Information Processing*, vol. 3, n° 1, p. 39-44, 2012.
- Backer E. et Jain A. K., A clustering performance measure based on fuzzy set decomposition , *IEEE Transactions on Pattern Analysis & Machine Intelligence*, n° 1, p. 66-75, 1981.
- Bakar Z. A. et al., A comparative study for outlier detection techniques in data mining, *2006 IEEE Conference on Cybernetics and Intelligent Systems*, Bangkok, Thailand, 2006.
- Bar-Joseph Z. et al., A new approach to analyzing gene expression time series data, *In proceedings of the Sixth Annual International Conference on Computational Biology*, New York, USA: ACM, p. 39-48, 2002.
- Barnett V. et Lewis T., Discordancy tests for outliers in univariate samples, *Outliers in statistical data*, vol. 3, p. 120-121, 1984.
- Bartezzaghi E., The evolution of production models: is a new paradigm emerging?, *International Journal of Operations & Production Management*, vol. 19, n° 2, p. 229-250, 1999.
- Begum N., Ulanova L., Wang J., et Keogh E., Accelerating Dynamic Time Warping Clustering with a Novel Admissible Pruning Strategy, *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, ACM, New York, NY, USA, p. 49-58, 2015.
- Ben-Dor A., Shamir R. et Yakhini Z., Clustering gene expression patterns, *Journal of Computational Biology*, vol. 6, n° 3-4, p. 281-297, 1999.
- Berkhin P., A Survey of Clustering Data Mining Techniques, *Grouping Multidimensional Data: Recent Advances in Clustering*, Springer Berlin Heidelberg, Berlin, p. 25-71, 2006.
- Berry M. J. et Linoff G., *Data Mining Techniques : For Marketing, Sales, and Customer Support*, New York, NY, USA : John Wiley & Sons, Inc., 1997.
- Billor N. et Kiral G., A comparison of multiple outlier detection methods for regression data, *Communications in Statistics - Simulation and Computation*, vol. 37, n° 3, p. 521-545, 2008.

- Bishop C. M. et al., *Neural networks for pattern recognition*, Oxford university press, New York, NY, USA, 1995.
- Bowyer K. et Ahuja N., *Advances in image understanding: A festschrift for Azriel Rosenfeld*, IEEE Computer Society Press, 1996.
- Bradley P. S. et Fayyad U. M., Refining initial points for k-means clustering, *In Proceedings of the Fifteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, vol. 98., p. 91-99, 1998.
- Brown R., *Smoothing, Forecasting and Prediction of Discrete Time Series*, Englewood Cliffs, N.J., Prentice-Hall, 1963.
- Cardoso M. G., de Carvalho A. P. d. L. et al., Quality indices for (practical) clustering evaluation, *Intelligent Data Analysis*, vol. 13, n° 5, p. 725-740, 2009.
- Cassagnes-Brouquet S. et Dousset-Seiden C., Genre, normes et langages du costume, *Clio. Femmes, Genre, Histoire* [En ligne], n° 36, p. 7-18, 2012.
- Cefrio., « Netendances 2018 — le commerce électronique au Québec. », [En ligne]. Disponible : <https://cefrio.qc.ca/fr/enquetes-et-donnees/netendances2018-commerce-electronique-au-quebec>, 2019.
- Chadwick N., Mcmeekin D. et Tan T., Classifying eye and head movement artifacts in eeg signals, *IEEE International Conference on Digital Ecosystems and Technologies*, 2011.
- Charrad M., Ghazzali N., Boiteau V. et Niknafs A., NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set, *Journal of Statistical Software*. vol. 61, n° 6, p.1-36, 2014.
- Cordier F., Seo H., et Magnenat-Thalmann N., Made-to-measure technologies for an online clothing store, *IEEE Computer graphics and applications*, vol. 23, n° 1, p. 38-48, 2003.
- Croston J. D., Forecasting and stock control for intermittent demands, *Journal of the Operational Research Society*, vol. 23, n° 3, p. 289-303, 1972.
- Del Mondo G., *A Spatio-temporal graph-based model for the evolution of geographical entities*, Theses, Université de BretagneOccidentale-Brest, 2011.

- Dhillon I. S., Fan J. et Guan Y., *Efficient Clustering of Very Large Document Collections*, MA : Springer, Boston, US, p. 357-381, 2001.
- Dimitriadou E., Dolničar S., et Weingessel A., An examination of indexes for determining the number of clusters in binary data sets, *Psychometrika*, vol. 67, n° 1, p. 137-159, 2002.
- Dixon W. J., Analysis of extreme values, *The Annals of Mathematical Statistics*, vol. 21, n° 4, p. 488-506, 1950.
- Duda R. O. et Hart P. E., *Pattern classification and scene analysis*, Wiley, New York, vol. 81, p. 269-296, 1973.
- El Ayeb S., Agard B., Développement d'un outil de segmentation de marché en se basant sur l'évolution des données morphologiques, *13ème Congrès International de Génie Industriel – GI 2019*, Montréal (Québec), Canada, 2019.
- Encyclopedia. « Garment industry, dictionary of American history. », [En ligne]. Disponible : <https://www.encyclopedia.com/history/dictionaries-thesauruses-pictures-and-press-releases/garment-industry>, 2019.
- Ester M. et al., A density-based algorithm for discovering clusters in large spatial databases with noise, *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, Portland, Oregon, vol. 96, n° 34, p. 226-231, 1996.
- Euzenat J., *Granularité dans les représentations spatio-temporelles*, INRIA, Rapport technique RR-2242, 1994.
- Everitt B., Landau S. et Leese M., *Cluster analysis*, A Hodder Arnold Publication, Taylor & Francis, London, 2001.
- Fashion History Love To Know, E. Wilson, «Evolution of the fashion industry», [En ligne]. Disponible : <https://fashion-history.lovetoknow.com/fashion-clothing-industry/evolution-fashion-industry>, 2005.
- Fisher D., Iterative optimization and simplification of hierarchical clusterings, *Journal of Artificial Intelligence Research*, vol. 4, p. 147-178, 1996.
- Frieden B. R., A new restoring algorithm for the preferential enhancement of edge gradients, *Journal of the Optical Society of America*, vol. 66, n° 3, p. 280-283, 1976.



- Fu T., Chung F. et Ng C., Financial time series segmentation based on specialized binary tree representation, *International Conference on Data Mining*, vol. 2006, p. 26-29, 2006.
- Gaudin R. et Nicoloyannis N., Apprentissage non supervisé de séries temporelles à l'aide des k-means et d'une nouvelle méthode d'agrégation de séries, *Conférence Extraction et Gestion des Connaissances (EGC)*, p. 201-212, Paris, France, 2005.
- Ghobbar A. A. et Friend C. H., Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model, *Computers & Operations Research*, vol. 30, n° 14, p. 2097-2114, 2003.
- Grabisch M. et al., Aggregation functions: Construction methods, conjunctive, disjunctive and mixed classes, *Information Sciences*, vol. 181, n° 1, p. 23-43, 2011.
- Grabisch M. et al., Aggregation functions: means, *Information Sciences*, vol. 181, n° 1, p. 1-22, 2011.
- Grubbs F. E., Procedures for detecting outlying observations in samples, *Technometrics*, vol. 11, n° 1, p. 1-21, 1969.
- Hair J. et al., *Multivariate Data Analysis*, 7th edition, Pearson Education Limited, Harlow, United Kingdom, 2013.
- Halkidi M., Batistakis Y. et Vazirgiannis M., On clustering validation techniques, *Journal of Intelligent Information Systems*, vol. 17, n° 2, p. 107-145, 2001.
- Han H. et Nam Y., Automatic body landmark identification for various body figures, *International Journal of Industrial Ergonomics*, vol. 41, n° 6, p. 592-606, 2011.
- Han, J. and Kamber, M., *Data Mining: Concepts and Techniques*, 2nd Edition, Morgan Kaufmann Publishers, San Francisco, 2006.
- Hansen P. et Jaumard B., Cluster analysis and mathematical programming, *Mathematical Programming*, vol. 79, n° 1, p. 191-215, 1997.
- Hartigan J. A. et Wong M. A., Algorithm as 136 : A k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, n° 1, p. 100-108, 1979.

- He L., Agard B. et Trépanier M., A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method, *Transportmetrica A : Transport Science*, p. 1-20, 2018.
- Huber P. J. et al., Robust regression : asymptotics, conjectures and monte carlo, *The Annals of Statistics*, vol. 1, n° 5, p. 799-821, 1973.
- Hui C. L. et Ng S. F., Predicting seam performance of commercial woven fabrics using multiple logarithm regression and artificial neural networks , *Textile Research Journal*, vol. 79, no° 18, p. 1649-1657, 2009.
- Jain A. K. et Dubes R. C., *Algorithms for clustering data*, Englewood Cliffs : Prentice Hall, 1988.
- Jiantong Z. et Biyu L., Forecasting intermittent demand based on grey theory, *Second International Conference on Intelligent Computation Technology and Automation*, vol. 2. IEEE, p. 49-52, 2009.
- Jin, H., & Reed, D. Network and Parallel Computing, *IFIP International Conference, Springer Science & Business Media*, vol. 3779, Beijing, China, 2005.
- Johnston F. R. et Boylan J. E., Forecasting for items with intermittent demand, *The Journal of the Operational Research Society*, vol. 47, n° 1, p. 113-121, 1996.
- Kanungo T. et al., An efficient k-means clustering algorithm : Analysis and implementation, *EEE Transactions on Pattern Analysis & Machine Intelligence*, n° 7, p. 881-892, 2002.
- Karypis G., Han E. et Kumar V., *Multilevel refinement for hierarchical clustering*, Minnesota Univ Minneapolis Dept Of Computer Science, Rapport technique, 1999.
- Kass R. E., Uri T. E., et Emery N. B., Analysis of neural data, *Springer Science & Business Media*, vol. 491, New York, USA, 2014.
- Kaufman L. et Rousseeuw P. J., *Finding groups in data : an introduction to cluster analysis*, John Wiley & Sons, vol. 344, 2009.
- Kaufman L. et Rousseeuw P. J., Partitioning around medoids (program pam), *Finding groups in data : an introduction to cluster analysis* , vol. 344, p. 68-125, 1990.
- Kennedy K., What size am i? decoding women's clothing standards, *Fashion Theory*, vol. 13, n° 4, p. 511-530, 2009.

- Kim B., Kim J. et Yi G., Analysis of clustering evaluation considering features of item response data using data mining technique for setting cut-off scores , *Symmetry*, vol. 9, n° 5, 2017.
- Kim M., et Ramakrisna R. S., New indices for cluster validity assessment, *Pattern Recognition Letters*, vol. 26, n° 15, p. 2353-2363, 2005.
- Kotler P. et al., *Le marketing : de la théorie à la pratique*, Gaetan Morin Editeur Limitee, Canada, 1991.
- Kourentzes N., Intermittent demand forecasts with neural networks, *International Journal of Production Economics*, vol. 143, n° 1, p. 198-206, 2013.
- Kurbalija V. et al., Time-series mining in a psychological domain, *In Proceedings of the Fifth Balkan Conference in Informatics*, New York, USA : ACM, p. 58-63, 2012,.
- Liu L.-M. et al., Forecasting and time series analysis using the SCA statistical system, *Scientific Computing Associates DeKalb, IL*, vol. 1, n° 2, 1992.
- Liu Z. et al., Predicting detailed body sizes by feature parameters, *International Journal of Clothing Science and Technology*, vol. 26, n° 2, p. 118-130, 2014.
- MacQueen J. et al., Some methods for classification and analysis of multivariate observations, dans *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, n°14. Oakland, CA, USA, p. 281-297, 1967.
- Miller H. J. et Han J., *Geographic Data Mining and Knowledge Discovery*, Bristol, PA, USA : Taylor & Francis, Inc., 2001.
- Moore A. W. et Jorgenson J. W., Median filtering for removal of low-frequency background drift , *Analytical Chemistry*, vol. 65, n° 2, p. 188-191, 1993.
- Muñoz-Garcia, J., et al. Outliers : A Formal Approach, *International Statistical Review / Revue Internationale De Statistique*, vol. 58, n° 3, p. 215-226, 1990.
- Murray P. W., Agard B. et Barajas M. A., Asact - data preparation for forecasting: A method to substitute transaction data for unavailable product consumption data, *International Journal of Production Economics*, vol. 203, p. 264-275, 2018.
- Ng R. T. et Han J., Clarans : A method for clustering objects for spatial data mining, *IEEE Transactions on Knowledge & Data Engineering*, n° 5, p. 1003-1016, 2002.

- O'Connell L., *Total revenue of the global sports apparel market 2012-2025*, [En ligne]. Disponible : <https://www.statista.com/statistics/254489/total-revenue-of-the-global-sports-apparel-market>, 2019.
- Pal N. R., Bezdek J. C. et Tsao E. C., Generalized clustering networks and kohonen's self-organizing scheme, *IEEE Transactions on Neural Networks*, vol. 4, n° 4, p. 549-557, 1993.
- Paparrizos J., Gravano L., k-Shape: Efficient and Accurate Clustering of Time Series, *In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD n° 15, pp. 1855-1870. ACM, New York, NY, USA, 2015
- Pelleg D. et Moore A. W., X-means : Extending k-means with efficient estimation of the number of clusters, *In Proceedings of the Seventeenth International Conference on Machine Learning*, Pat Langley (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 727-734, 2000.
- Peña D. et Yohai V. J., The detection of influential subsets in linear regression by using an influence matrix, *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 57, n° 1, p. 145-156, 1995.
- Pena J. M., Lozano J. A. et Larranaga P., An empirical comparison of four initialization methods for the k-means algorithm, *Pattern recognition letters*, vol. 20, n° 10, p. 1027-1040, 1999.
- Persson F. et al., Using simulation to determine the safety stock level for intermittent demand, *Winter Simulation Conference (WSC). IEEE*, p. 3768-3779, 2017.
- Petropoulos F., Kourentzes N. et Nikolopoulos K., Another look at estimators for intermittent demand, *International Journal of Production Economics*, vol. 181, p. 154-161, 2016.
- Planchon V., *Traitement des valeurs aberrantes : concepts actuels et tendances générales*, BASE [En ligne], volume 9, n° 1, p.19-34, Disponible : <https://popups.uliege.be:443/1780-4507/index.php?id=13859>, 2005.
- Price M. N., Dehal P. S. et Arkin A. P., FastTree: computing large minimum evolution trees with profiles instead of a distance matrix., *Molecular biology and evolution*, vol. 26, p. 1641-1650, 2009.

- Rabiner L, Sambur M. et Schmidt C., Applications of a nonlinear smoothing algorithm to speech processing, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, n° 6, p. 552-557, 1975.
- Ragot J., Mesures aberrantes : détection, localisation, suppression, acceptation et robustesse, *8ème Colloque Interdisciplinaire en Instrumentation*, C2I, Bordeaux, France, 2019.
- Rahm E. et Do H., Data cleaning : Problems and current approaches, *IEEE Data Eng. Bull.*, vol. 23, p. 3-13, 2000.
- Ratanamahatana C. A. et Keogh E., Multimedia retrieval using time series representation and relevance feedback, *In In Proceedings of the 8th international conference on Asian Digital Libraries: implementing strategies and sharing experiences (ICADL'05)*, Springer-Verlag, Berlin, Heidelberg, p. 400-405, 12-15 décembre, 2005.
- Rehm J. et Gmel G., Aggregate time-series regression in the field of alcohol , *Addiction*, Abingdon, England, vol. 96, n° 7, p. 945-954, 2001.
- Rendón E., Abundez I., Alejandra A., et Quiroz E. M., Internal versus external cluster validation indexes, *International Journal of computers and communications*, vol. 5, n° 1, p. 27-34, 2011.
- Sadahiro Y. et Kobayashi T., Exploratory analysis of time series data: Detection of partial similarities, clustering, and visualization, *Computers, Environment and Urban Systems*, vol. 45, p. 24-33, 2014.
- Sakoe H. et Chiba S., Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Tcoustics Speech, and Signal Processing*, vol. 26, n° 1, p. 43-49, 1978.
- Savitzky A. et Golay M. J. E., Smoothing and differentiation of data by simplified least squares procedures, *Analytical Chemistry*, vol. 36, n° 8, p. 1627-1639, 1964.
- Sen A., The us fashion industry: A supply chain review, *International Journal of Production Economics*, vol. 114, n° 2, p. 571-593, 2008.
- Shen H.-b., Yang J. et Wang S.-t., Outlier detecting in fuzzy switching regression models, *Artificial Intelligence : Methodology, Systems, and Applications*, Springer Berlin Heidelberg, p. 208-215, Berlin, Heidelberg, 2004.

- Stone D. C., Application of median filtering to noisy data , *Canadian Journal of chemistry*, vol. 73, n° 10, p. 1573-1581, 1995.
- Struyf A. et al., Clustering in an object-oriented environment, *Journal of Statistical Software*, vol. 1, n° 4, p. 1-30, 1997.
- Syntetos A. A. et Boylan J. E., On the bias of intermittent demand estimates , *International Journal of Production Economics*, vol. 71, n° 1-3, p. 457-466, 2001.
- Syntetos A. A. et Boylan J. E., On the variance of intermittent demand estimates , *International Journal of Production Economics*, vol. 128, n° 2, p. 546-555, 2010.
- Syntetos A. A., et al., Supply chain forecasting : Theory, practice, their gap and the future, *European Journal of Operational Research*, vol. 252, n° 1, p. 1-26, 2016.
- Syntetos, A. A. et Boylan J. E., The accuracy of intermittent demand estimates, *International Journal of forecasting*, vol. 21, n° 2, p. 303-314, 2005.
- Thalmann N. et al., Modeling of bodies and clothes for virtual environments, *International Conference on Cyberworlds*, Tokyo, Japan, p. 201-208, 2004.
- Tiao G. C., Asymptotic behaviour of temporal aggregates of time series, *Biometrika*, vol. 59, n° 3, p. 525-531, 1972.
- Tokatli N., Global sourcing : insights from the global clothing industry-the case of Zara, a fast fashion retailer, *Journal of Economic Geography*, vol. 8, n° 1, p. 21-38, 2007.
- Trendex North America. «Canadian apparel market.», [En ligne]. Disponible : <http://www.trendexna.com/canadian-market>, 2018.
- Tukey J., *Exploratory Data Analysis*, Addison-Wesley Publishing Company, 1977.
- Van der Laan M., Pollard K. et Bryan J., A new partitioning around medoids algorithm, *Journal of Statistical Computation and Simulation*, vol. 73, n° 8, p. 575-584, 2003.
- Velmurugan T. et Santhanam T., Computational complexity between k-means and kmedoids clustering algorithms for normal and uniform distributions of data points, *Journal of computer science*, vol. 6, n° 3, p. 363-368, 2010.

- Vliegthart R., Moving up. applying aggregate level time series analysis in the study of media coverage, *Quality & Quantity*, vol. 48, n° 5, p. 2427-2445, 2014.
- Wang J. et Lin L., Improved median filter using minmax algorithm for image processing, *Electronics Letters*, vol. 33, n° 16, p. 1362-1363, 1997.
- Wang W., Yang J. et Muntz R., Sting+ : An approach to active spatial data mining, *In Proceedings 15th International Conference on Data Engineering IEEE*, p. 116-125, 1999.
- Wang X. et al., Experimental comparison of representation methods and distance measures for time series data, *Data Mining and Knowledge Discovery*, vol. 26, n° 2, p. 275-309, 2013.
- Wang X., Smith K., et Hyndman R., Characteristic-based clustering for time series data, *Data mining and knowledge Discovery*, vol. 13, n° 3, p. 335-364, 2006.
- Ward Jr J. H., Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, vol. 58, n° 301, p. 236-244, 1963.
- Wedel M. et Kamakura W. A., *Clustering Methods*, Springer, Boston, USA, p. 39-73, 2000.
- Willemain T. R., Smart C. N. et Schwarz H. F., A new approach to forecasting intermittent demand for service parts inventories, *International Journal of forecasting*, vol. 20, n° 3, p. 375-387, 2004.
- Winks J. M., *Clothing sizes : International standardization*, Textile Institute, Manchester, 1997.
- Xu R. et Wunsch II D., Survey of clustering algorithms, *IEEE Transactions on Neural Networks*, vol. 16, n° 3, p. 645-678, 2005.
- Yang Q. et Wu X., 10 challenging problems in data mining research, *International Journal of Information Technology & Decision Making*, vol. 5, n° 4, p. 597-604, 2006.
- Zadeh L. A., Fuzzy sets, *Information and control*, vol. 8, n° 3, p. 338-353, 1965.
- Zhang Z., Huang K. et Tan T., Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes, *In Proceedings of the 18<sup>th</sup> International Conference on Pattern Recognition*, vol. 03, Washington DC, USA, 2006.

Zheng R, Yu W. et Fan J., Development of a new chinese bra sizing system based on breast anthropometric measurements, *International Journal of Industrial Ergonomics*, vol. 37, n° 8, p. 697-705, 2007.