Market segmentation through data mining: A method to extract behaviors from a noisy data set

Paul W. Murray, Bruno Agard, Marco A. Barajas

École Polytechnique de Montréal, Montréal, QC H3C 3A7, Canada
Fisheries & Marine Institute of Memorial University of Newfoundland, St. John’s, NL, Canada

Abstract

Strategic business planning requires forecasted information that contains a sufficient level of detail that reflects trends, seasonality, and changes while also minimizing the level of effort needed to develop and assess the forecasted information. The balance of information is most often achieved by grouping the customer population into segments; planning is then based on segments instead of individuals.

Ideally, separating customers into segments uses descriptive variables to identify similar behavior expectations. In some domains, however, descriptive variables are not available or are not adequate for distinguishing differences and similarities between customers. The authors solved this problem by applying data mining methods to identify behavior patterns in historical noisy delivery data. The revealed behavior patterns and subsequent market segmentation are suitable for strategic decision-making. The proposed segmentation method demonstrates improved performance over traditional methods when tested on synthetic and real-world data sets.

Keywords: Market segmentation, Time-series clustering, Delivery transactions, Demand behavior, Data mining, Pattern extraction

1. Introduction

Strategic planning relies on the ability to gather information regarding the customer base and being able to process that information into predictions of future requirements. In situations where the customer base is large, a method to reduce the number of possible scenarios is necessary. Scenario reduction is commonly achieved in business through a process known as market segmentation. Early research has demonstrated that forecasting groups of customers (or products) is more accurate than aggregated individual forecasts due to the positive effects of error smoothing and error cancelation (Armstrong, 1985; Dangerfield & Morris, 1992). Market segmentation was first proposed by Smith (1956) as a method of differentiating customers based on the customers’ individual preferences and desires. However, despite extensive literature regarding market segmentation, there is little guidance to managers on how to accomplish it (Dibb & Simkin, 2001; Hung & Tsai, 2008).

The traditional segmentation methods found in the literature rely on identification or creation of a set of descriptive variables from which the differences (distances) between customers are calculated, for example, see (Calvet, Ferrer, Gomes, Juan, & Masip, 2016). These methods calculate distances between members based on attribute-based variables or feature-based variables. An attribute-based distance relies on descriptive attributes such as sex, age, location, industry, or size. Attributes are sometimes converted into numerical variables and organized into standard forms prior to processing (Kantardzic, 2011). With feature-based distance, variables are created based on statistical features such as median, kurtosis, sum, or purchase frequency, from historical data. The features are weighted, summarized, and normalized to become the variables from which distance is calculated (Y. Chen, Zhang, Hu, & Wang, 2006).

Creating variables, from either attributes or features benefits the data analysis in two key areas. First, high variance and outliers are reduced or eliminated by condensing the data into a predetermined set of variables. Second, the resulting dataset is much smaller than the original data and algorithms run much faster. Traditional segmentation methods share two fundamental assumptions: the first assumption is that the data-munging process used to create, modify, and weight the data has resulted in a set of variables that truly reflect the intentions and behaviors of the customers. The second assumption is that the information that describes the demand behavior is present within the available data (Longbign, 2014; Verdu, Garcia, Senabre, Marin, & Franco, 2006). If the assumptions are incorrect, the variables will not adequately describe the customers and the analysis will not produce robust results.
When descriptive variables are not available and cannot be adequately created from the available data, the traditional segmentation methods do not work. We propose to solve this problem by identifying behavior patterns directly from the historical delivery data and then using those patterns to segment the market. A major challenge with this approach is that delivery data can appear noisy when the delivery frequencies and quantities are driven partly by consumption demand and partly by transportation logistics decisions. The noise effect is especially evident when a point-of-use inventory exists between the delivery point and the consumption point. The level of noise in the data increases as the point of data recording moves up the supply chain (Chen, Ryan, & Simchi-Levi, 2000).

The scientific relevance of this work is as follows: A method for identifying market segments in the absence of descriptive variables is developed and tested; market segments are based on consumption behavior patterns which are extracted from noisy delivery data. Testing the new market segmentation method on synthetic data demonstrates that the method can identify behavior patterns from a noisy data set and create useful groups based on those patterns. The application of the method to real-world data generates results that are useful in strategic planning applications.

The remainder of the paper is organized as follows: Section 2 presents a review of the state of the art. Section 3 presents the proposed method. Sections 4 and 5 apply the proposed method to both synthetic and real-world data and compare the proposed method to traditional segmentation method. The paper concludes in Section 6 with a discussion of the scientific contribution, limitations of the research, and suggestions for future research.

2. Literature review

2.1. Market segmentation

Market segmentation is a topic with both a long history and active contemporary investigation; its benefits were identified in a seminal paper by Smith (1956). Market segmentation for planning, marketing, and forecasting has become nearly universal practice in business; in fact, Bain & Company’s “Management Tools & Trends 2015” listed it as one of the top ten executive management tools globally (Rigby & Bilodeau, 2015). The importance of market segmentation can be traced to guidance given by Aristotle to tailor your message to the audience. A more recent example is General Motors Corporation’s approach to build “a car for every purpose and purse” (Gann, 1996). Armstrong (1985) found that forecasting segments rather than individual customers resulted in higher accuracy due to the reduced effect of outliers and irregularities in the data. Regardless of the specific techniques and steps used, all segmentation methods follow the same essential sequence: information or data relating to the customers is gathered, differences and similarities between customers are determined, and finally, the customer base is divided into groups with similar behavior. Fig. 1 illustrates five segmentation strategies, including a priori, key attributes, descriptive attributes, statistical features, and behavior model. The segmentation strategies are ordered according to the level of rigor and analytical effort of each strategy. At one end, the a priori method is purely qualitative; its results depend entirely on the knowledge of the practitioner. At the other end, the behavior extraction methods are rigorous enough that they can theoretically be applied by different practitioners in different domains and produce satisfactory results. The following subsections will develop the relevant segmentation strategies in more detail.

![Fig. 1. Segmentation strategies.](image-url)
2.1.1. A priori segmentation

The a priori method, followed by Aristotle and General Motors, relies on the analyst instinctively knowing how to separate the groups (Gann, 1996). The flexibility of qualitative decision making makes a priori segmentation useful when the available data does not consistently support the observed results (Randle & Dolnicar, 2009). A priori method can be fast and when the analyst’s knowledge is sufficient, it can be very accurate. However, since the a priori approach is entirely qualitative and depends on the level of knowledge the analyst has about the customer base; it can be slow and prone to errors (Ettl, Zadrozny, Chowdhary, & Abe, 2005).

2.1.2. Key attribute segmentation

The banking industry began to recognize in the 1970s that market segmentation solely based on demographics and socio-psychological characteristics was not accurate, and began using quantifiable determinant attributes. In their seminal paper, Anderson, Cox, and Cooper (1976) defined a list of customers’ “determinant attributes” (such as sex, married, age, or education level) that should be used to calculate customers’ similarity. Tsiptsis and Chorianopoulos (2011) presented several frameworks for segmentation, each beginning with identifying attributes that are used to describe the customer’s expected behavior. Athanassopoulos
(2000) expanded the attribute-based approach by adding simple statistical features (such as: staff size, assets, or years of operation) to improve the description of the expected customer behaviors. Key attribute segmentation is only suitable when the necessary attributes are available. The method also heavily depends on the assumption that the key attributes actually distinguish differences between groups.

2.1.3. Descriptive attribute segmentation

Descriptive attribute segmentation builds on key attributes by converting the attributes into numeric variables. The numeric variables can then be standardized and analyzed using statistical analysis techniques (Kantardzic, 2011). While descriptive attributes are better for quantitative analysis, the drawbacks of key attribute segmentation method remain.

2.1.4. Statistical feature segmentation

Feature-based distance involves extracting statistical features such as median, kurtosis, sum, or purchase frequency, from historical data (Bala, 2012). The features are normalized and become the variables from which distance is calculated (Chen et al., 2006). The tasks of creating, normalizing, and importance-weighting variables—often referred to in literature as “data munging”—do not follow a set of rules or procedures. The accuracy of its results and of subsequent analysis depends on the skills of the analyst. Data munging is considered as much an art as science (Heer & Kandel, 2012). A criticism of using statistical features is that the method is not resistant to outliers. Features such as mean and variance can be greatly compromised by outliers (Park & Leeds, 2016).

Attribute-based and feature-based segmentation share two fundamental assumptions. The first assumption is that the data munging has resulted in a set of variables that reflect the customers’ behaviors. The data analyst must decide the attributes or features to include as variables and the weight to be applied to each. When deciding what variables to create, the data analyst must be cognizant of computer limitations; the computational expense of some clustering algorithms such as neural networks leads some analysts to dimensionally reduce the variables. For example, in preparation for feeding a neural network, Romdhane, Fadhel, and Ayeb (2010) calculate information entropy for each variable and then retained only the most informative attributes. Byrnes (2014) suggests that as the number of attribute variables increases the ability of an algorithm to detect useful information decreases. Variability creation is a qualitative decision process that can greatly influence the outcome of the analysis and may produce inaccurate or misleading results.

The second assumption is that the available data contains enough information to describe the behaviors adequately (Longbing, 2014; Verdu et al., 2006). Attribute information, such as the age of the customer might be essential for accurate segmentation, but if the data is not available, the missing variable might be problematic. Information about strongly influential exogenous factors may be missing from the available data—for example, in a vendor-managed inventory arrangement (VMI), the supplier’s logistics decisions might influence demand patterns and yet no information with regard to those decisions exists in the data.

2.1.5. Behavior model segmentation

Customer behavior can be used to classify customers based on pre-determined segments, which reflect whether a customer is increasing level of business, stable, or at risk of leaving (Ettl et al., 2005). In situations where customers do not have regular interaction, variables are necessary to measure and describe the behavior. Historical transaction data may contain patterns that can be useful to predict behaviors. A large customer base may lead to a large and noisy historical dataset from which behavior patterns are difficult to distinguish. Depending on the point in the supply chain where data is collected, there may be more or less noise incorporated. Research into the Bullwhip effect indicates that as the point of data collection moves upstream from the point of

![Fig. 5. Methodology.](image-url)
consumption, the data incorporates more noise (Chen et al., 2000). Simple aggregation of individual predictions is undesirable since higher level forecasts generally do not equal the aggregated summed forecasts (Hyndman, Lee, & Wang, 2014), and one great model for the entire group will give little information due to the noise in the data.

Chao, Fu, Lee, and Chang (2008) use historical data to measure customer behavior, employing an intermediate preprocessing step which converts the data into RFM (recency, frequency, monetary) variables and then weight those variables. The conversion to RFM and weighting involves subjective decisions and hence has similar limitations to the variable creation strategies discussed previously. As with key attribute segmentation, behavior model segmentation relies on the availability of variables or other information that adequately describes the behavior.

2.2. Measures of similarity

With most traditional segmentation methods, variables are created and then distances between customers are calculated to create a distance matrix. Once a distance matrix has been calculated, general-purpose clustering algorithms can be applied to identify clusters (Liao, 2005). The most popular distance metric in the literature is Euclidean straight-line distance. Simplicity, robustness, and relatively good performance make Euclidean a frequent comparison baseline or foundation for other methods (for examples, see Chang, Tay, and Lim (2015), Hung and Tsai (2008), or Serra and Zárate (2015)). While Euclidean distance performs well with static data, it does not do well when directly applied to time-series data. Euclidean distance is constrained to point-to-point evaluation; similar patterns such as examples “C5” and “C8” in Fig. 2, which have temporal shifts, are not recognized. Euclidean distance treats each time-period as a separate, unrelated variable.

Evaluating time-series with temporal shift is possible when the distance is calculated with cross correlation (CCor) distance. The CCor distance has the characteristic that noise in the data does not significantly impact the results (Golay et al., 1998). The CCor method is limited in that it makes global adjustments in an attempt to find the best correlation between time-series. In effect, it is similar to Euclidean distance, but with added constants (Agrawal, Faloutsos, & Swami, 1993).

Variations in time-series data are never as simple as the examples shown in Fig. 2. Typical variations include changes in frequency, magnitude, and duration. While shifting dissimilar time-series with CCor is often quite successful (Höppner & Klawonn, 2009), there are other techniques, such as dynamic time warping, which perform well on time-series datasets (Wang et al., 2013).

Dynamic time warping (DTW) was first developed as a technique for speech recognition (Sakoe & Chiba, 1978); it was intro-

![Fig. 6. Synthetic data aggregated.](image-url)
Fig. 7. Expected clusters in synthetic data.
duced in the 1990s into data mining research for comparing time-series (Keogh & Pazzani, 1999). DTW is different from other distance measures as it uses non-linear mapping to compare pairs of time-series. While an excellent description of the details of the DTW algorithm can be found in Keogh and Ratanamahatana (2005), a brief illustration is presented in Fig. 3. Although the patterns of objects “A” and “B” in Fig. 3 are very similar, the relationship is not linear. The DTW algorithm warps the time series to achieve a best match between objects and then calculates the distances.

While DTW delivers good results, some researchers are reluctant to use it. A major criticism of DTW is its high computational cost which sometimes results in its elimination as a viable distance measure (Zhang, Kaiqi, & Tieniu, 2006). Two developments offset DTW’s high computational cost. First, as computers become faster and less expensive, computational cost becomes more of an abstract consideration. Second, the continual refinement and development of DTW algorithms make the distance calculations faster (Ratanamahatana & Keogh, 2005). Extensive comparison testing of DTW against other distance methods has demonstrated that criticisms that DTW is exceedingly slow are mostly unfounded (Wang et al., 2013). The authors believe that despite DTW’s slow performance—compared to Euclidean distance calculation—the concern is not great enough to prevent its use.

2.3. Cluster methods

Obtaining accurate segmentation results depends on the selection of a good clustering method (Kashwan & Velu, 2013); of the many available clustering methods, some are too complex for practical applications (Agard, Partovi-Nia, & Trepanier, 2013) and not all are suitable for time-series data. A review by Liao (2005) lists the most common time-series related clustering methods, including partitional, artificial neural networks, and hierarchical. Partitional clustering methods, such as K-means are the most

<table>
<thead>
<tr>
<th>Method</th>
<th>Raw data</th>
<th>Normalized data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical features</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>K-means</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>Euclidean</td>
<td>13</td>
<td>29</td>
</tr>
<tr>
<td>ANN</td>
<td>6</td>
<td>27</td>
</tr>
<tr>
<td>CCor</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>DTW</td>
<td>17</td>
<td>31</td>
</tr>
<tr>
<td>Visual (expected results)</td>
<td>36</td>
<td>na</td>
</tr>
</tbody>
</table>

* Normalized data does not affect CCor results.

While DTW delivers good results, some researchers are reluctant to use it. A major criticism of DTW is its high computational cost which sometimes results in its elimination as a viable distance measure (Zhang, Kaiqi, & Tieniu, 2006). Two developments offset DTW’s high computational cost. First, as computers become faster and less expensive, computational cost becomes more of an abstract consideration. Second, the continual refinement and development of DTW algorithms make the distance calculations faster (Ratanamahatana & Keogh, 2005). Extensive comparison testing of DTW against other distance methods has demonstrated that criticisms that DTW is exceedingly slow are mostly unfounded (Wang et al., 2013). The authors believe that despite DTW’s slow performance—compared to Euclidean distance calculation—the concern is not great enough to prevent its use.

### Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Raw data</th>
<th>Normalized data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical features</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>K-means</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>Euclidean</td>
<td>13</td>
<td>29</td>
</tr>
<tr>
<td>ANN</td>
<td>6</td>
<td>27</td>
</tr>
<tr>
<td>CCor</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>DTW</td>
<td>17</td>
<td>31</td>
</tr>
<tr>
<td>Visual (expected results)</td>
<td>36</td>
<td>na</td>
</tr>
</tbody>
</table>

* Normalized data does not affect CCor results.
commonly used and often employed as a baseline for comparison of other clustering methods, see for example, Hautamaki, Nykanen, and Franti (2008). Although partitional clustering methods create clusters efficiently, the results do not always provide useful information (Murray, Agard, & Barajas, 2015). Additionally, K-means algorithm requires a pre-determination on the number of clusters, which is a challenge in most applications.

Artificial neural networks (ANNs), also known as self-organizing maps (SOMs) (Altintas & Trick, 2014) do not need the number of clusters pre-defined and often give good cluster results. ANNs have generated interest in the forecasting field recently (Co & Boosarawongse, 2007), however, they have drawbacks. One criticism of ANNs is that the algorithms are computationally expensive (Ye & Li, 2002) leading some analysts to employ dimensional reduction of variables prior to computation. See, for example, Bala (2012) or Romdhane (2010). While the dimensional reduction allows the algorithm to operate more efficiently, it also incorporates a subjective decision process into the analysis, which may reduce the accuracy of the results. ANNs characteristic of not needing the number of clusters pre-defined makes the method suitable for the first stage of a two stage clustering analysis where the first stage defines the number of clusters (Kuo, Ho, & Hu, 2002).

A third type of clustering method is hierarchical clustering (Ward, 1963). In hierarchical clustering, the first step is the calculation of similarity (distance) between customers. The distance matrix is used to construct a hierarchy that begins with all customers in individual clusters and ends with all customers in one cluster. The hierarchy is often displayed in a dendrogram as illustrated in Fig. 4. The dendrogram facilitates visual assessment of the cluster results and enables the incorporation of a “cut-line” which established the number of clusters in the result. Unlike K-means, the number of clusters does not have to be pre-determined.

### 2.4. Cluster evaluation

There are many methods to create clusters, ranging from simple a priori to complex algorithms, and nearly every cluster method shares the common trait that clusters (good or bad) are created. No algorithm guarantees that genuine clusters are created and different algorithms often create different cluster results for no discernable reason (Huang, Cheung, & Ng, 2001). An important step in the segmentation process is to evaluate the clusters’ goodness of fit. The literature generally agrees that segmentation results can be evaluated through comparison of inter-cluster homogeneity and intra-cluster heterogeneity (Ding, Trajcevski, Scheuermann, Wang, & Keogh, 2008; Liao, 2005; Seret, Maldonado, & Baesens, 2015). Wang et al. (2013) propose using a one-nearest neighbor classifier on labeled data to generate quantifiable results. The evaluation methods share a common limitation in that they compare one cluster method to others; there is an underlying assumption
that the best clusters are “good” clusters. Keogh and Kasetty (2003) conclude that due to data and implementation biases, “many of the results claimed in the literature have very little generalizability to real world problems”. The lack of accepted evaluation criteria and ambiguity in testing confuses the selection of a “good method” for an evaluation of cluster methods.

Humans’ natural ability to detect patterns visually is excellent, and in some cases superior to computers’ abilities (Chellappa, Wilson, & Sirohey, 1995; Esling & Agon, 2012; Van Wijk & Van Selow, 1999). It would appear then that visual evaluation of cluster results should prove effective—provided that the data is presented in a format humans can visualize (Huang et al., 2001). Unfortunately, high dimensional datasets often contain so many points that it becomes impractical to display graphically precluding visual evaluation (Hung & Tsai, 2008; Kantardzic, 2011).

2.5. Market segmentation through data mining

Market segmentation is both an important part of business management and an active area of contemporary research. Traditional methods employ a variety of strategies with varying degrees of a priori knowledge necessary for successful application. Data mining breaks from traditional paradigms and explores the discovery of knowledge without the preconceptions pre-established hypothesis (Agard & Kusiak, 2004). Successful application of data mining, however, requires careful selection of data mining tools that are able to measure differences between objects (or customers) and then arrange those objects into sensible groups. Finally, a means to evaluate the results of the analysis is required. Unfortunately for the data mining practitioner, there is no tool or prescription for selecting the best technique at each step of the data mining process—rather, it is an experimental and iterative process that includes subjective decisions at different points.

Market segmentation through data mining relies not only on selection of suitable algorithms to analyze the data, but also on suitable inputs to feed into the algorithms. Extracting behaviors from the data requires careful consideration of how the data should be processes so that it actually reflects the behavior (Kantardzic, 2011).

3. Methodology

Market segmentation is traditionally based on descriptive attributes of the customers or on a set of variables created from various data sources. Our method, illustrated in Fig. 5 utilizes only historical data and does not convert the data into variables. This section describes the method to import the raw data, clean it, and extract behavior patterns from it. To validate the methodology, we use a synthetic dataset in section four and a real dataset in section five.

Fig. 10. Industrial dataset – normalized.
Fig. 11. Cluster evaluation.

Fig. 12. Dendrogram of industrial data.
We compare our results with a traditional method (presented in Section 5.4).

3.1. Data preprocessing

The industrial data in its raw format is a set of delivery records; each record contains a date stamp, quantity of product delivered, origin & destination locations, and a customer identifier. In preparation for analysis, the data is first cleaned per Fig. 5(a). Outliers exist in the data due to administrative adjustments (account debits & credits) and errors in data recording. Negative and unreasonably large values are considered outliers and removed from the dataset. Delivery records are then aggregated into monthly bins (monthly bins suit the industrial partner’s planning activities). The resulting sets of customer specific time-series contain a significant level of noise due to irregular delivery frequencies and quantities. Customers with less than monthly deliveries exhibit intermittent patterns due to months with no activity. Newly acquired and lost customers exhibit long periods of zero activities. And, casual or low-consumption customers have very infrequent and unpredictable patterns. Although the analysis would be much easier if all the difficult to predict customers were removed, it would not produce useful real-world results. Therefore, outlier removal was limited and the resulting dataset appears very noisy.

3.2. Segmentation

The literature contains many different segmentation methods, which are summarized in the literature review above. Nearly all of these segmentation methods depend on the availability of several attributes that when analyzed together can lead to useful groupings. In the context of this research, the industrial data contains only time and quantity; there is not enough information for most segmentation methods. The available data can be interpreted to reflect how much product was consumed during each period and therefore is suitable for behavior model segmentation method.

Distance calculation, per Fig. 5(b), is a critical component to the segmentation process. Once the distances are calculated, segments are identified via a hierarchical clustering algorithm per Fig. 5(c). Hierarchical clustering generates good results and enables visualization of the relationship between clusters via the dendrogram output (Barirani, Agard, & Beaudry, 2013). Some researchers have found that hierarchical clustering does not give accurate results, but they were unable to determine the extent that variable selection and distance calculations affected their results (Maqbool & Babri, 2007). Liao (2005) suggests that the actual clustering calculation method does not significantly have an effect on the outcome. Therefore, hierarchical clustering is used for comparing all distance methods.

![Proposed segmentation method, 8 clusters.](image-url)
The segmentation tests were conducted with “R” language (R Core Team, 2015) using Euclidean, ANN (Wehrens & Buydens, 2007), DTW (Giorgino, 2009) and CCor (Mori, Mendiburu, & Lozano, 2015). The most common methods are extensively reviewed in Mori et al. (2015) and Wang et al. (2013).

Artificial neural networks (ANN) and K-means were also tested to validate the proposed methodology. Unlike the methods discussed above, ANN and K-means do not have separate steps to calculate a distance matrix and then build clusters; rather, they produce clusters directly from the raw data. Lastly, a statistical feature-based distance method is included. The feature-based method is a traditional variable creation segmentation method.

3.3. Cluster evaluation and patterns extraction

When deciding how to best evaluate the cluster results, the analyst must determine whether the ground-truth cluster assignment exists, as illustrated in Fig. 5(d). Although there are many statistical methods available for evaluating cluster results, the goal of this research is not to compare and conclude whether the clusters satisfy some arbitrary statistical measure. Rather, the research goal is to determine if one (or many) method can produce “good” clusters in terms of extracted behaviors. The problem of visualizing high-dimensional data is resolved by using a synthetic dataset to test various clustering methods; the synthetic dataset is simple enough to enable visual evaluation of the resulting clusters. While simple, the synthetic data is sufficiently complex that not all algorithms will return a satisfactory solution. We leverage humans’ natural ability for pattern recognition in the evaluation.

Determining a good number of clusters is a challenge yet to be solved and remains as much of an art as a science (Kantardzic, 2011). Many distance and cluster options were tested in this research to evaluate the effects of different cluster sizes while retaining the ability to display the results. For this research, the synthetic data described in the following section is developed and evaluated with eight clusters. For the industrial data, 8 and 24 clusters appeared pertinent in the dendrogram; this level of segmentation suits the managerial needs of the industrial partner. This research uses cluster homogeneity to evaluate the effectiveness of the cluster methods.

3.3.1. Cluster evaluation – synthetic dataset

For the synthetic data, the ground-truth cluster assignment is known, so the analysis can proceed according to Fig. 5(e). We developed the synthetic dataset with the intention that clusters can be readily distinguished by human visual evaluation. Predetermined clusters establish the ground-truth of the expected cluster results from which the results of each clustering method are compared.

\[ \text{Fig. 14. Traditional segmentation method, 8 clusters.} \]
pared. For the second step of evaluation, the results from each segmentation method are graphically displayed for visual evaluation. The two-step evaluation established not only which segmentation method performs best, but also whether the methods produce “good” clusters.

3.3.2. Cluster evaluating – industrial dataset
Unlike the synthetic dataset, the industrial dataset has no ground-truth segmentation and the analysis must rely on a subjective assessment, per Fig. 5(f). The subjective evaluation is enhanced by utilizing a baseline for comparison. For the baseline, variables are extracted from the raw data using simple statistical analysis; sum, mean, kurtosis, skew, frequency, and standard deviation are used as variables. Euclidean distance is applied to calculate a distance matrix based on the baseline’s variables. The baseline and the proposed method both begin with the same raw data; segmentation for both is done with hierarchical clustering. Once clusters are created, the resulting graphical output are inspected for evidence of patterns in the data. Although a visual evaluation lacks the rigor of quantifiable measure, it offers a valid assessment of whether or not an industrial practitioner can detect patterns such as seasonality, increasing trends, or lost customers. The kind of results expected are explained in Table 2.

4. Synthetic dataset
Selecting a suitable distance measure and clustering algorithm are critical to the overall method in this research. A synthetic dataset was created to enable easy and effective evaluation of the results of each candidate distance measure.

4.1. Research dataset
The synthetic dataset was developed by plotting customers with eight different types of simple behaviors patterns. Within each group, the customer exhibit behaviors that are similar, but occur at different magnitude levels, time lag, or have slight variations and noise. When viewed as a single group in Fig. 6, the synthetic data appears complex; visually, it is difficult to detect more than two or three distinct patterns in the data.

It is expected that a good clustering method will detect the similar behaviors and not be incorrectly influenced by where or when the behavior occurred. Fig. 7 shows the synthetic data manually separated into the expected clusters based on behavior similarities—we consider the cluster assignment in Fig. 7 as the ground-truth for evaluating segmentation strategies. Although the synthetic data appears extremely simple, some distance methods...
and clustering algorithms preformed surprisingly poorly. The test results are in the following section.

### 4.2. Segmentation – synthetic dataset

The goal of the research is to find a method that leads to logically matched cluster members. The evaluation of the different methods was based on how closely the results are to the expected results from Fig. 7. The synthetic data was passed to the distance calculation and clustering algorithms and cluster assignments were generated by each method. For each test, the number of clusters (K) was set to eight (the expected number as illustrated in Fig. 7). Table 1 shows the results of each method when applied to the synthetic data. The maximum value for the results in Table 1 is 36—indicating a complete match to the visual baseline; lower values indicate that the strategy does not produce an expected result. The data is normalized to remove the undue effect of the size of each customer; behavior patterns are preserved while the effect of magnitude is eliminated. The results in Table 1 show that, with the exception of CCor, every method performs significantly better with normalized data.

Some data mining methods require normalization of the data to calculation. Normalization prevents overweighting features that have larger average values (Kantardzic, 2011). In the tests conducted, most methods performed better with normalized data. CCor distance produces the same results with or without normalized data due to its inherent scaling within the correlation calculations.

### 4.3. Discussion on synthetic dataset

The synthetic dataset is intentionally simple to ensure human visual evaluation could be used to conclude whether clusters were “good” or not. The simplicity enabled the establishment of the baseline expected cluster results. We observed that DTW with scaled data produce the best results; DTW produced 31 correct results when compared to the baseline. However, DTW detected similarities that were not expected. For example, the bottom member (blue) of Cluster 2 in Fig. 8 was expected to be included in Cluster 1 due to its nearly flat pattern. The small variation is visually similar to other members of Cluster 2, although it does not follow the expected results, the solution appears to be valid. We conclude from the results that DTW with scaled data is able to generate acceptable clusters.

### 5. Industrial dataset

#### 5.1. Context

A supplier of bulk liquid materials—used for manufacturing, food packaging, and medical services—provided the industrial

---

**Fig. 15 (continued)**
The dataset used in our research. The supplier maintains point-of-use inventories of the bulk liquid and is responsible for ensuring uninterrupted product availability. The bulk liquids require specialized supplier-specific point-of-use storage tanks and multi-year purchase contracts are the norm; customers typically do not purchase from alternate suppliers during the contract period. Delivery dates and quantities for a five-year period are available for the study; actual customer-level consumption data is not available.

Fig. 9 represents the original data set aggregated into monthly bins. Each line represents delivery pattern for one specific customer (units on the y-axis are suppressed for confidentiality). We can observe a large variation between customers, in terms of quantities, number of deliveries, and delivery dates, yet no patterns appear between customers—this is the data that is actually available to the company to make prediction models, with more or less success.

The customers represent many different industries that use the product for different purposes. Various internal and external factors influence each customer’s operation and in turn affect their specific product consumption. The customers’ location and industry type are known; however, segmentation based on these attributes does not result in any similarities in consumption behavior patterns. Therefore, knowing who the customers are and their industry type does not provide useful information to predict future consumption rates.

In the VMI arrangement of the case study, delivery requirements are triggered when sensors in the point-of-use storage tanks indicate a low level. The supplier attempts to both minimize delivery frequency and avoid point-of-use inventory stock-outs. Although the ability to adjust delivery quantities and frequency affords the supplier flexibility in its logistics, the resulting historical data, based on deliveries, does not necessarily reflect the customers’ actual consumption behavior. For example, while two customers may have identical consumption behavior, the supplier might use different delivery strategies due to distance from warehouse or proximity to other customers. The resulting historical data can indicate different demand behavior for customers with identical consumption behavior.

The industrial partner views its production and delivery of products as a continuous process with monthly production volumes as the preferred measure. The available data, however, is not a measure of a continuous process, but instead it is a set of delivery records. While some customers have frequent and consistent deliveries, others have very infrequent deliveries and/or highly varied delivery quantities. Additionally, the total demand quantity varies from very small customers to very large. When viewed as a monthly aggregated time-series format in Fig. 9, the data reveals little information; customers with relatively small consumption quantities are all banded together in the lower region of the graph.

Fig. 16. Traditional segmentation method, 24 clusters.
Preprocessing the industrial data began with normalizing to remove the effect of overall magnitude; normalization preserves behavior patterns, and produces superior cluster results. The normalized data in Fig. 10 is no longer influenced by customer size leaving only behavior patterns, however, the normalized data reveals no discernable patterns.

The following normalization formula is used:

\[ x_{\text{norm}} = \frac{(x - \bar{x})}{\sqrt{\sum (x - \bar{x})^2}} \]

where \( x \) is the value of the delivery quantity, \( \bar{x} \) is the mean of all delivery quantities, and \( x_{\text{norm}} \) is the normalized value. The chosen normalization method is well suited to this type of data due to its resistance to the distorting effect of outliers.

### 5.2. Calculating a distance matrix

In order to group the customers into segments with similar behavior patterns, a measure of similarity is needed; DTW is employed to calculate the distances. The industrial dataset comprises a matrix of many time-series data; a wrapper algorithm enables direct application of the DTW calculations to the time-series matrix (Mori et al., 2015). Within the DTW algorithm, the time-series are elastically stretched (warped) to create a best fit between time-series (Keogh & Ratanamahatana, 2005). The warping is constrained by a window of two to prevent over-manipulation and loss of important behaviors such as seasonality. After warping, the algorithm uses Euclidean distance to calculate the pair-wise distances.

### 5.3. Segmentation – industrial dataset

We applied the clustering methodology (using DTW and hierarchical clustering) to the industrial dataset to separate it into sensible groups. Unfortunately, there is no automatic method for determining the best number of clusters; selecting number of clusters is one of the most difficult problems in data mining (Jain, 2010). The industrial partner in this research is interested in limiting the number of clusters to a relatively small and easily managed number while still having enough separations to enable detection of different behaviors. To enable an objective selection of the number of clusters, we calculated within cluster homogeneity and in-between cluster heterogeneity.

- Within cluster homogeneity is the mean value of all pairwise distances between all elements within the same cluster
- In-between cluster heterogeneity between clusters A and B is the mean distance of all pairs of elements \( a_i \) and \( b_j \) where \( a_i \) belongs to A and \( b_j \) belongs to B.

![Fig. 16 (continued) Ind. Dataset – Clus # 5](image1)

![Fig. 16 (continued) Ind. Dataset – Clus # 13](image2)

![Fig. 16 (continued) Ind. Dataset – Clus # 21](image3)

![Fig. 16 (continued) Ind. Dataset – Clus # 6](image4)

![Fig. 16 (continued) Ind. Dataset – Clus # 14](image5)

![Fig. 16 (continued) Ind. Dataset – Clus # 22](image6)

![Fig. 16 (continued) Ind. Dataset – Clus # 7](image7)

![Fig. 16 (continued) Ind. Dataset – Clus # 15](image8)

![Fig. 16 (continued) Ind. Dataset – Clus # 23](image9)

![Fig. 16 (continued) Ind. Dataset – Clus # 8](image10)

![Fig. 16 (continued) Ind. Dataset – Clus # 16](image11)

![Fig. 16 (continued) Ind. Dataset – Clus # 24](image12)
Homogeneity is a measure of similarity of members of a cluster, and heterogeneity is a measure of dissimilarity between clusters. When distance is used as the measure, it is desirable to minimize heterogeneity distance and homogeneity distance (Fisher, 1958). The cluster evaluation displayed in Fig. 11 indicates that for less than eight clusters (K < 8) the results do not look stable. Additionally, for K > 25, there is only little change. Therefore, considering the industrial partner’s preferences and the quality of information that we anticipate, we tested “K” equal to 8 and 24 for this research.

Fig. 12 is a dendrogram that shows the relationship between elements of the industrial data. The red rectangles illustrate the makeup of the clusters when split into 8 clusters.

5.4. Results of segmentation attempts

Segmentation is performed using both a traditional segmentation method and the proposed method. The results in Fig. 13 are for K = 8 clusters using the proposed method. Although the data remains quite noisy, the sub-groups begin to show some limited information in the behavior patterns. Most clusters indicate stable demand, although Cluster 5 and Cluster 8 appear more stable than the others. Cluster 3 shows an increasing trend over the entire period of the data; Clusters 2 and 4 also show an increasing trend, although it is not as evident as in Cluster 3 and only occurs during the first few years. Cluster 7 contains customers with a mixture of no activity and spikes in activity.

The same dataset is segmenting using a traditional segmentation method; variables are created by extracting statistical information from the data. The results of the traditional method are shown in Fig. 14. Unlike the results from the proposed method, the traditional segmentation method produces clusters that offer no apparent indication of behavior pattern. The traditional method has assigned most customers to only three clusters while leaving Cluster 7 and Cluster 8 with only one customer each.

Both segmentation methods are again tested, this time using K = 24 clusters. With the proposed method, the clusters become clearer and more patterns emerge, illustrated in Fig. 15. Clusters 3 & 17 both indicate increasing trends, however, they have slightly different patterns. Cluster 2 shows a multi-year cycle; a similar cycle may also be evident in Cluster 7. Cluster 6 and 24 both appear stationary over time (not increasing or decreasing), however, Cluster 24 has less variation and should produce predictions with a higher level of confidence.

The traditional method of extracting statistical information to create variables is used to produce 24 clusters, as shown in Fig. 16. The traditional method does not give good results. With a few exceptions, none of the clusters produced by the traditional segmentation method indicate any useful patterns. As with the attempt to produce eight clusters, the traditional method assigns

\[ \text{Ind. Dataset – Clus # 1} \]

\[ \text{Ind. Dataset – Clus # 2} \]

\[ \text{Ind. Dataset – Clus # 3} \]

\[ \text{Ind. Dataset – Clus # 4} \]

\[ \text{Ind. Dataset – Clus # 9} \]

\[ \text{Ind. Dataset – Clus # 10} \]

\[ \text{Ind. Dataset – Clus # 11} \]

\[ \text{Ind. Dataset – Clus # 12} \]

\[ \text{Ind. Dataset – Clus # 17} \]

\[ \text{Ind. Dataset – Clus # 18} \]

\[ \text{Ind. Dataset – Clus # 19} \]

\[ \text{Ind. Dataset – Clus # 20} \]

Fig. 17. Data smoothed with Croston’s method.
most customers to a few clusters and leave other clusters with a single member.

5.5. Discussion on industrial dataset

While several clusters in Fig. 15 exhibit very distinct patterns, not all clusters are easy to interpret. Aggregating the original delivery data into monthly periods has created a set of very noisy time-series, which are difficult to visualize. Since some customers have less than monthly delivery frequencies, the monthly aggregation has resulted in intermittent time-series for those customers. We compensate for the noise in the data by applying a classical smoothing technique prior to segmentation. A commonly applied correction for intermittent time-series data is Croston’s method (Croston, 1972). When we apply Croston’s method to smooth the data, the resulting clusters reveal much more information than before. Fig. 17 shows the 24 clusters with smoothed data.

Most clusters in Fig. 17 reveal now several clear behavior patterns. For example, seasonal changes are evident at the beginning of the year for most segments, but not in Clusters 13, 15, 17 and 24. Other changes in demand indicate either monthly or quarterly seasonality or possible a cyclic behavior. Changes in overall trends are also evident; some customers have stable consumption, some increasing, and others decreasing. Table 2 summarizes the segments based on trend patterns. From Table 2, we observe that a significant portion of the customer base is either stable or increasing. Conversely, we also observe that 23% of the customers are decreasing; the decreasing customers account for nearly a third of the historic delivery quantities; the decreasing trend indicates a significant population of customers at risk of decreasing or ceasing business. Table 2 also reveals more subtle details such as the point in time when a change occurred and the magnitude of the change.

Column k of Table 2 lists the homogeneity of cluster based on the distance between the clusters members calculated per formula 2

$$H_{\text{homogeneity}} = \sum \sqrt{(x_i - y_i)^2}$$

where $H$ is the sum of distances between observations and $x_i$ and $y_i$ are numerical values of two observations. The values of $H$ are all similar in magnitude, ranging from 40 to 52. The similarity of homogeneities indicates that all clusters have similar density and uniformity.

The methodology that we proposed permits to extract patterns in the data that show distinctive behaviors of the customers. These patterns are not visible in the combined data in Fig. 9. Once visible, past behaviors can be used to make predictions for future behavior to support the development of strategic forecasts. In some cases, additional information is evident and may promote other manage-
ment activities, such as identifying at-risk customers before they become lost customers. The initial dataset has a high level of noise that reveals little information. From the patterns in Fig. 17, we can detect new customers, lost customers, growing customers, and decreasing customers. Although some clusters, such as Cluster 7 and Cluster 9 in Fig. 17 reveal distinct patterns, others are more difficult to assess. Column (d) of Table 2 reflects how much noise or variation exists within the cluster. The assessment of noise in each cluster will help to establish the level of confidence in subsequent analysis performed on each cluster. Clusters 11 and 23 contain customers with a high level of noise, these observations share only the common trait that they do not fit well into any other clusters.

6. Conclusion

The engineering challenge was to provide information to facilitate the strategic forecasting process. The available data was limited to historical delivery data that initially revealed little information due to high noise in the data. It was proposed to create sub-groups of customers with similar behavior patterns. Traditional market segmentation methods rely on availability of descriptive variables; they could not work given a lack of information about the customers. The situation was further complicated due to vendor managed inventory arrangement whereby customers' consumption behavior and the supplier's logistics decisions both influence delivery patterns.

The proposed methodology for market segmentation relied on detecting patterns of behavior in customers' historical data. By using dynamic time warping, we were able to measure similarities (distance) between customers. Sub-groups were subsequently established based on the distances. Behavior patterns, indistinguishable in the overall population of customers, were revealed in the sub-groups. We resolved some of the problems of high noise level in the data by smoothing it by applying Croston’s method (Croston, 1972); the resulting market segments revealed a variety of information including seasonality, trends, and cycles. Information produced by the method proposed in this research has direct management applications for identifying segments of customers with similar behaviors—from the identified behaviors, management can make informed decisions on future actions within the segments.

Market segmentation is a commonly applied management tool, and while there is a growing abundance of data available regarding industrial, economic, and social trends, there is little guidance to managers on how to interpret and convert the data into actionable information. Our proposed method extracts useful information using only historic delivery data. This method could be successfully applied to any domain that has historic delivery data available. Analysis of intermittent time-series is an active area of research that attempts to address noise in data due to aggregating delivery or transaction data into time-series format. Our proposed method should be effective on any dataset that originates from transaction data.

While this study gives insight into a method to segment customers, it has limitations in that the number of segments has not been optimized. Further research is necessary to determine a suitable method to define the best number of segments and an analytical method to evaluate the resulting clusters. Future research should also investigate ways to incorporate external data information. Our proposed method extracts useful information using only historic delivery data. This method could be successfully applied to any domain that has historic delivery data available. Analysis of intermittent time-series is an active area of research that attempts to address noise in data due to aggregating delivery or transaction data into time-series format. Our proposed method should be effective on any dataset that originates from transaction data.

Acknowledgement

The authors would like to acknowledge our industrial partner (Air Liquid) and the National Sciences and Engineering Research Council of Canada (NSERC) for funding under grant RDCPJ 492021-15 and other support for this research.