DEVELOPPEMENT D'UN OUTIL DE SEGMENTATION DE MARCHE EN SE BASANT SUR L'EVOLUTION DES DONNEES MORPHOLOGIQUES

SAFA EL AYEB[1,2], BRUNO AGARD[1,2]

[1] Laboratoire en Intelligence des Données
Département de mathématiques et génie industriel,
École Polytechnique de Montréal, CP 6079, succursale Centre-Ville, Montréal, Québec, Canada
safa.elayeb@polymtl.ca, bruno.agard@polymtl.ca

[2] Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport (CIRRELT)

***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. Cette analyse est fortement utilisée par les entreprises afin de les guider dans leurs décisions opérationnelles et stratégiques reliés à des produits et services existants ou voire éventuels. Cet article présente un des outils d'analyse des marchés qui est la segmentation des clients, en considérant des variables externes. Cette approche vise à diviser un ensemble d'individus hétérogènes en groupes plus homogènes en se basant sur les données morphologiques. Dans ce travail, nous avons utilisé des méthodes de « data mining » afin de comprendre les tendances dans l'évolution des mensurations des clients. L'analyse se base sur la segmentation des séries temporelles, construites à partir des données historiques. Cette étude a été menée sur les données morphologiques des clients, fournies par notre partenaire industriel. Plusieurs tests basés sur différents critères ont eu place afin de d'obtenir des groupes ayant des comportements similaires. Finalement, une évaluation de la segmentation a été effectuée.**

***Abstract* – Market analysis is nowadays considered more a managerial field than a scientific one. It is based on collecting clients' data and exploring them in order to understand their behavior. Industries use this analysis in order to guide their operational and strategic decisions, related to existent products or services as well as the future ones. This paper presents a market analysis tool that is market segmentation based on morphological data. Segmentation consists on dividing heterogeneous populations into homogeneous groups that have the same patterns. This work aims to discover tendencies in clients' behavior, based on their historical data. The data used in this study are morphological data, provided by our industrial partner. Segmentation on different sets of time series will take place. The consistency of the results will be afterwards assessed through a clusters' evaluation.**

***Mots clés* – Séries temporelles, demande intermittente, agrégation, segmentation de marché.**
***Keywords* – Time series, intermittent demand, aggregation, market segmentation.**

1 INTRODUCTION

L'analyse de données ou le « data mining » est un concept qui existe depuis 1960, bien qu'il n'ait eu sa vraie extension en tant qu'outil liant les domaines scientifiques et industriels qu'en 1970 [Zighed, Rakotomalala, 2002]. [Berson et al., 2000], [Lejeune, 2001] et [Linoff et Berry, 2011] définissent le data mining comme étant l'outil permettant de fouiller les bases de données massives en vue d'en extraire des connaissances utiles et profitables pour les entreprises. Dans ce sens, la collecte et l'exploitation des données deviennent une nécessité plutôt qu'un luxe. Non seulement elles sont indispensables pour l'adaptation aux besoins du marché par l'analyse des comportements des clients [Ngai et al., 2009], mais aussi elles jouent un rôle important aidant les industries à renforcer leurs stratégies de concurrences, en fournissant les outils nécessaires à la prédiction des futures tendances [Ahmed, 2004].

Les domaines d'application et les types de données exploitables par le data mining sont ainsi innombrables [Bellanger et Tomassone, 2014], parmi lesquelles on trouve les données morphologiques. Ces dernières concernent la forme ou l'anatomie des individus, et sont exprimées via différentes mensurations. Ces données ont déjà été abordées dans la littérature, mais révèlent encore de nombreux atouts. En effet dans le domaine de vente en détail, en particulier celui en ligne, parmi les raisons les plus importantes des retours de produits, figure le choix de la mauvaise taille. Dans ce contexte, une étude menée par l'Institut français du textile et de l'habillement en 2005, a affirmé que « près d'une femme sur trois, et près d'un homme sur six trouvent difficilement des vêtements à leurs tailles » [IFTH, 2006]. Pour cela, plusieurs technologies ont émergé afin de guider les consommateurs à trouver « la taille parfaite ». Également, de plus en plus

d'industries ont commencé à être conscient de l'importance des données de morphologie.

À cet égard, il existe dans la littérature plusieurs techniques servant l'analyse de ces données en vue d'en extraire de l'information utile. [Fayyad et al., 1996] et [Kantardzic, 2011] divisent les objectifs de ces méthodes en deux grandes familles, la description et la prédiction. Alors que la description vise à identifier des modèles et les présenter sous forme visuelle et interprétable, la prédiction quant à elle, consiste à prévoir des comportements futurs et inconnus. Parmi les méthodes d'analyse de données les plus utilisées, on peut citer la classification, la régression, la segmentation et l'analyse discriminante.

Dans le présent article, nous proposons une méthode de segmentation des individus en se basant principalement sur leurs données morphologiques. L'objectif est de comprendre d'abord les tendances des clients et l'évolution de leurs mensurations dans le temps. Par la suite, on mettra en œuvre un algorithme pour prédire les comportements futurs de ces mensurations. Ce travail a été réalisé en collaboration avec un partenaire industriel, spécialisé dans la confection, la vente et la distribution des uniformes, Logistik Unicorp.

La structure de l'article se décompose de la manière suivante : la section 2 présente l'état de l'art, dans laquelle nous introduirons d'abord la notion de demande intermittentes (2.1). Nous présenterons également les principes de l'agrégation des séries temporelles à la section (2.2). La segmentation du marché est présentée à la section (2.3), notamment la segmentation de la clientèle (2.2.1) et les métriques de distance (2.2.2). La section 3 présentera la méthodologie suivie pour l'analyse du marché. La section 4 exposera le cas d'étude qui a servi à la validation de la méthode retenue. Nous présenterons le contexte du cas d'étude (4.1), l'application de la méthodologie (4.2), et l'analyse des résultats (4.3). Finalement, la section 5 rappellera les résultats et limites de la méthode utilisée et proposera quelques perspectives.

2 ÉTAT DE L'ART

2.1 Demande intermittente

La demande intermittente est un phénomène hautement observable en milieu industriel. En fait, une demande est dite intermittente lorsqu'elle est irrégulière et diffère grandement d'une période à l'autre tout en ayant plusieurs intervalles de demandes nulles [Syntetos et al., 2001]. [Bartezzaghi et al., 1999] explique l'intermittence par cinq facteurs qui sont : le nombre des clients, leur hétérogénéité, ainsi que la fréquence, la variation et la corrélation entre leurs demandes. Ainsi, une demande intermittente est plus difficile à suivre, à analyser et à prédire comparée à une demande régulière. Pour pallier les problèmes de la prédiction avec des demandes intermittentes, une approche classique consiste à lisser les données. Dans la littérature, il existe plusieurs méthodes pour lisser les demandes intermittentes afin de rendre possible leurs exploitations. 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, elles 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 fins, 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.

2.2 Agrégation des séries temporelles

Dans l'industrie, la collecte des données de commandes des clients est un processus de plus en plus important et répandu. La collecte de ces données se fait généralement sous forme de liste, où chaque ligne présente une commande individuelle. Le problème avec ce format de données est le fait qu'il est difficilement manipulable. Pour pouvoir analyser ces données, il est nécessaire de les transformer en séries temporelles [Liu et al., 1992]. Une série temporelle est définie comme étant la succession de la même observation sur une échelle de temps. L'utilisation de ce type de données est de plus répandue dans le domaine scientifique. Cela est dû aux nombreuses méthodes et algorithmes développés facilitant leurs analyses. Pourtant, l'un des défis liés aux séries temporelles est leurs grands volumes et leurs complexités [Stolojescu, 2012]. À cet effet, il est nécessaire d'agréger les séries temporelles lorsque le niveau de détail y est important ou comme mentionné dans la section précédente lorsque celles-ci présentent des demandes irrégulières ou intermittentes. Parmi les choix à faire est celui du niveau d'agrégation temporel ou la fréquence. En effet, la fréquence peut être faible telle qu'à l'année ou au semestre, ou haute telle qu'à la semaine ou le mois. Le problème c'est qu'en agrégeant à faible fréquence, on risque de perdre les comportements de la population, alors qu'en agrégeant à haute fréquence, on risque d'avoir des séries temporelles intermittentes [Murray et al., 2018]. Cependant, l'agrégation n'a pas été exemptée de critiques. [Rehm et Gmel, 2001] supportent le fait que travailler avec des données agrégées risque de faire perdre les de l'information, et affaiblit la consistance des modèles obtenus. Aussi [Tiao, 1972] a trouvé que l'agrégation ne permettait pas d'améliorer les résultats des prédictions, par rapport aux données non agrégées. D'une autre part [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. De plus [Jin et al., 2015] ont prouvé que l'agrégation permettait de réduire la variabilité et de donner des prédictions plus exactes. Ainsi, pour l'étude des tendances d'une population de clients où on cherche à discerner des tendances globales plutôt que des comportements individuels, il est avantageux de considérer des données agrégées.

2.3 Segmentation du marché

Dans cette partie, nous allons aborder le concept de segmentation de marché et les objectifs de son déploiement. Par la suite, nous passerons en revue les méthodes les plus utilisées dans la communauté scientifique pour la segmentation et les métriques de calcul de distances sur lesquelles elles peuvent se baser.

2.3.1 Segmentation de la clientèle

La segmentation consiste à diviser les données en groupes en se basant sur leur similarité [Berkhin, 2006]. Ce concept est utilisé depuis longtemps, et dans tous les domaines, notamment le domaine industriel. Il a aidé les entreprises à comprendre les tendances au sein du marché et par la suite à planifier efficacement leurs stratégies de commercialisation [Smith et al., 1956]. Dans des marchés généralement hétérogènes, la segmentation consiste à découper l'ensemble des clients en sous-groupes homogènes selon des critères quantifiables tels que l'âge, la taille, le sexe, etc. Il existe plusieurs méthodes pour la segmentation, les plus communes étant k-means [MacQueen, 1967], les réseaux de neurones [Moreno et al., 1997] et la classification hiérarchique [Langfelder et al., 2008]. Chaque méthode a ses avantages et inconvénients. Pour une revue complète des méthodes de segmentation, le lecteur est invité à se référer aux travaux de [Berkhin, 2006] et [Xu, Wunsch, 2005].

2.3.2 Métriques de distances

Afin de pouvoir regrouper les clients en segments, il est nécessaire de calculer une notion de distance entre les clients, deux à deux [Price et al., 2009]. Il existe dans la littérature plusieurs métriques de calcul d'écarts. Parmi les distances les plus utilisées, on trouve la distance euclidienne [Berkhin, 2006] et la distance de Manhattan [Bakar et al., 2006]. Bien qu'elles soient très efficaces, le problème avec ces distances est qu'elles ne sont pas adaptées pour des séries où on travaille avec des données qualitatives ou lorsqu'on cherche à analyser un comportement [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 celle-ci sont décalées dans le temps. Pour remédier à ce problème [Skoe et Chiba, 1978] ont développé une technique nommée « Dynamic Time Warping » (DTW). Cette méthode a été conçue en premier lieu pour des systèmes de reconnaissance vocale. En effet, 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]. Néanmoins, beaucoup de critiques ont été adressées à cette méthode, concernant surtout son temps de calcul couteux [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.

3 METHODOLOGIE

Notre travail se base sur des données historiques collectées par Logistik Unicorp, notre partenaire industriel. Ces données rassemblent l'historique de commandes de tous ses clients, pendant toute la période de collecte de données. Initialement, chaque ligne de la base présente une commande séparée. La figure 1 illustre la méthodologie suivie dans ce travail. Elle se compose de 4 phases principales. (1) Il est nécessaire de commencer par nettoyer et filtrer les données historiques. Par la suite, (2) les données sont transformées en séries temporelles, et l'effet d'intermittence est corrigé. (3) Une agrégation permet de segmenter et de créer des groupes de clients ayant des comportements similaires. Finalement (4), une évaluation de ces groupes va nous renseigner sur les différentes natures des évolutions des données morphologiques des clients. Ceci non seulement permettra à l'entreprise de prédire l'évolution des tailles de ses clients, mais aussi de mieux gérer son inventaire, et donc d'offrir un meilleur service à ses clients.

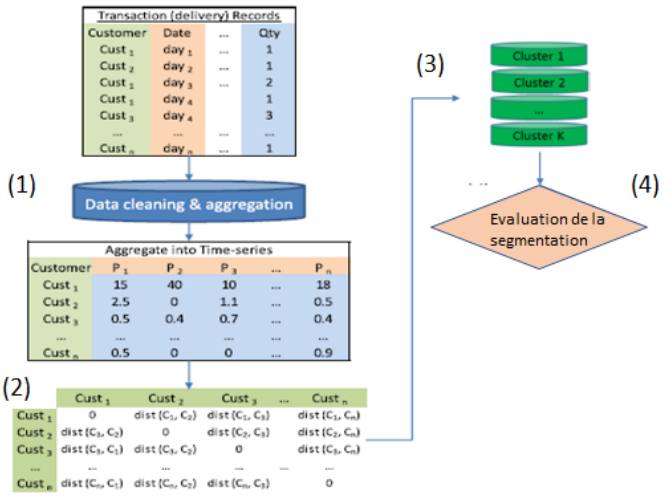

Figure 1. Méthodologie

3.1 Phase 1 : Nettoyage et préparation des données

La première phase de ce travail est la préparation de la base de données sur laquelle sera effectuée la segmentation. Les données historiques sous leur forme brute sont groupées dans une liste, où chaque ligne présente une commande, avec une date et une quantité commandée ou retournée. La base contient aussi des informations concernant le client telles que la localisation et le sexe, et des informations concernant la commande telle que la taille. Seules les variables utiles pour l'étude ont été sélectionnées. Ceci a servi à diminuer grandement la taille de la base. Toutes les valeurs aberrantes ont été supprimées. Chaque commande qui manque des informations importantes à notre étude, tel que le sexe a été éliminée. Aussi, pour quelques types de vêtements, des valeurs de mensurations trop écartées de la médiane ont été considérées comme aberrantes. D'autre part, seules les commandes individuelles ont été retenues. Ceci exclura les commandes de groupes, soit les commandes faites avec de grandes quantités. Du fait que ces commandes sont désignées pour des profils de gens différents, et basées sur l'anticipation des besoins, nous ne pourrons les utiliser pour observer les profils des clients. Les inclure dans l'analyse pourra donc biaiser les résultats. Aussi, les clients avec un nombre de commandes trop bas durant la période d'étude ont aussi été éliminés. Ces individus ne permettront pas d'étudier une vraie évolution ou tendance à cause du manque d'informations.

3.2 Phase 2 : Transformation en série temporelle

Notre objectif est d'avoir une base où chaque client présente une série, représentant toutes ses commandes au cours du temps. Comme chaque commande est caractérisée par sa date, on se retrouve avec des séries temporelles à échelle quotidienne. Le problème avec des séries temporelles à un niveau de granularité aussi fin, c'est que les données sont fortement intermittentes. Les séries contiennent beaucoup de zéros. De ce fait, la correction de ce problème est établie sur deux étapes. La première étape est une agrégation au niveau mensuel. Ainsi, en considérant le même type d'article acheté, si plusieurs commandes ont été effectuées pendant un même mois, une valeur moyenne est retenue. Après

cette agrégation, plusieurs zéros sont encore observables dans la base. Le défi, c'est que travailler avec les variables de tailles est un peu différent de la manipulation des données quantitatives, telles que des quantités. Avec ce type de données, un lissage par calcul de moyenne ou par la méthode de Croston [Croston ,1972] n'est pas possible. Ceci nous ramène à la deuxième étape, qui est le lissage des données. En effet, afin de corriger l'intermittence des données, nous nous sommes basés sur un principe simple et particulier à notre cas d'étude : tant qu'un client n'effectue pas une nouvelle commande, sa taille est considérée comme constante, et égale à la dernière taille qu'il a commandée. Le déroulement complet de ces deux étapes nous permet d'obtenir des séries temporelles, agrégées et continues.

3.3 Phase 3 : Segmentation des séries temporelles

La segmentation des séries temporelle passe d'abord par le calcul de la matrice de distance. Comme présentée dans la figure 1 étape (2), cette matrice présente la distance entre les clients deux à deux. Tel qu'expliqué dans la partie de revue de littérature, avec des données présentant des comportements, la métrique la plus adéquate à utiliser est la « DTW ». Plusieurs méthodes de segmentation existent dans la littérature. Pour notre étude, on choisit de travailler avec une classification hiérarchique ascendante. Cette méthode présente les classes sous forme de dendrogramme [Berry and Linoff, 2004]. Les classes peuvent être construites par agglomération ou par division [Barirani et al., 2013]. L'agglomération considère chaque individu comme une classe et regroupe les individus similaires à chaque itération. La division quant à elle, commence par un seul groupe, et à chaque itération construit de plus petits groupes tout en maximisant l'hétérogénéité entre eux. Le choix de la classification hiérarchique ascendante est basé sur sa flexibilité vis-à-vis des métriques de similarité aussi et son adaptation à tous les types de données [Berkhin, 2006]. Dans notre travail, la classification hiérarchique va utiliser la méthode de « WARD » [Ward, 1963] pour la construction des segments. Son principe est de grouper les individus en s'assurant à chaque pas de minimiser l'inertie intra-classe et de maximiser l'inertie interclasse [Gonzalez, 2008]. L'inertie étant la moyenne des carrés des distances entre les centres de gravité des points.

3.4 Phase 4 : Évaluation de la segmentation

Cette étape construit la dernière étape présente dans cette analyse. Plusieurs critères peuvent être utilisés pour l'évaluation de la segmentation obtenue. L'homogénéité entre les membres d'une classe et l'hétérogénéité entre les classes sont parmi les critères les plus utilisés. Pour notre travail, on a choisi le critère de Silhouette (Sil) [Arbelaitz et al., 2013]. 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 et la distance moyenne du plus proche voisin, 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. Nous avons choisi cet indice pour sa capacité à donner les meilleurs résultats surtout avec la classification hiérarchique [Arbelaitz et al., 2013].

4 CAS D'APPLICATION

Dans cette partie nous allons appliquer la méthodologie présentée dans la section précédente sur les données industrielles réelles. Nous allons commencer par une présentation du contexte industriel de l'étude et une analyse statistique des données industrielles dans la section 4.1. Nous allons par la suite revoir les étapes de nettoyage et filtrage des données dans 4.2. Ceci nous permettra d'effectuer par la suite l'étape de segmentation. Finalement, une analyse des résultats et une évaluation des segmentations obtenues auront lieu dans 4.3.

4.1 Contexte industriel

Notre partenaire industriel, Logistik Unicorp, est une compagnie de service qui est spécialisée dans la confection, la production, la vente et la distribution d'uniformes pour différents corps de métier. Il gère aussi les retours de ses clients. Le but de ce travail est d'analyser les données morphologiques des clients en se basant sur l'historique de leurs commandes. Cette étude a pour but de segmenter les clients en vue d'identifier les comportements morphologiques. Ceci est réalisé par l'observation de l'évolution de leurs mensurations pour des types de vêtements différents.
La base de données fournie par notre partenaire contient l'historique de commandes de tous ses clients. Elle couvre une durée de six ans (entre 2012 et 2018). Durant cette période, plus d'un million de commandes ont été effectuées, avec un taux de retours inférieur à 10%. Le nombre moyen de commandes par client est proche de 10. Alors que le nombre moyen de retours par client est inférieur à 1. Ainsi, la base de données contient plus de 100 000 clients dont 82.9% hommes et 17.1% femmes. Pour des commandes couvrant six années, ces clients ont acheté près de 10 000 types de vêtements, dont à titre d'exemple des chemises, des pantalons, des chaussures, etc. Toutes ces statistiques ont été groupées dans le Tableau 1.

Tableau 1. Résultats – Statistiques descriptives des données de commandes

	Nombre de commandes par client	Nombre d'articles par commande	Nombre de retours par client
Min	1	1	1
25%ile	3	15	1
Médiane	5	24	2
75%ile	9	30	4
Max	4735	2103	78

4.2 Application de la méthodologie

4.2.1 Phase 1 : Nettoyage et préparation des données

Comme décrit dans la partie méthodologie (3.1), la première étape de l'analyse est le nettoyage et le filtrage des données. Cette phase est primordiale, étant donné que la base contient plusieurs variables, et que pas toutes seront utilisées dans ce cadre particulier d'étude. Par conséquent, cette phase commence tout d'abord par discerner les attributs pertinents par rapport à une analyse de données morphologiques. Effectivement, la sélection des données adéquates permettra d'en tirer des résultats plus fiables [Agard et Kusiak, 2005]. Ainsi, pour pouvoir analyser les données de tailles, il a été d'abord convenu que seules les commandes individuelles seraient retenues.

ORDER_NUM	SPECIFIC_NSN	LINE_NUMBER	QTY_ORDERED	QTY_SHIPPED	QTY_RETURNED	ENTRY_DATE	POSTAL_CODE	CLIENT_CODE	CLIENT_SEX	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

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	45		49	49	49	49
5	45	45	45	45		45	45	45	45
7	47	47	47	47		51	51	51	51
9	46.5	46.5	46.5	46.5		46.5	46.5	46.5	46.5

Figure 2. Transformation de la base de données

Une fois la liste des clients filtrée, la liste d'attributs retenus a été définie. La base de données finale avait moins de colonnes, parmi lesquelles on retrouve les informations sur les commandes et les mensurations de chaque produit commandé ou retourné. Cependant, les données contiennent au total quinze familles de produits différents. Cependant, il existe des produits où une analyse de mensurations est impossible, tel que des cravates ou des bas. Ces derniers seront donc écartés de l'étude. Les principales familles à être considérées par notre étude seraient principalement les pantalons, les chemises et les manteaux ou vêtements d'extérieur. L'un des défis rencontrés par la suite est le fait que chaque article est identifié par un nombre de mensurations qui varie entre deux et cinq mesures différentes. De plus des produits appartenant à la même famille ne comportent pas nécessairement le même nombre de mensurations. À titre d'exemple, les catégories de vêtements de type chemise peuvent avoir ou bien deux mensurations ou bien trois mensurations. Ceci pose un problème dans la mesure où on désire grouper tous les vêtements semblables dans une seule catégorie qui sera étudiée de manière globale. Donc en se basant sur les informations présentées ci-haut, il a été convenu de considérer une seule mensuration pour chaque type de vêtement. Dans ce qui suit, on fera référence à cette mensuration en tant que « la mensuration retenue ».

À la fin de cette étape, la base de données ne contient que les clients dont on peut suivre l'évolution de taille dans le temps. Pour chacune des commandes de ces derniers, seuls les attributs contributifs à l'analyse sont conservés.

4.2.2 Phase 2 : Segmentation des séries temporelles

Cette phase peut être divisée en deux étapes importantes. La première serait de construire à partir de la base de données actuelle une base groupant un ensemble de séries temporelles. La deuxième étape serait la segmentation de ces dernières, en regroupant ensemble les clients ayant les comportements similaires.

Ainsi, le premier choix à faire est le choix de l'échelle temporelle ou la granularité. Choisir une granularité fine, soit au niveau quotidien, non seulement résultera en une base de données très grande, mais aussi nous mènera à travailler avec des données intermittentes. D'un autre côté, augmenter le niveau de granularité, soit au trimestre ou à l'année, risque d'induire la perte d'information ou des tendances [Ravat et al., 2000]. Dans le cadre de notre étude, le niveau de granularité choisi est le niveau mensuel. La figure 2, présente les changements apportés à la base de données initiale.

Partant de ces données, une matrice de distance est calculée. Cette dernière se construit en calculant la distance entre chaque paire de clients, deux à deux. Il existe plusieurs métriques de distance permettant de calculer la matrice de similarité. Mais à cause du caractère comportemental de nos données, et le fait qu'elles sont « shape-based », la distance choisie est basée sur le « Dynamic Time Warping » (DTW). Ce qui caractérise cette méthode est surtout le fait qu'elle calcule la similarité des séries, de points de vue forme, sans prendre pour autant en considération les décalages temporels qui peuvent avoir lieu [Berndt et Clifford, 1994]. Ensuite, nous avons appliqué une classification hiérarchique ascendante. Cette méthode opère de manière itérative. À chaque pas, des groupes sont construits de manière à minimiser la perte d'inertie interclasses, permettant à la fin de construire un arbre, aussi appelé dendrogramme. Cet arbre nous permet d'afficher la relation entre les classes [Barirani et al., 2013], mais aussi de voir le nombre de classes optimal qu'il faut, en fonction des écarts de distance faits à chaque agglomération. Toutefois, il n'est pas toujours évident de décider quant au nombre de segments à définir seulement à travers le dendrogramme. C'est pourquoi on utilise la courbe des sauts d'inertie. Cette courbe permet de visualiser la mesure de la perte d'inertie, en fonction du nombre des segments produits. Ainsi, l'implémentant de l'algorithme de classification a été effectué avec le logiciel « R studio ». En premier lieu, nos essais ont commencé par une classification globale, prenant en considération l'ensemble de tous les clients de même sexe, pour un type de vêtement considéré, soit les chemises. La figure 3 permet de voir l'état initial des séries temporelles avant la segmentation. On peut y voir que peu d'informations peuvent être lues sur ce graphe, les données sont beaucoup bruitées. Par la suite, des segmentations sur des groupes plus spécifiques ont été accomplies. Autant dire, pour cette étape, nous avons choisi de segmenter des individus ayant le même corps métier, tout en séparant les hommes et les femmes. L'étude de ces cas spécifiques cherche à mettre en exergue si la considération des groupes plus homogènes permettait d'avoir une meilleure segmentation, et par la suite mieux observer les tendances de chaque groupe en ce qui concerne le comportement vis-à-vis des données morphologiques.

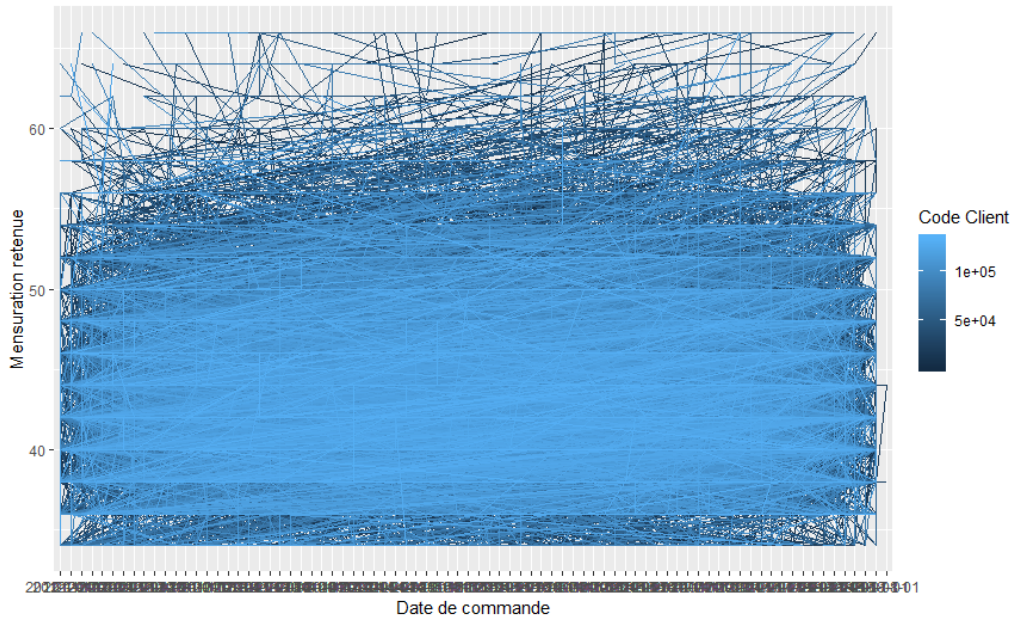

Figure 3. Allures des séries temporelles avant segmentation

4.3 Analyse des résultats

Dans cette partie, nous allons présenter les résultats de deux tests effectués. Le premier consiste à considérer les clients d'un sexe en particulier. Tandis que le deuxième concerne un métier en spécifique. Pour les deux tests, un seul type de vêtement a été considéré à la fois. Puisque la nature des mensurations diffère d'un type à l'autre, il est impossible de les considérer tous simultanément. L'implémentation de notre algorithme a commencé par la génération du dendrogramme et de la courbe des sauts d'inertie, illustrés respectivement par les figures 4 et 5. À partir de ces graphes, le nombre de classes a été établi à six groupes. Les groupes de clients ont été ensuite construits. Le résultat est présenté dans la figure 6. Ce qu'on peut remarquer, c'est qu'en dépit de l'agrégation temporelle faite, visant surtout à éliminer les redondances dans les données et de réduire le niveau de détail et surtout l'intermittence, les groupes obtenus contiennent tout de même beaucoup de bruit. Mais malgré ce bruit, des tendances peuvent se voir au niveau des courbes. En effet, Figure6, les groupes (1), (2), (3) et (5) sont caractérisés par une croissance, suivie d'une stabilisation dans le temps. Le groupe (4), en revanche, regroupe les individus dont la taille reste constante pendant une importante durée, et qui a fini par augmenter à la fin. Finalement, le groupe (6), quant à lui, présente des individus dont la taille a augmenté au début, mais qui a par la suite diminué graduellement. Ce qui attire aussi notre attention, c'est que pour les six groupes, des sauts discrets au niveau de la taille ont été observés. De plus, les clients commencent à partir d'un certain seuil à avoir des mensurations plutôt constantes dans le temps. Cela peut être expliqué par le fait qu'au début, les clients ne savent pas trop les tailles adéquates à se procurer pour chaque type de vêtements, mais au terme de deux voire trois années, ils ont tendance à connaitre leurs tailles exactes.

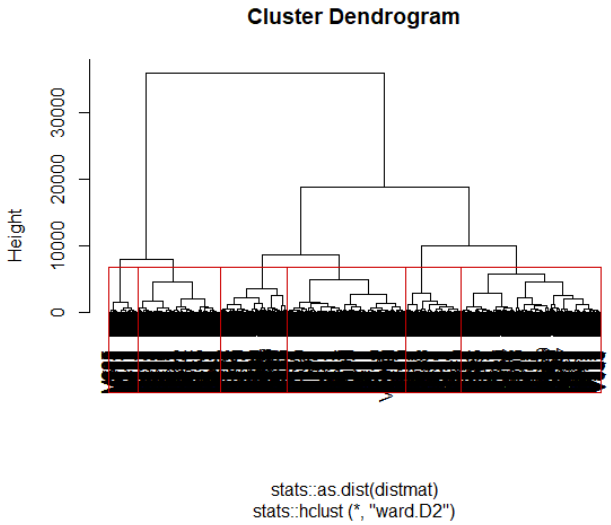

Figure 4. Dendrogramme

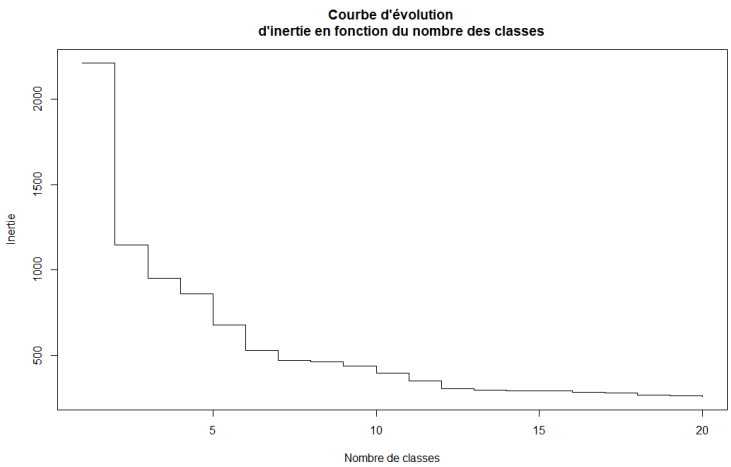

Figure 5. Choix du nombre de segments

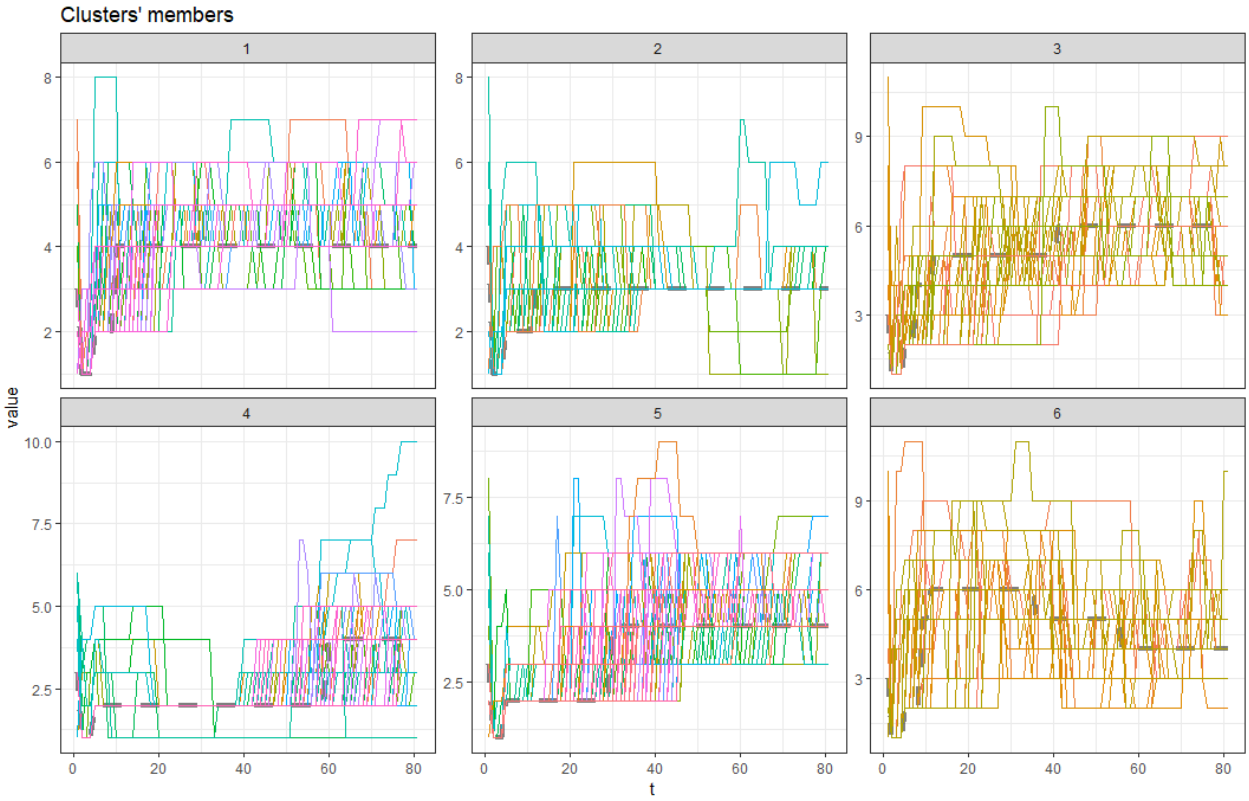

Figure 6. Résultats de la segmentation

Quant aux valeurs de mensurations, on observe qu'à vue globale, les clients appartiennent au même intervalle de mesures.

Par la suite, nous avons appliqué la segmentation sur des individus appartenant au même métier. En revanche, cette fois, nous avons considéré parallèlement les clients de sexe féminin et ceux du sexe masculin. Comparer ces deux groupes, nous permettra de voir s'il y a une différence en comportement entre les deux. Ainsi, similairement à l'exemple précédent, nous avons commencé par une classification hiérarchique ascendante. Les dendrogrammes obtenus sont présentés par les figures 7 et 8. Les courbes de pertes d'inertie sont illustrées par les figures 9 et 10. En nous basant sur ces deux diagrammes et les courbes de pertes d'inertie correspondants, nous avons choisi de segmenter les deux groupes en cinq segments.

Les deux résultats de segmentations concernant le sexe féminin et le sexe masculin sont proposés par les figures 11 et 12 respectivement. Commençant par considérer la première segmentation. Sur la figure 11, on peut voir des groupes où le comportement est assez clair. Par exemple le groupe (5) présente des clients avec un comportement décroissant. Le groupe (2) présente une légère croissance. Les groupes (3) et (4) sont caractérisés par un comportement plutôt constant, mais avec de grands sauts. Le comportement des individus appartenant au groupe (1) est pourtant difficilement observable. Les données y sont beaucoup bruitées. Aussi, alors que la moyenne des valeurs de mensurations des quatre premiers groupes appartiennent au même intervalle, le dernier groupe commence par des valeurs plus importantes, observables aussi sur les sauts discrets des autres groupes.

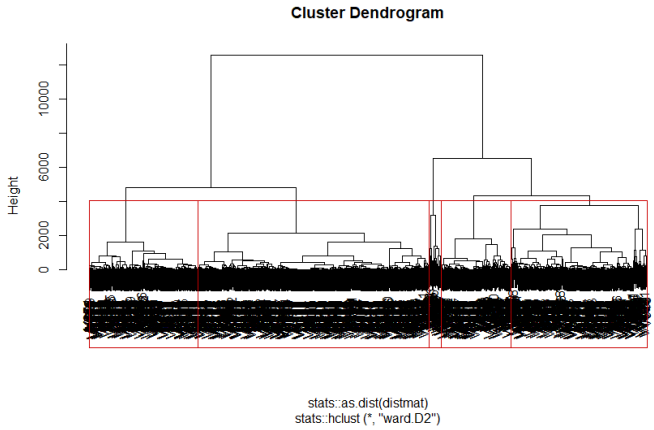

Figure 7. Dendrogramme

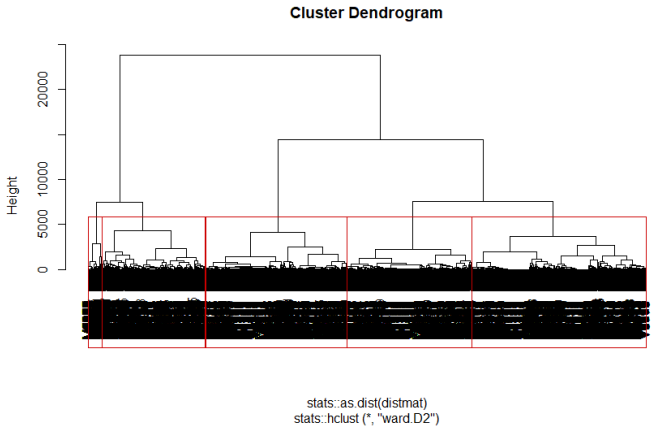

Figure 8. Dendrogramme

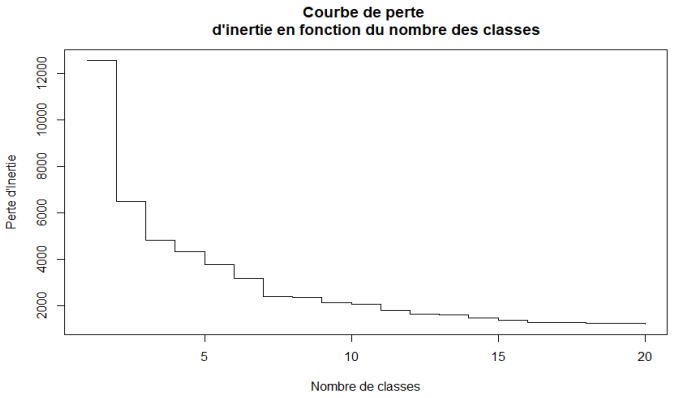

Figure 9. Choix du nombre de segments

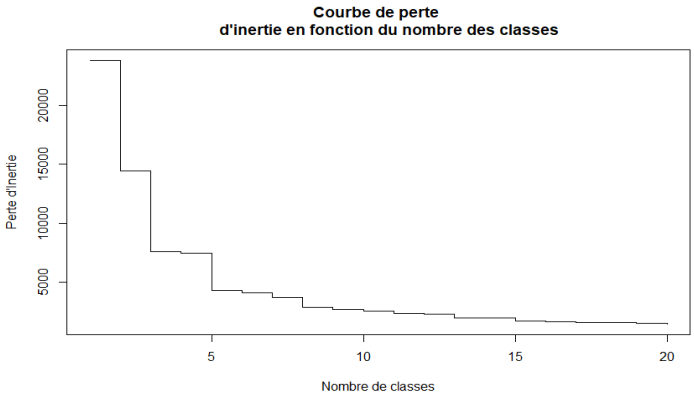

Figure 10. Choix du nombre de segments

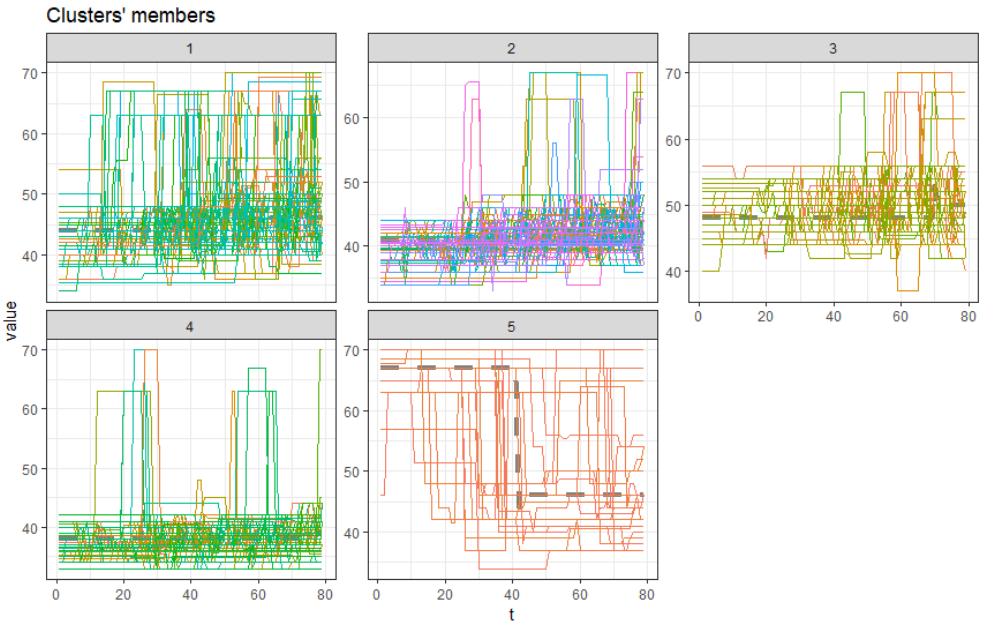

Figure 11. Résultats de la segmentation

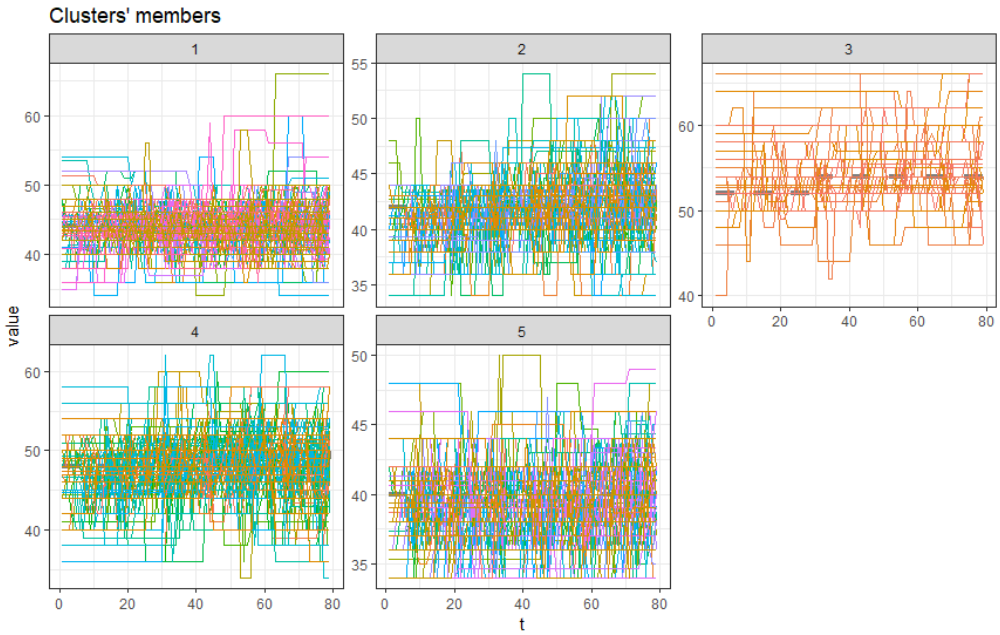

Figure 12. Résultats de la segmentation

Étant donné que nous manipulons des données d'individus appartenant au sexe féminin, ces sauts peuvent s'expliquer par une potentielle grossesse par exemple. D'un autre côté, la segmentation des clients de sexe masculin a donné la figure 12. Malgré le bruit des données, une légère croissance est observable au niveau de tous les groupes. Les séries du groupe (3) à l'encontre, présentent des valeurs plus importantes par rapport aux autres groupes.

Bien que les observations à l'œil nu permettent d'observer les comportements de chacun des groupes, un critère plus précis devra être utilisé pour mettre en évidence la qualité de la segmentation. L'indice de silhouette a été sélectionné dans le cadre de cette étude. Les valeurs de cet indice peuvent varier dans un intervalle [-1,1]. Plus l'indice est proche de 1, meilleure est la segmentation obtenue. Les valeurs du calcul de l'indice de Silhouette pour les trois segmentations effectuées sont groupées dans le tableau 2. Le cas présente la première segmentation effectuée sur tous les individus du même sexe. Le cas 2 et le cas 3 font référence aux segmentations des individus appartenant au même métier, et représentant le sexe féminin et masculin respectivement.

Tableau 2. Résultats de l'évaluation de la segmentation par l'indice de silhouette

	Indice de Silhouette
Cas 1	0.34
Cas 2	0.93
Cas 3	0.97

À partir des résultats présentés dans le tableau 2, il est visible que les segmentations dans le cas 2 et le cas 3 sont meilleures que celle du cas 1, bien que le résultat de cette dernière soit assez acceptable. Ceci montre en fait qu'il est plus avantageux d'effectuer l'analyse sur des groupes plus spécifiques, et donc plus homogènes, ce qui semble assez trivial.

5 CONCLUSION ET PERSPECTIVES

Le « data mining » est aujourd'hui un concept omniprésent dans le domaine industriel. Il est en fait indispensable pour des entreprises qui décident d'exploiter ces données, en général volumineuses. Parmi les outils les plus populaires de l'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.

Ce travail présente les étapes d'une segmentation des clients, en se basant sur leurs données morphologiques. La méthode appliquée dans ce papier s'appuie sur quatre phases : le nettoyage des données, la création des séries temporelles, la correction de l'intermittence des données par agrégation, et finalement la segmentation des clients et l'évaluation des groupes obtenus. Les résultats de la segmentation par classification hiérarchique ascendante sur trois groupes d'individus ont été présentés et évalués à l'aide de l'indice de silhouette. L'application de la segmentation dans notre cas d'études nous a permis de créer des groupes dont l'évolution des mensurations pour un produit défini est similaire et permet de montrer un comportement morphologique qui évolue dans le temps. Mais bien que les indices obtenus soient acceptables, voire bons, l'importance du bruit dans les séries rend difficile d'observer les tendances à travers les courbes. C'est pourquoi nos futures recherches vont se concentrer sur la réduction de ce bruit. Une méthode adéquate de lissage permettra de discerner mieux les tendances et mieux différencier les groupes.

Ce travail présente une étape préliminaire à une étude de prévision de la demande des clients. À travers nos futures recherches, nous allons aborder ce problème, qui consiste en premier lieu à créer des profils de clients, en se basant sur les historiques de mensurations et l'évolution de ces dernières. Ensuite, nous allons essayer de comprendre les corrélations entre chacun des groupes et des facteurs extérieurs tels que l'âge, la zone géographique ou la nature du travail. Finalement, nous allons choisir la méthode adéquate de prédiction, qui nous permettra de visualiser les comportements futurs de chacun des groupes de clients obtenus. Ceci aidera notre partenaire industriel à ajuster ses stratégies logistiques et mieux répondre aux besoins de ses clients.

6 REMERCIEMENTS

Les auteurs expriment leur reconnaissance au partenaire industriel Logistik Unicorp pour sa collaboration au projet et la fourniture des données. Les auteurs soulignent également le support financier de MITACS (projet IT 12058).

7 REFERENCES

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–A decade review. *Information Systems*, *53*, pp. 16-38.

Ahmed, S.R., (2004). Applications of data mining in retail business. International Conference on Information Technology: Coding Computing, ITCC. 2. Vol.2., pp. 455-459.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), pp. 243-256.

Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M. (2006, June). A comparative study for outlier detection techniques in data mining. In *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*, IEEE,pp. 1-6.

Barirani, A., Agard, B., & Beaudry, C. (2013). Competence maps using agglomerative hierarchical clustering. *Journal of Intelligent Manufacturing*, *24*(2), pp. 373-384.

Bartezzaghi, E., Verganti, R., & Zotteri, G. (1999). A simulation framework for forecasting uncertain lumpy demand. *International Journal of Production Economics*, *59*(1-3), pp. 499-510.

Bellanger, L., & Tomassone, R. (2014). *Exploration de données et Méthodes statistiques : data analysis & data mining avec R* ,Ellipses, p. 480.

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.

Berry, M., & Linoff, G. (2004). Data Mining Techniques: for marketing, sales, and customer relationship management, John Wiley & Sons, New York, USA.

Brown, R., (1963). Smoothing. *Forecasting and Prediction of Discrete Time Series*. Prentice-Hall, Englewood Cliffs, NJ.

Chadwick, N. A., McMeekin, D. A., & Tan, T. (2011, May). Classifying eye and head movement artifacts in EEG signals.

In *Digital Ecosystems and Technologies Conference (DEST), 2011 Proceedings of the 5th IEEE International Conference on*, pp. 285-291.

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, *23*(3), pp. 289-303.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Gonzalez, P. L. (2008). Méthodes de classification, [Site web le cnam], (page consultée le 15 Décembre 2018).

Grabisch, M., Marichal, J. L., Mesiar, R., & Pap, E. (2011). Aggregation functions: means. *Information Sciences*, *181*(1), pp. 1-22.

He, L., Agard, B., & Trépanier, M. (2018). 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*, pp. 1-20.

Ifth, (2006). Résultats de la Campagne Nationale de Mensuration, Conférence de presse, salon PRET à PORTER PARIS.

Jin, Y. H., Williams, B. D., Tokar, T., & Waller, M. A. (2015). Forecasting with temporally aggregated demand signals in a retail supply chain. *Journal of Business Logistics*, *36*(2), pp. 199-211.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

Lejeune, M. A. (2001). Measuring the impact of data mining on churn management. *Internet Research*, *11*(5), pp. 375-387.

Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Liu, L. M., Hudak, G. B., Box, G. E., Muller, M. E., & Tiao, G. C. (1992). *Forecasting and time series analysis using the SCA statistical system* (Vol. 1, No. 2). DeKalb, IL: Scientific Computing Associates.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, No. 14, pp. 281-297.

Murray, P. W., Agard, B., & Barajas, M. A. (2018). ASACT-Data preparation for forecasting: A method to substitute transaction data for unavailable product consumption data. *International Journal of Production Economics*, *203*, pp. 264-275.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), pp. 2592-2602.

Peter Langfelder, Bin Zhang, Steve Horvath; Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R, *Bioinformatics*, Volume 24, Issue 5, 1 March 2008, pp. 719–720.

Petropoulos, F., Kourentzes, N., & Nikolopoulos, K. (2016). Another look at estimators for intermittent demand. *International Journal of Production Economics*, *181*, pp. 154-161.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, *26*(7), pp. 1641-1650.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., & Keogh, E. (2013). Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *7*(3), 10.

Rehm, J., & Gmel, G. (2001). Aggregate time-series regression in the field of alcohol. *Addiction*, *96*(7), pp. 945-954.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, *26*(1), pp. 43-49.

Smith, W. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing*, *21*(1), pp. 3-8.

Stolojescu, C. L. (2012). *A Wavelets Based Approach for Time Serie Mining* (Doctoral dissertation, Télécom Bretagne, Université de Bretagne-Sud).

Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. *International journal of production economics*, *71*(1-3), pp. 457-466.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, *252*(1), pp. 1-26.

Tiao, G. C. (1972). Asymptotic behaviour of temporal aggregates of time series. *Biometrika*, *59*(3), pp. 525-531.

Torres Moreno, J. M. (1997). *Apprentissage et généralisation par des réseaux de neurones : étude des nouveaux algorithmes constructifs* (Doctoral dissertation, Grenoble INPG).

Vliegenthart, R. (2014). Moving up. Applying aggregate level time series analysis in the study of media coverage. *Quality & quantity*, *48*(5), pp. 2427-2445.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, *26*(2), pp. 275-309.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), pp. 236-244.

Xu, Rui & Wunsch, Donald. (2005). Survey of Clustering Algorithms. Neural Networks, IEEE Transactions *on Neural Networks*. *16*(3), pp. 645-678.

Zhang, Z., Huang, K., & Tan, T. (2006). Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance Scenes.International Conference on Pattern Recognition, pp. 1135-1138.Zighed, D. A., & Rakotomalala, R. (2002). Extraction de connaissances à partir de données (ECD). *Techniques de l'Ingénieur, H*, *3*, 744.