11e Congrès international de génie industriel – CIGI2015
Québec, Canada
26-28 octobre 2015

Segmentation de données de livraison pour la prévision de la demande des clients

PAUL W. MURRAY[1], LI HE[1], BRUNO AGARD[1], MARCO A. BARAJAS[2]

[1] ÉCOLE POLYTECHNIQUE DE MONTRÉAL & CIRRELT
Montréal, QC, H3C 3A7, Canada
paul.murray@polymtl.ca, li.he@polymtl.ca, bruno.agard@polymtl.ca

[2] FISHERIES & MARINE INSTITUTE OF MEMORIAL UNIVERSITY OF NEWFOUNDLAND
St. John's, NL, Canada
marco.barajas@mi.mun.ca

Résumé **- Les prévisions de la demande sont essentielles pour la gestion des activités de la chaîne d'approvisionnement, mais il est difficile de déterminer ces prévisions lorsqu'une partie de l'information manque. Beaucoup d'outils traditionnels et avancés de prévision sont disponibles, mais leur application à un très grand nombre de clients n'est pas gérable. Dans notre recherche, nous utilisons des techniques d'exploration de données pour identifier des segments de clients avec des comportements similaires de la demande, basés sur leurs données historiques. Une fois les segments de clientèle identifiés, un nombre raisonnable de modèles de prévision peut être construit pour représenter les clients dans chaque segment.**

Abstract **- Demand forecasts are essential for managing supply chain activities but are difficult to create when collaborative information is absent. Many traditional and advanced forecasting tools are available, but applying them to a large number of customers is not manageable. In our research, we use data mining techniques to identify segments of customers with similar demand behaviors. Historical usage is used to cluster customers with similar demands. Once customer segments are identified, a manageable number of forecasting models can be built to represent the customers within the segment.**

Mots clés **- Modèle de données, data mining, segmentation, prévisions, gestion de la demande.**
Keywords **- Data models, data mining, clustering, forecasting, Vendor Managed Inventory.**

1 INTRODUCTION

Les chaînes d'approvisionnement existent depuis le début de l'industrialisation, mais l'étude des chaînes d'approvisionnement est relativement nouveau (Janvier-James, 2012). L'objectif global d'une chaîne d'approvisionnement est de transformer et de transporter des matériaux ou des produits, pour y ajouter de la valeur, et satisfaire une demande à chaque étape du processus (Janvier-James, 2012).

Comprendre la demande en aval est une condition préalable à la réalisation de l'objectif global de la chaîne d'approvisionnement (Carbonneau, Laframboise, & Vahidov, 2008). La chaîne d'approvisionnement traditionnelle où les clients passent des commandes à l'avance, permettant une planification très en amont, n'est plus vraiment répandue; de nos jours, la responsabilité de la gestion et de la prévision des stocks s'est déplacée vers le vendeur. Pour cela, il existe deux principaux modes de gestion : Collaborative Planning Forecasting & Replenishment (CPFR) et Vendor Managed Inventory (VMI) où la responsabilité de la prévision est alors transférée au fournisseur (Holweg, Disney, Holmström, & Småros, 2005).

La prévision de la demande est parfois compliquée par le fait que l'ensemble des partenaires ne partagent pas l'information dont ils disposent (Holweg et al., 2005). (Kembro & Näslund, 2014) font même remarquer que le partage d'information entre les membres d'une chaîne logistique est un moindre mal quand on ne veut/peut pas partager les données de prévision. Cependant, quand le partage d'information n'est pas possible, il reste pour le fournisseur l'analyse des données historiques dont il dispose, pour construire ses modèles de prévision.

Le développement de modèles de prévision, basés sur des historiques, n'est pas une nouveauté, considérons par exemple, les recherches sur l'analyse des séries chronologiques publiées il y a près de 50 ans (Box & Jenkins, 1968). Bien que très performantes, ces méthodes deviennent moins efficaces quand il y a beaucoup de signaux à prédire, ce qui est souvent le cas quand il y a beaucoup d'éléments dans la chaîne d'approvisionnement et que, en plus, les clients finaux, en aval,

présentent une grande variété de modèles d'utilisation, parfois non-linéaire (Aburto & Weber, 2007).

Cet article propose une méthodologie de prévision de la demande, basée sur un historique de vente, à partir d'une segmentation selon les profils d'utilisation des clients.

Le reste de l'article est organisé comme suit: la section suivante est une rapide revue de la littérature sur les modèles CFPR, VMI, la segmentation de la clientèle, et les méthodes de prévision. Sont ensuite présentés la méthodologie, les résultats d'études de cas, et la discussion. Nous terminons avec une conclusion et présentons des suggestions pour les recherches futures.

2 2. REVUE DE LITTÉRATURE

2.1 Prévision collaborative, de la planification et reconstitution

La communication entre les membres de la chaîne d'approvisionnement est bénéfique pour la construction de meilleures prévisions et l'amélioration de la compétitivité (Vachon, Halley, & Beaulieu, 2009). La collaboration est la clé du succès du CPFR, et l'information peut être transmise par une variété de modes, y compris l'échange électronique de données, les prévisions, et l'interaction entre les différentes personnes partenaires de la chaîne d'approvisionnement.

Malgré les avantages du CPFR, la recherche a montré qu'en pratique, la collaboration était peu fréquente (Holweg et al., 2005). Un frein au CPFR est le besoin de technologie de l'information suffisamment sophistiquée pour permettre le partage des données. Le partage de données est pas possible si les systèmes informatiques des partenaires de la chaîne d'approvisionnement sont insuffisants ou incompatibles (Hernández, Mula, Poler, & Lyons, 2014). Lorsque le partage de données électroniques n'est pas possible, le CPFR peut encore être accompli grâce à la préparation manuelle et le partage des informations. La main-d'œuvre et les efforts limitent alors le champ d'application de cette méthode (Holweg et al., 2005). Enfin, un catalyseur important de CFPR est la confiance et les relations personnelles entre les intervenants au sein des organisations partenaires (Wang, Ye, & Tan, 2014), l'absence de relations de confiance entrave le partage de l'information. Le CFPR est un élément important et bénéfique pour la performance de la chaîne d'approvisionnement, cependant, d'autres stratégies peuvent être utilisées pour construire une prévision. Ces méthodes sont discutées dans les sections suivantes.

2.2 Vendor Managed Inventory (VMI)

Avec VMI la responsabilité de la prévision est entièrement transférée du côté du fournisseur. Pour cela, idéalement, le fournisseur a accès aux sources d'information en aval, y compris les taux de consommation, les niveaux de stocks et les prévisions (Achabal, McIntyre, Smith, & Kalyanam, 2000). La recherche a montré que les stratégies de gestion des stocks sont efficaces pour accroître l'efficacité de la chaîne d'approvisionnement, réduire les coûts globaux de la chaîne d'approvisionnement, et augmenter la compétitivité de la chaîne d'approvisionnement (Achabal et al., 2000; Jung, Chang, Sim, & Park, 2005). Malgré un lien direct entre la performance de la chaîne et le partage de l'information (Forslund & Jonsson, 2007), les partenaires sont parfois réticents ou incapables de partager des informations de prévision utiles (Holweg et al., 2005; Kembro & Näslund, 2014). Lorsque l'information collaborative est absente, le fournisseur doit compter sur d'autres méthodes pour assurer sa capacité à satisfaire la demande.

2.3 Segmentation de la clientèle

Dans un environnement VMI où il n'y a pas d'informations de prévision, le fournisseur doit développer son propre modèle avec les informations dont il dispose, généralement ce sont des données historiques. De nombreuses méthodes statistiques existent pour développer des prévisions basées sur des données historiques, il est cependant impossible de tenter de construire des prévisions individuelles pour un grand nombre de demandes de clients uniques. Par conséquent, le regroupement des clients en segments logiques est nécessaire de sorte qu'un petit nombre de modèles de prévision peut être utilisé pour représenter la population totale.

Il existe plusieurs méthodes pour segmenter une population de clients, et il est important de sélectionner avec soin un algorithme de segmentation approprié pour obtenir des résultats précis (Kashwan & Velu, 2013). Une segmentation rudimentaire basée sur la localisation géographique ou le type d'industrie des clients est tentant car relativement facile. Cependant, les caractéristiques de la demande au sein de chaque groupe obtenu risquent d'être très variables et donc peu appropriés au but recherché (Shapiro, 2007). Il existe heureusement plusieurs méthodes de segmentation plus avancées par partitionnement, hiérarchiques, basées sur la densité, basées sur des grilles et/ou à base de modèles (Han & Kamber, 2006). Les méthodes les plus communes sont cependant les méthodes hiérarchiques et par partitionnement (Chakraborty, 2013; Le, Agard, & Deveault, 2009).

Pour la manipulation de très grands ensembles de données, la segmentation par partitionnement est appropriée en raison de besoins plus faibles en puissance de calcul et en mémoire. La méthode de partitionnement la plus commune est la k-moyenne (k-means) qui apparaît fréquemment dans la littérature. Ce n'est pas une technique nouvelle, elle est apparue dans Fisher (1958) et a été développée par MacQueen (1967). Avec k-means, un ensemble de N points de données est regroupé en K grappes centrés sur la moyenne de chaque groupe. Malgré son utilisation fréquente, k-means a néanmoins des limites importantes, notamment le nombre de groupes doit être connu à l'avance, toutes les « formes » ne sont pas prises en compte, importance des points de départ, affectation de manière rigide des points à des groupes individuels (MacKay, 2003).

Les réseaux de neurones artificiels (ANN) sont une alternatives aux méthodes de type k-means, notamment les cartes auto-organisés de Kohonen (SOM) (Altintas & Trick, 2014). SOM a été officiellement développé par Kohonen (Kohonen, 1990) comme alternative à K-means, avec comme avantage que le nombre de clusters n'est pas nécessairement connu à l'avance.

Beaucoup de variantes à K-means et ANN ont été développées pour surmonter certains de leurs défauts respectifs. Celles-ci comprennent l'ajout de la logique floue pour éviter l'affectation rigide des points aux groupes (MacKay, 2003). Dans ces cas, la logique floue permet l'identification de groupes ayant des attributs similaires (Barajas & Agard, 2015) et lorsque combiné avec la segmentation par partition cela donne la flexibilité d'attribuer des points proportionnellement à la taille des groupes. Les variantes sont nommées de différentes manières, selon leur stratégie : k-means gazeuses, k-means floues, et RNA floues…

2.4 méthodes de prévision

Les outils traditionnels de prévision de la demande, basés sur des méthodes statistiques élémentaires et/ou avancées, se sont avérés utiles dans de très nombreux cas (Kuo, 2001). Dans certains cas, des méthodes statistiques plus récentes, telles que les algorithmes génétiques à bases de réseaux flous (GFNN) ont permis d'améliorer les résultats fournis par les méthodes traditionnelles (Kuo, 2001). Ces méthodes ont une mise en œuvre facile, et fournissent des informations utilisables. Cependant, la recherche a montré que les méthodes statistiques simples ont tendance à amplifier l'effet de coup de fouet (Carbonneau et al., 2008).

Compte tenu des contraintes précédentes, les méthodes de séries temporelles telles que la moyenne mobile autorégressive intégrée (ARIMA), et la méthode de Winter ont été développés (Chang, Fan, & Lin, 2011). ARIMA est largement utilisée pour la prévision, mais est moins performante pour les modèles non-linéaires (Pai & Lin, 2005). Les chercheurs ont développé des méthodes plus sophistiquées que des outils statistiques traditionnels, par exemple, Ferbar et al. (2009) a réussi à obtenir de meilleurs résultats avec le lissage exponentiel en incorporant une étape de débruitage en ondelettes. Les résultats obtenus ont montré une amélioration sur le lissage exponentiel, mais n'a pas été comparé avec d'autres outils de prévision.

Les variations de niveaux des ventes historiques peuvent donner une indication du comportement de la demande, mais ils ne permettent pas de faire ressortir les variables exogènes. L'exploration de données avec des algorithmes décisionnels est une approche à la compréhension des données historiques, intégrant les facteurs exogènes, pour finalement produire des résultats compréhensibles et utiles (Kusiak, 2007).

Dans les années 1990, (Leung, 1995) a identifié les réseaux de neurones artificiels (ANN) comme potentiellement appropriés pour la prévision de la demande dans la chaîne d'approvisionnement. L'architecture RNA connue comme perceptron multicouche (MLP) avec rétro-propagation est couramment utilisée (Beccali, Cellura, Lo Brano, & Marvuglia, 2004). Cependant, Chang et al. (2011) soulignent certaines lacunes avec des modèles de retro-propagation. Ils suggèrent qu'une ANN modifiée peut fournir un meilleur résultat. La modification proposée par Chang et al. (2010) consiste à intégrer les algorithmes génétiques dans l'ANN. Les algorithmes génétiques sont aussi parfois utilisés pour configurer la topographie des RNA (Kaylani et al., 2010).

Le problème de l'établissement du niveau désiré de complexité dans un modèle ANN est discuté par Efendigil et al (2009). Dans la recherche effectuée, ils concluent que la complexité du modèle est déterminée par le nombre de couches cachées, qui pourrait être déterminées de manière heuristique. De plus, ils utilisent la logique floue dans un ANN pour surmonter les limitations de l'utilisation des outils mathématiques classiques. Le système d'inférence floue basé sur un réseau adaptatif (ANFIS) a ainsi été utilisé. L'avantage de ANFIS est qu'il permet d'intégrer les connaissances que l'utilisateur possède sur les entrées et sorties (Jang, 1993). L'inclusion d'une connaissance *a priori* permet d'influencer l'orientation du modèle ANN plutôt que de laisser le modèle fonctionner sans surveillance.

Une autre modification à la prévision avec RNA qui apparaît fréquemment dans la littérature consiste à intégrer une composante floue (Chang et al., 2011; Efendigil et al., 2009; Kuo, 2001; Neto, da Costa Junior, Bitar, & Junior, 2011). La logique floue est alors utilisée pour améliorer les processus de prise de décision en évitant la contrainte de oui / non des décisions binaires (Barajas & Agard, 2014).

Alors que les modèles hybrides ont reçu beaucoup d'attention dans la littérature, Pai et Lin (2005) ont utilisé une approche différente. Plutôt que de combiner RNA et la logique floue, ils ont combiné les machines à support vectoriel (SVM) avec ARIMA. Dans leur recherche, Pai et Lin ont constaté que le modèle hybride donne de meilleurs résultats que SVM ou ARIMA. Cependant, leur modèle n'a pas été comparé à d'autres outils de prévision.

3 METHODOLOGIE

La revue de la littérature montre que l'élaboration de prévisions sans information collaborative est possible, et que les méthodes hybrides peuvent fournir une précision améliorée par rapport aux méthodes traditionnelles. Cependant, quelle que soit la méthode utilisée, les données doit être suffisamment complètes et précises avant de construire un modèle de prévision. Ici, nous développons une méthode dans laquelle nous disposons de données complètes et précises, mais en très grande quantité, en tirant parti des progrès réalisés dans la segmentation de la clientèle et dans la prévision.

Notre méthodologie comprend trois étapes (figure 1): le prétraitement des données, la segmentation et la prévision (Murray, Agard, & Barajas, 2015).

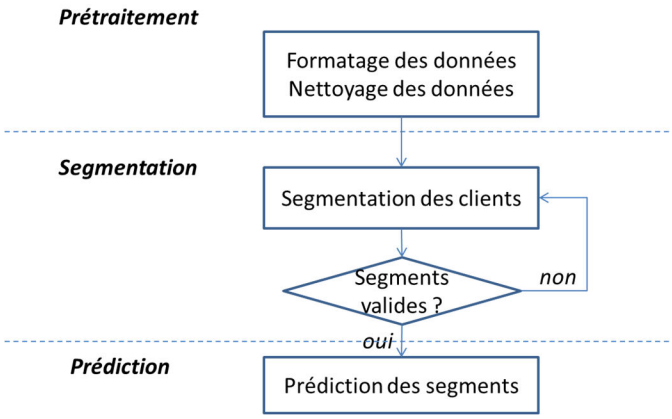

Figure 1 : Méthodologie

Chaque étape sera détaillée dans les prochains paragraphes.

3.1 prétraitement

Avant de traiter les données industrielles pour une éventuelle prévision, il est nécessaire de s'assurer de la qualité de ces données (Kantardzic, 2011). Dans notre étude de cas, le fournisseur était en mesure de fournir quatre années d'activités de livraison, y compris les dates de livraison, les quantités, les emplacements et le code d'activité industriel du client (CTI).

Bien que les données aient été extraites à partir d'activités réelles de livraisons, un certain nombre de correctifs sur les informations a été nécessaire (dans notre cas, par exemple, différentes activités administratives telles que des remboursements, qui représentent des quantités livrées négatives ou de fausses pointes de livraisons, ou alors des clients qui changent de noms, identifiés cette fois par de fausses périodes de zéro utilisation, …).

Lors du nettoyage des données, il faut être prudent pour bien retirer un maximum de signaux aberrants, tout en ne retirant pas de précieuses informations. Pour faciliter cela, une connaissance approfondie du problème est nécessaire, et de nombreux aller-retours avec les experts de l'industrie sont effectués. Pour cela, les clients ont été évalués sur une base annuelle, les valeurs aberrantes ont été identifiées par plusieurs méthodes statistiques simples. Les clients dont la consommation totale était presque nulle ou négative ont été enlevés. Les clients avec des pointes en quantités ont été identifiés et éliminés en évaluant les médianes et écarts standards dans les quantités. Pour tous les autres clients « aberrants » une validation a été effectuée avec le partenaire pour la décision à adopter.

3.2 Segmentation

Les clients ont été segmentés en utilisant le programme informatique R (R Core Team, 2014) en utilisant différents méthodes de calcul de distance et de segmentation. La segmentation a été basée sur le volume mensuel de produit livré et calculé sur la base de différentes métriques. Plusieurs nombre de groupes K ont été essayés (Kantardzic, 2011), les résultats ont été évalués, et plusieurs itérations de segmentations ont été effectuées.

3.3 Prévision

Une fois que les clients ont été convenablement segmentés, les modèles de prévision sont appliqués pour modéliser chaque segment. La création de segments de clientèle basés sur le comportement de consommation permet d'utiliser des méthodes de prévision simples adaptées à chaque segment. Par exemple, des segments avec effet saisonnier sont traités différemment de ceux avec d'autres types de comportements.

Dans le présent article, l'étape 2 : segmentation est détaillée. Les modèles de prévision seront la prochaine étape à effectuer.

4 CAS D'ÉTUDE

4.1 Contexte

Le contexte de l'étude de cas est un fournisseur de matériaux en vrac (matière première X) utilisés dans une variété de processus de fabrication notamment différentes productions industrielles, l'alimentation et les services hospitaliers. Seulement aux États-Unis (cas d'étude) il livre plusieurs milliers de clients partout à travers le territoire.

Le fournisseur est responsable de la gestion des stocks de X chez chacun de ses clients. Il doit s'assurer que chacun a suffisamment de X à chaque instant. Aucune rupture de stock n'est envisageable, pour des raisons de sécurité notamment. Le fournisseur rempli les réserves de chaque client, et le client se sert dans ses réserves.

Pour le fournisseur, il s'agit d'évaluer au mieux les tendances, la saisonnalité, et les profils d'utilisation de chaque client pour développer un meilleur modèle de prévision. Les profils d'utilisation des clients sont influencés par différent facteurs internes et externes, tels que le type d'industrie, l'emplacement géographique du client, la saisonnalité de la demande, et d'autres facteurs pris en considération.

Dans certains secteurs, par exemple la pêche, la consommation a des motifs saisonniers, dans d'autres, par exemple de fabrication aérospatiale, elle est plutôt liée à des indicateurs macro-économiques, elle peut aussi être assez stable en croissante (secteurs hospitaliers).

Le transport et le stockage de X exige des installations spécialisées coûteuses et pas rapidement disponibles. Par conséquent, une variation de la capacité de stockage chez le client n'est pas une option.

Dans notre étude de cas, le fournisseur est responsable de maintenir un approvisionnement ininterrompu au point d'utilisation de chaque client. Des prévisions de la demande à court terme sont générées en fonction de l'historique d'utilisation; la demande et les livraisons peuvent être déclenchées par des données télémétriques obtenues à partir de capteurs de niveau au point de l'utilisation des réservoirs de stockage (lorsque disponible). Un engagement à assurer un inventaire ininterrompu sur les sites des clients et le manque de données de prévisions à moyen et à long terme oblige le fournisseur à être conservateurs avec les décisions futures de capacité.

Les données de chaque livraison sont enregistrées : le client, la date, l'emplacement et la quantité (Cf fig 2). Le nombre d'enregistrements par année est supérieur à 200.000. Le fournisseur souhaite améliorer ses modèles de prévision, en considérant les impacts des facteurs externes tels que l'emplacement, de la saison, l'industrie de l'utilisateur, et le climat sur la demande du produit X.

Date de livraison	client	adresse	volume
jour 1	1	18 chemin…	15 236
jour 1	2	5 rue …	14 250
jour 1	3	6 impasse …	8 183
jour 2	1	18 chemin…	4 360
…			
jour n	c	23 rue …	56 874

Figure 2 : données brutes

4.2 Prétraitement des données

Nous disposons de quatre ans d'histoire avec près d'un million d'observations d'événements de livraison. Un premier prétraitement des données de livraisons vise à représenter les livraisons mensuelles par client, résultant en environ 3500 vecteurs de séries chronologiques (Cf figure 3).

client	adresse	janvier	fevrier	…	decembre
1.année 1	18 chemin…	1 543 625	1 453 620		850 698
2.année 1	5 rue …	584 362	356 284		365 254
3.année 1	6 impasse …	52 634	0		25 635
…					
1.année 2	18 chemin…	1 265 897	1 587 458		956 325
…					
c.année 4	23 rue …	3 568 965	3 255 698		2 569 845

Figure 3 : données prétraitées

Premier biais, les données disponibles représentent les livraisons du fournisseur, pas la consommation par le client. En fonction de la capacité de stockage et de la consommation à chaque point d'utilisation, le fournisseur optimise son calendrier de livraison. Cela lui permet de minimiser ses coûts de livraison en optimisant la gestion de ses camions de livraison, mais cela biaise l'information disponible, qui n'est plus vraiment la consommation réelle du client.

4.3 Segmentation

On cherche maintenant à obtenir des groupes de clients qui partagent le même profil d'utilisation.

Compte tenu du nombre assez élevé de données, en première approche, nous essayons la segmentation par k-moyenne (rapidité d'exécution) d'autant plus que le nombre de groupes peut être contrôlé par l'utilisateur.

Première difficulté : comment identifier la distance entre deux clients ?

Pour simplifier, considérons non pas le volume livré, mais le fait d'avoir été livré Dans le tableau figure 4, 13 clients sont livrés (1) ou pas (0) dans 7 périodes de temps (V1 à V7).

	V1	V2	V3	V4	V5	V6	V7
1	**1**	0	0	0	0	0	0
2	0	**1**	0	0	0	0	0
3	0	0	**1**	0	0	0	0
4	0	0	0	**1**	0	0	0
5	0	0	0	0	**1**	0	0
6	0	0	0	0	0	**1**	0
7	0	0	0	0	0	0	**1**
8	**1**	**1**	0	0	0	0	0
9	**1**	0	**1**	0	0	0	0
10	0	**1**	**1**	0	0	0	0
11	**1**	0	0	0	**1**	0	0
12	0	0	0	0	**1**	**1**	0
13	0	0	0	0	0	**1**	**1**

Figure 4 : livraisons simplifiées

D'un point de vue « ingénieur », voici quelques caractéristiques (A) que nous aimerions faire ressortir :

- les clients 1 et 2 se ressemblent,
- les clients 2 et 3 se ressemblent, etc …
- 1 est plus proche de 2 que de 3 (livraisons en début de période mais espacées).
- 1 ne ressemble pas, ou peu à 6 ou 7 (livraison en fin de période).
- 1 ressemble à 8 (1 livraison ou 2, mais un début de période)
- 8 ressemble à 1, 2, 9 et 10 mais pas à 11 ou 12…

Afin de capturer ces « ressentis », différentes métriques sont testées. Nous calculons les distances 2 à 2 entre tous les clients :

a) Distance euclidienne

La distance euclidienne est simple à comprendre, et elle est couramment utilisée lorsque l'on cherche la distance entre deux vecteurs V1 et V2 (de coordonnées respectifs $x1_i$ et $x2_j$):

$$d(V1,V2) = \sqrt{\sum_{i=1}^{n}(x1_i - x2_i)^2}$$

Avec les données de la figure 4, obtenons les résultats suivants (figure 5)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1,41	1,41	1,41	1,41	1,41	1,41	1	1	1,73	1	1,73	1,73
2	1,41	0	1,41	1,41	1,41	1,41	1,41	1	1,73	1	1,73	1,73	1,73
3	1,41	1,41	0	1,41	1,41	1,41	1,41	1,73	1	1	1,73	1,73	1,73
4	1,41	1,41	1,41	0	1,41	1,41	1,41	1,73	1,73	1,73	1,73	1,73	1,73
5	1,41	1,41	1,41	1,41	0	1,41	1,41	1,73	1,73	1,73	1	1	1,73
6	1,41	1,41	1,41	1,41	1,41	0	1,41	1,73	1,73	1,73	1,73	1	1
7	1,41	1,41	1,41	1,41	1,41	1,41	0	1,73	1,73	1,73	1,73	1,73	1
8	1	1	1,73	1,73	1,73	1,73	1,73	0	1,41	1,41	1,41	2	2
9	1	1,73	1	1,73	1,73	1,73	1,73	1,41	0	1,41	1,41	2	2
10	1,73	1	1	1,73	1,73	1,73	1,73	1,41	1,41	0	2	2	2
11	1	1,73	1,73	1,73	1	1,73	1,73	1,41	1,41	2	0	1,41	2
12	1,73	1,73	1,73	1,73	1	1	1,73	2	2	2	1,41	0	1,41
13	1,73	1,73	1,73	1,73	1,73	1	1	2	2	2	2	1,41	0

Figure 5 : distances euclidiennes entre les clients de la figure 4

Le client 1 est alors à égale distance des clients 8, 9, et 11. Tous les clients 1 à 7 sont équidistants… et plein d'autres caractéristiques qui ne capturent pas les caractéristiques souhaitées (A).

Le dendrogramme (figure 6) montre la proximité des différents clients calculés à la figure 5. On observe un mélange de profils de clients très différents (les commandes à différentes périodes de temps sont regroupées).

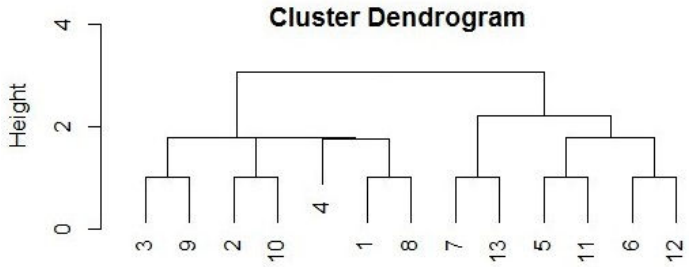

Figure 6 : Segmentation hiérarchique avec distances euclidiennes entre les clients de la figure 4

Cette métrique ne permet pas d'obtenir les caractéristiques escomptées.

b) Corrélation croisée (Mori, Mendiburu, Álvarez, & Lozano, 2014)

La corrélation croisée entre deux séries V1 et V2 (de coordonnées respectifs $x1_i$ et $x2_j$) se calcule de la manière suivante :

$$d_{CC}(V1,V2) = \sqrt{\frac{(1 - CC_0(V1,V2)}{\sum_{k=1}^{max} CC_k(V1,V2)}}$$

Avec

$$CC_k(V1,V2) = \frac{\sum_{i=0}^{n-1-k}(x1_i - \overline{x1})\,(x2_{i+k} - \overline{x2})}{\sqrt{(x1_i - \overline{x1})^2}\sqrt{(x2_{i+k} - \overline{x2})^2}}$$

Où k est le décalage (*lag*) maximum permis entre les deux « signaux » V1 et V2.

Cette fois-ci nous obtenons les résultats de la figure 7.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1,72	6,72	6,72	6,72	6,72	5,88	1,35	1,98	2,53	1,98	4,25	4,15
2	1,72	0	1,69	5,39	5,39	5,39	6,72	1,47	1,69	1,41	2,45	3,41	3,93
3	6,72	1,69	0	1,69	5,39	5,39	6,72	2,45	1,81	1,41	3,93	3,41	3,93
4	6,72	5,39	1,69	0	1,69	5,39	6,72	3,93	2,45	2,31	2	2,31	3,93
5	6,72	5,39	5,39	1,69	0	1,69	6,72	3,93	3,93	3,41	1,81	1,41	2,45
6	6,72	5,39	5,39	5,39	1,69	0	1,72	3,93	3,93	3,41	2,45	1,41	1,47
7	5,88	6,72	6,72	6,72	6,72	1,72	0	4,15	4,15	4,25	4,15	2,53	1,35
8	1,35	1,47	2,45	3,93	3,93	3,93	4,15	0	2,38	1,62	2,78	2,41	2,6
9	1,98	1,69	1,81	2,45	3,93	3,93	4,15	2,38	0	2,81	3,5	2,41	2,6
10	2,53	1,41	1,41	2,31	3,41	3,41	4,25	1,62	2,81	0	2,97	2,09	2,41
11	1,98	2,45	3,93	2	1,81	2,45	4,15	2,78	3,5	2,97	0	3,56	3,37
12	4,25	3,41	3,41	2,31	1,41	1,41	2,53	2,41	2,41	2,09	3,56	0	1,62
13	4,15	3,93	3,93	3,93	2,45	1,47	1,35	2,6	2,6	2,41	3,37	1,62	0

Figure 7 : corrélations croisée entre les clients de la figure 4

Les caractéristiques (A) sont respectées. Le dendrogramme (figure 8) permet d'obtenir des regroupements qui ont du sens pour la suite de l'étude. On obtient ici deux groupes indépendants : 7, 13, 6, 5 et 12 d'un côté, le reste dans un second groupe.

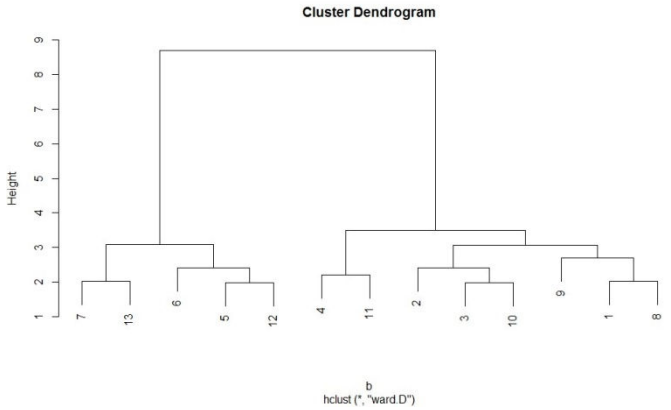

Figure 8 : Segmentation hiérarchique avec corrélations croisée entre les clients de la figure 4

Cette fois les caractéristiques recherchées sont obtenues. De nombreuses autres méthodes, notamment DTW (Dynamic Time Warping) ont été expérimentées, certaines donnent des résultats semblables, mais avec des besoins en calculs bien supérieurs. Sur les données réelles, très nombreuses, nous utiliserons donc la corrélation croisée.

4.4 Prévision

La segmentation des données brutes, nous permet maintenant d'obtenir des groupes de comportement similaires pour des ensembles de clients. La prochaine étape consiste maintenant à choisir un modèle de prévision adapté à chaque groupe obtenu. Pour ensuite faire de la prévision, client par client, à partir d'un nombre limité de modèles, compréhensibles par l'utilisateur.

Sur le cas industriel, à partir d'un échantillon de données (données de la Californie seulement), nous obtenons différents type de groupes, comme celui présenté (figure 9). Les éléments d'un même groupe partagent un même profil de demande qui sera utile pour la prévision.

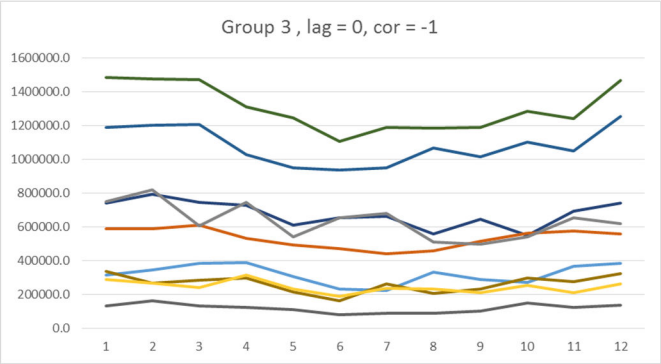

Figure 9 : exemple de segments de client obtenu

Par exemple, sur les données de consommation réelles de différents clients sur une période de 12 mois, obtenues dans le groupe 3 (figure 9), il est facile de trouver un modèle de prévision qui prenne en compte la saisonnalité.
Il reste maintenant à choisir un modèle de prévision pour chaque groupe de comportement et de le calibrer pour chaque client.

5 CONCLUSIONS ET PERSPECTIVES

Ce document a présenté une méthode pour transposer un nombre élevé de clients individuels dans un petit nombre de groupes de clients ayant des comportements de la demande similaire.
La prévision de la demande pour les groupes est une tâche gérable. De nombreuses stratégies de prévision utilisent des outils statistiques pour prédire la demande future des clients ou des produits spécifiques. Dans le cadre de notre étude de cas, il est impossible de tenter de construire des prévisions spécifiques au client en raison du grand nombre de clients. Nous avons résolu ce problème en segmentant les clients en fonction de leurs comportements. Nous sommes alors en mesure de construire un nombre gérable de modèles de prévision et de les appliquer au sein de chaque segment de clientèle.
L'état de l'art montre une quantité importante de recherches sur les méthodes de segmentation de données. Cependant le prétraitement des données est encore une problématique très importante (Kantardzic, 2011; Miller, 2009; Witten, Farnk, & Hall, 2011) et n'as pas été discutée en détail. Une prochaine étape visera à identifier le meilleur modèle de données pour obtenir un niveau de prévision souhaité.
Au cours de cette étude, plusieurs techniques de classification différentes, tels que K-means, classification hiérarchique, cartes auto-organisatrices (SOM), et la classification floue ont été explorées, quel modèle est le plus représentatif pour le fournisseur ?
Une fois que des groupes de clients ont été établis, ils peuvent être résumés dans un modèle de comportement unique. Cependant, un même client peut, en fonction des années être dans différents groupes. Quel(s) modèle(s) de prévision utiliser ? Comment ?
Comprendre la cause d'un changement global de la demande permet au fournisseur de décider si un changement dans la capacité est nécessaire en raison de tendances, ou si le changement est dû à un effet saisonnier et aucun changement de capacité ne serait nécessaire.

Dans l'étude de cas, les clients ayant un comportement de la demande qui ne suivent aucun modèle ont été traités comme des valeurs aberrantes et retirés. Des recherches supplémentaires sont nécessaires pour développer une méthode pour classer ces clients et les réincorporer dans le modèle.

6 REFERENCES

Aburto, L., & Weber, R., (2007) Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing,* 7(1), pp. 136-144.

Achabal, D. D., McIntyre, S. H., Smith, S. A., & Kalyanam, K., (2000) A decision support system for vendor managed inventory. *Journal of retailing,* 76(4), pp. 430-454.

Altintas, N., & Trick, M., (2014) A data mining approach to forecast behavior. *Annals of Operations Research,* 216(1), pp. 3-22.

Barajas, M., & Agard, B., (2015) A methodology to form families of products by applying fuzzy logic. *International Journal on Interactive Design and Manufacturing (IJIDeM)* online.

Beccali, M., Cellura, M., Lo Brano, V., & Marvuglia, A., (2004) Forecasting daily urban electric load profiles using artificial neural networks. *Energy Conversion and Management,* 45(18-19), pp. 2879-2900.

Boucheham, B., (2010) Reduced data similarity-based matching for time series patterns alignment. *Pattern Recognition Letters,* 31(7), pp. 629-638.

Box, G. E. P., & Jenkins, G. M., (1968) Some Recent Advances in Forecasting and Control. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 17(2), pp. 91.

Carbonneau, R., Laframboise, K., & Vahidov, R., (2008) Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research,* 184(3), pp. 1140-1154.

Chakraborty, G. (2013) Customer segmentation using SAS Enterprise Miner, SAS Institute Inc.: Cary, NC.

Chang, P.-C., Fan, C.-Y., & Lin, J.-J., (2011) Monthly electricity demand forecasting based on a weighted evolving fuzzy neural network approach. *International Journal of Electrical Power & Energy Systems,* 33(1), pp. 17-27.

Efendigil, T., Önüt, S., & Kahraman, C., (2009) A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. *Expert Systems with Applications,* 36(3, Part 2), pp. 6697-6707.

Ferbar, L., Čreslovnik, D., Mojškerc, B., & Rajgelj, M., (2009) Demand forecasting methods in a supply chain: Smoothing and denoising. *International Journal of Production Economics,* 118(1), pp. 49-54.

Fisher, W., (1958) On grouping for maximum homogeneity. *American Statistical Association Journal*.

Forslund, H., & Jonsson, P., (2007) The impact of forecast information quality on supply chain performance. *International Journal of Operations & Production Management,* 27(1), pp. 90-107.

Han, J., & Kamber, M. (2006) Data Mining, Southeast Asia Edition: Concepts and Techniques, Morgan kaufmann.

Hernández, J. E., Mula, J., Poler, R., & Lyons, A. C., (2014) Collaborative Planning in Multi-tier Supply Chains Supported by a Negotiation-Based Mechanism and Multi-agent System. *Group Decision and Negotiation,* 23(2), pp. 235-269.

Holweg, M., Disney, S., Holmström, J., & Småros, J., (2005) Supply Chain Collaboration:: Making Sense of the Strategy Continuum. *European Management Journal,* 23(2), pp. 170-181.

Jang, J.-S. R., (1993) ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man and Cybernetics,* 23(3), pp. 665-685.

Janvier-James, A. M., (2012) A New Introduction to Supply Chains and Supply Chain Management: Definitions and Theories Perspective. *International Business Research,* 5(1), pp. 194-207.

Jung, S., Chang, T., Sim, E., & Park, J., (2005) Vendor Managed Inventory and Its Effect in the Supply Chain Systems Modeling and Simulation: Theory and Applications: Springer Berlin Heidelberg.

Kantardzic, M. (2011) Data mining : concepts, models, methods, and algorithms, Second edition Ed., Wiley-IEEE Press: Hoboken, NJ.

Kashwan, K. R., & Velu, C. M., (2013) Customer Segmentation Using Clustering and Data Mining Techniques. *International Journal of Computer Theory and Engineering,* 5(6), pp. 856-861.

Kaylani, A., Georgiopoulos, M., Mollaghasemi, M., Anagnostopoulos, G. C., Sentelle, C., & Mingyu, Z., (2010) An Adaptive Multiobjective Approach to Evolving ART Architectures. *Neural Networks, IEEE Transactions on,* 21(4), pp. 529-550.

Kembro, J., & Näslund, D., (2014) Information sharing in supply chains, myth or reality? A critical analysis of empirical literature. *International Journal of Physical Distribution & Logistics Management,* 44(3), pp. 179-200.

Kohonen, T., (1990) The self-organizing map. *Proceedings of the IEEE,* 78(9), pp. 1464-1480.

Kuo, R., (2001) A sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. *European Journal of Operational Research,* 129(3), pp. 496-517.

Kusiak, A., (2007) Data mining in industrial applications and innovation. *ICS News* pp. 17-21.

Le, T., Agard, B., & Deveault, S. (2009). Application du data mining à la segmentation du marché des meubles aux États-Unis. *Eighth Congres International de Genie Industriel*

Leung, H. C. (1995). Neural networks in supply chain management. *Engineering Management Conference, 1995. Global Engineering Management: Emerging Trends in the Asia Pacific., Proceedings of 1995 IEEE Annual International*

MacKay, D., (2003) An example inference task: clustering. *Information Theory, Inference and Learning Algorithms* pp. 284-292.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*

Miller, H. J. H. J. (2009) Geographic data mining and knowledge discovery, CRC: Boca Raton, Fla.

Mori, U., Mendiburu, A., Álvarez, M., & Lozano, J. A., (2014) A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transportmetrica A: Transport Science*(ahead-of-print), pp. 1-39.

Murray, P., Agard, B., & Barajas, M. (2015). Forecasting supply chain demand by clustering customers. *INFORM 2015*, Ottawa, Canada

Neto, J. C. d. L., da Costa Junior, C. T., Bitar, S. D. B., & Junior, W. B., (2011) Forecasting of energy and diesel consumption and the cost of energy production in isolated electrical systems in the Amazon using a fuzzification process in time series models. *Energy Policy,* 39(9), pp. 4947-4955.

Pai, P.-F., & Lin, C.-S., (2005) A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega,* 33(6), pp. 497-505.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Shapiro, J. F. (2007) Modeling the supply chain, 2nd Ed., Thompson Brooks/Cole: Belmont, CA.

Vachon, S., Halley, A., & Beaulieu, M., (2009) Aligning competitive priorities in the supply chain: the role of interactions with suppliers. *International Journal of Operations & Production Management,* 29(4), pp. 322-340.

Wang, Z., Ye, F., & Tan, K. H., (2014) Effects of managerial ties and trust on supply chain information sharing and supplier opportunism. *International Journal of Production Research,* 52(23), pp. 7046.

Witten, I., Farnk, E., & Hall, M. (2011) Data Mining, Third edition Ed., Morgan Kaufmann / Elsevier Inc.: Burlington, MA.