Dynamic Time Warping For Behavioral Similarity Clustering For Retail Sales Forecasting And Insight Generation

by

Efe Pınar

A Thesis Submitted to the

Graduate School of Engineering

in Partial Fulfillment of the Requirements for

the Degree of

Master of Science

in

Industrial Engineering

Koc University

September 2013

Koc University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Efe Pınar

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by the final

examining committee have been made.

Committee Members:

_____

Assoc. Prof. Dr. Özden Gür Ali (Advisor)

_____

Prof. Dr. Serpil Sayın

_____

Asst. Prof. Dr. Onur Kaya

Date: _____

ABSTRACT

Retail industry is a dynamic industry with many observable and hidden drivers of sales. Accurately forecasting short and long term sales is a major advantage to retailers as it helps with decision and policy making. Along with forecasting, determining and understanding hidden drivers of sales is invaluable for identifying indicators of change in trends.

In this thesis, we propose a method for increasing forecasting accuracy of the model proposed by Gür Ali and Pınar (2013), and provide a new method for automatic similarity identification to gain insights. The base model is a multi-period two-layer pooled regression model. The first layer considers marketing, inflation and seasonal effects. In the second layer, the residuals from the first layer are extrapolated to identify trend-cyclical components of the sales. The pooling is done according to characteristics of stores, predetermined by the company.

We propose behavioral pooling to improve the accuracy of the forecasts. Behavioral pooling groups the stores according to their similarity in movement patterns. We use Dynamic Time Warping method to quantify similarity between stores. The resulting pooling significantly improves the accuracy of the base model.

Dynamic Time Warping is also used as a similarity measure between the residuals and environmental and socio-economic indicator time series to gain insights about other potential drivers of sales.

ÖZETÇE

Perakende sektörü çok dinamik ve tahmin etmesi zor bir sektördür. Özellikle Türkiye gibi büyüyen ekonomilerde sektörün ilerleyişini, sektörü etkileyen faktörlerin anlaşılması ve tahminlenmesi zordur. Perakende satış tahminleri ne kadar doğru olursa, perakendecilerin kısa ve uzun vadeli karar mekanizmaları daha düzgün çalışır ve büyüme gözlemlenir.

Bu tezde, Gür Ali ve Pınar (2013) tarafından uygulanan satış tahmini modelini geliştirmek için davranışsal gruplandırma yöntemi önerilmiştir. Davranışsal gruplandırma amacıyla Dinamik Zaman Bükme (DZB) yöntemi kullanılmıştır. Ayrıca içgörü geliştirme amacıyla da gene DZB yöntemini temel olarak kullanan bir yöntem geliştirilmiştir.

Temel olarak kullanılan model çok dönemli iki seviyeli bağlanım analizi yöntemidir. İlk seviye satışları etkileyen pazarlama, enflasyon ve mevsimsellik etkilerini göz önüne alır. İkinci seviye, ilk seviyeden arta kalan sapma değerlerini geceleğe dönük tahmin ederek eğilim ve çevrimsel hareketleri modellemektedir.

Temel modelde seriler şirketçe belirlenmiş özelliklerine gore gruplanmaktadır. Bu tezde önerilen davranışsal gruplama, zaman serilerini hareket ve örüntülerine gore gruplamaktadır. Bu gruplama için kullanılan, iki seri arasındaki benzerlik değeri DZB mesafesinden gelmektedir. Sonuçlar incelendiğinde DZB temelli davranışsal gruplamanın daha doğru tahminler verdiği gözlemlenmiştir.

İçgörü geliştirme amacıyla, DZB mesafelerini temel alan bir Monte-Carlo deneyi geliştirilmiştir. Mesafelerin dağılımı incelenerek istenilen sayıda zaman serisini birbiriyle karşılaştırarak benzer serileri tanımlamayı otomatik hale getiren bi program oluşturulmuştur. Yapılan deneyler sonucunda geliştirlen yöntemin doğru ve içgörü oluşturan benzerlikler olduğu görülmüştür.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Assoc. Prof. Dr. Özden Gür Ali; without whom I woud not have the chance to learn and master the subjects and lessons I love.

I would like to thank my parents for their preparation of me to the world with care, dedication and most importantly patience.

I would like to thank all of my friends that I have gained in my education in Koç University; too numerous to name. We have worked and laughed together and I appreciate all the times we have spent together.

I would like to thank my girlfriend Ayşe Ece Durmaz; without her continuous support and joy I would have tried to quit countless times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Retail industry demand is highly dynamic with many observable and hidden drivers with different affects. Especially in countries with growing economies and inherently dynamic economy, retail industry is especially hard to work in. Retailers that can accurately forecast their demand and understand the impact of drivers have a great advantage; however these accurate forecasts are hard to achieve in such environments.

The main goal in retail forecasting is achieving accurate forecasts for all of the business aggregation levels, from store and product category level to national level sales. Gür Ali and Pınar developed a multi-period two-layer pooled regression model for forecasting retail sales [33]. That model takes into consideration marketing expenses and seasonal effects and extrapolates the residuals into the future to achieve forecasts. They forecast up to 12 month ahead and pool the store and category level sales according to the business hierarchy to improve model accuracy. To achieve higher level sales, the base level forecasts are aggregated to the desired level [33]. The data source for this paper as well as the thesis is from the leading retailer in Turkey.

In this thesis, we propose two improvements to the model developed. Behavioral pooling is implemented to cluster similarly behaving stores and fit models for sub groups of stores. To quantify the behavioral similarity between stores, Dynamic Time Warping (DTW) method is used. DTW is a method that is used in speech recognition, pattern matching in electronic wavelengths and to a lesser extent in image matching. Implementing DTW to find similarities between economic time series have been

proposed before but using it as a similarity measure in clustering time series is a new contribution to the literature. We use DTW as the dissimilarity measure between the series and implement k-medoids to cluster the series. Through this process, behaviorally similar store clusters are formed. The pooled models are known to have improved forecast accuracy.

Another important contribution is a method for discovering underlying effects and hidden drivers of sales and/or demand. Identifying hidden drivers gives retailers opportunity to develop indicators or set red flags from outer sources when considering their sales. DTW is implemented for series matching as well. In the current state, there is no solid "benchmark" for DTW that indicates whether the series are similar or not. We develop a Monte-Carlo simulation to identify such benchmarks and use them to identify significant similarity between residual series numerous socio-economical indicators. This method is useful when there are numerous series to be matched and an automated method is needed to identify significant similarities.

The rest of this thesis is as follows; first we review other research done on the subject of retail forecasting with multiple series and DTW method in chapter 2. Then in chapter 3 we explain and formulate the base regression method Gür Ali and Pınar developed, the data used, and provide the forecast accuracy of the regression method. In chapter 4, we describe in detail our behavioral clustering method, how it is implemented, including the prerequisite methods. We give the results regarding the accuracy improvement with the proposed behavioral clustering method in chapter 5. In chapter 6, the method for gaining insights and discovering drivers of sales is explained. The results with the data set at hand are also given. Finally we conclude with a summary of our aim, the method, and the benefits of our methodology in chapter 7.

Chapter 2

LITERATURE REVIEW

This thesis focuses on multi-period ahead forecasting of store-category level retail sales. To achieve these forecasts, we implement a regression forecasting method that uses behavioral pooling to improve the accuracy and model quality. For the behavioral pooling, we use Dynamic Time Warping, a method capable of identifying similarities in patterns and series with different speeds or lags.

There are different streams of research present in the literature that tackle a similar problem or uses a similar methodology. We have identified the main streams that are relevant to our research.

- First we go over the forecasting literature, types of methods along with their advantages and disadvantages and the research on each method
- A comparison between independent and/or single time series and multiple time series forecasting is given.
- The concept of analogous time series and panel data are given and the relation between multiple time series forecasting is established.
- To improve forecasting accuracy in the presence of analogous time series, pooling time series is an option. Types and methods of pooling are explained.
- Similarity measures between time series are investigated for the goal of pooling and Dynamic Time Warping is explained in further detail.
- The research on single and multiple time period ahead forecasts are given.

- Aside from forecasting, an automated method for similarity identification between series is proposed. The literature on similarity measures, their efficiency and their use in retail and socio-economic time series are investigated.

These main streams are explained in detail in this chapter.

2.1. Forecasting Methods

There are three different methods of forecasting. There are 1) Judgmental models, where expert judgment is used to forecast, 2) Extrapolation methods where past data is used to forecast future and 3) Causal methods where outside data and effects are used to forecast series. There are many methods for each type of forecasting methods.

Judgmental models do not require an estimation or parameter determination. Experts try to identify the past and current trends, effects and driver of the time series at hand. Then they make a calculated guess about the future of the time series at hand. However a method for modeling the expert decision under different effects have been developed by McIntyre et. al. [43]. They have provided a method for forecasting the results of promotions through case-based reasoning model. This model tries to simulate experts' decision input in case-to-case basis. The provided method is more of a decision support system rather than forecasting decisions; however the principle of leveraging similarities of analogous series are present. They expect to learn from the past decisions by investigating the reactions of individual incidents, then propose the correct decision based on the similarity of a new event to the past events. Collopy and Armstrong provided a like-wise solution; modeling the experts' judgment with 99 questions and making forecasts through these questions of the situation [5]. The drawback of these methods is that judgmental methods are objective and dependent on the opinions of different experts.

Extrapolation is commonly used to forecast time series. They use the past behavior and effects in the data and forecast under the assumption that the series will behave with the same trend-cyclical components in the future. Exponential smoothing methods, ARIMA models all use solely past data to forecast future. Lee et. al. compared the judgmental and extrapolation methods to forecast Conference Board's buying plan. They concluded that extrapolation methods outperform judgmental methods [40]. In 1984, Armstrong investigated the extrapolation methods up to that year and concluded that sophisticated extrapolation methods provide negligible improvements. He recommends using simple models and combination of the resulting forecasts in extrapolation methods [4]. The main difficulty in extrapolation is that these methods have consistent trend and seasonal components forecasting into the future, thus have difficulty following changes in the trend or seasonal components.

Causal methods identify and investigate outside drivers and indicators of the time series. These methods use the identified drivers and indicators, determine their future values and forecast the original time series under the assumed level of drivers and indicators. Regression methods are mainly used in causal methods, however more complicated models like Neural Networks may also be implemented. A comparison of some of the causal methods yield that segmentation, damped seasonality and trend and decomposition improve the accuracy of the causal methods in time series forecasting. [3]. Gür Ali proposed a driver moderator method for retail sales prediction [30]. In their method, marketing effects are used to forecasts retail sales. Drivers' degree of effect is set through moderator variables that take into account socio-economic factors and store properties that affect the sales of a specific SKU. Their method outperform regression, exponential smoothing and neural network models with parameters they have established. A disadvantage of causal methods is that they require the forecast of the outside effects beforehand, and these methods forecast from the forecasts of outside effects.

2.2. Independent Series and Multiple Series

The main component of forecasting is individual forecasting against forecasting multiple series. When there is a single time series, the information is of a single series and they are all used to model and forecasts the same series. To forecast single series, different models have been proposed. A method was proposed by Chen and Ou; using Grey Relation Analysis and Extreme Learning Machines to forecast retail [13]. They are forecasting a single time series and Grey Relation Analysis method is used to identify the degree of outside effects on a time series; Chen and Ou discovers the most important effects and gives these inputs to ARIMA models and learning machine methods to conclude that learning machines outperform traditional methods when the input series are selected well. Alon et. al. used neural networks for forecasting retail sales [2]. They used a single time series to compare the exponential smoothing, regression, ARIMA and Neural Network models and conclude Neural Networks perform best for multi-period forecasting. Gao et. al. also proposed using neural networks for forecasting retail sales for a single time series, using neural networks to adjust exponential smoothing methods' parameters. They report an increase in the accuracy of forecasts [26].

Aside from forecasting single series, there are multiple series where information and data points are multiple. The series may or may not be independent. There are three main ways of dealing with multiple time series forecasting; 1) treating each series independently, 2) aggregating the series into a single time series and 3) pooling the series to create models that use information from multiple series.

Treating each time series independently ignores the possible relations between series and reducing the forecasting process to multiple individual time series forecasting.

Aggregating the series lower the number of different series to the required number and may even be used to create a single series to forecast. The proper level of

aggregation for such circumstances were discussed and different metrics are proposed to help determine the level of aggregation to be forecasted [61]. A method to achieve lower level forecasts from this aggregated point is distributing the aggregate forecasts to lower dimensions [19].

Leveraging the relation and information of the series is part of an important concept, the presence of analogous time series.

2.3. Analogous Time Series

Analogous series are group of series that have similar patterns, properties and reactions to the same effects, e.g. marketing expenses, economical indicators [10], [21]. These similarities may either result from having similar properties in the first place, thus creating similar movement patterns overall; or they may result from having same kinds of reactions to an economical or social event. In econometrics, multiple analogous series are called panel data. Multiple analogous time series are merged into a matrix consisting of individual series as vectors, creating a combined matrix of information [22].

There are established and widely used methods for modeling analogous time series and panel data. To model panel data, regression methods are widely used to understand and model impact of individual effects and time series. Understanding the effect of individual estimators is invaluable; e.g. it can be used for policy analysis [28], [29].

There are methods other than regression for forecasting panel data as well. In 2003, Chu and Zhang provided a review and comparison of linear and non linear methods for aggregate retail forecasting [16]. They have concluded that nonlinear models are able to outperform linear models. They also state seasonal adjustment improves

forecasting accuracy in the neural network models; the best model was neural networks with deseasonalised series amongst the methods they have tried. Another important result they have discovered is that using seasonal dummies improves the accuracy of regression methods.

2.4. Pooling Time Series

2.4.1 Benefits of Pooling for Forecasting

Pooling is grouping individual members of a group, whether they are single observation or time series, and creating a bigger "pool" of observations. This is done to gain more observations to model and improve the fit of the model with more observations to learn from. Pooling the time series also gives the forecasting model an opportunity to consider inter-dependency between the series. The models built uses information from multiple series to create estimators. There are different types of pooling including 1) correlational co-movement groups, 2) clustering locations based on their characteristics and 3) pooling from expert judgment [21].

The benefits of clustering the time series themselves in analogous property leveraging have been investigated before. Lu and Wang clustered demand time series from computer industry in a single time period and use supper vector regression as the modeling approach. They use the weights of independent components in each individual time series to cluster the series with Growing Hierarchical Self Organizing Maps. Then, different SVR models are created for each cluster, and finally to implement forecasting, a classification algorithm is developed to identify the cluster of a point before using the appropriate SVR model [42]. Torgo and Costa propose a clustered regression method that cluster observations and creating individual models for each cluster group to improve the accuracies. They use cluster membership

probability; a measure for identifying how similar the series are to the cluster; as the weighing factor in the regression as well, and they have observed improvement in regression models when training set is clustered pre-modeling [56]. The order of the clustering and regression steps was investigated and it is found that clustering followed by regression improves accuracy as well [39].

The retail sales for individual store-categories may be noisy and irregular and Pooled models where different series are taken into consideration is proven to be beneficial when the series are noisy [21]. In their methodology, analogous time series that have similar patterns due to their reaction similarities are pooled together; they are then scaled and models that consider trend at both individual and aggregated levels are created. On the other hand, Corberan-Vallet et al argue that series subject to random disturbances do not necessarily have a common movement pattern and suggest a MCMC simulation for exponential smoothing models of multiple series [17].

There are many different model types for pooled regression of panel data. There are 1) single intercept with multiple variables for individual series, group specific intercepts with fixed effects, and 2) random parameters model that is able to formulate heterogenousity in variable patterns. In marketing, econometric models of panel data are used to deduce promotion response [22], [28].

To group the individual series according to their similarities, different approaches are possible. These are 1) clustering of series based on business hierarchy, 2) clustering individual constants and 3)clustering the time series themselves [10]. Grouping by business hierarchy means grouping series based on properties pre-determined by the business or experts. For clustering individual constants; they can be averaged/aggregated per group [18] or the series can be aggregated and the seasonal constants derived from this point [57] .

2.4.2. Pooling Accuracy

The better the pooling is done, the more accurate the model of the specific pool can be. As we have analogous time series, groups according to the inherent similarities of time series can be achieved. We investigated the research on the subject of defining and quantifying similarities between time series. There are 1) lock-step measures, 2) elastic measures, 3) threshold-based measures and 4) pattern-based measures for defining similarities [20]. Fu provided a review on time series data mining, focusing on identifying similarities and patterns present in time series [24]. They provide a compendium of the methods instead of the comparison between each method. They state there is whole sequence matching and sub sequence matching. Sub sequence matching deals with sub patterns and part of the series to identify similarities.

There are a lot of different methods for identifying patterns in different series. TS-Tree method is proposed for retrieving and identifying similarity between time series [6]. Deriving new series that focus a single aspect or remove some of the underlying effects from existing series and then classifying the new series by similarity have been proposed [7]. Chiu et. al. offered a probabilistic method to discover the motifs inherent in the series; these motifs can be used to match and find similar series [15].

2.4.3 Similarity Measures for Pooling - Dynamic Time Warping

In dynamic environments like retail, lagged relations and temporal dependencies are important to identify. Also sub patterns in the series are more important to find drivers that have had an effect in the past for a short time span.

These lagged relations, temporal dependencies and sub patterns identification require more elastic methods for finding similarities as lag and reaction speeds are major differencing factors. Tang et. al. proposed a method for discovering lagged similarities when there are temporal dependencies [54]. Dynamic Time Warping is another method proposed for discovering elastic similarities. Its' roots go back to speech recognition with dynamic programming [58] and have been further improved for both speech recognition [51] and general time series matching [50]. Dynamic Time Warping methods are used for identifying cross-similarity between data streams [57]. Jeong et. al. proposed a weighted Dynamic Time Warping based method for clustering time series [35]. Using DTW distances are shown to perform better when coupled with k-medoids instead of k-means [46]. An application of DTW with Neural Networks has been proposed for stock trading decision making; an environment even more dynamic than retail sales [11]. Many different researches on DTW have been done regarding the efficiency and accuracy of the methods. Rakthanmanon et. al. provides a breakthrough optimization for DTW enabling processing trillions of data points in a short time, thus combining the advantage of proper identifying with efficiency [48]. An alternative of DTW, Derivative Dynamic Time Warping is developed by Keogh and Pazzani, to better model the changes and cyclical patterns between the series [37]. On the other side, it is proposed that using elastic methods like DTW with lock-step measures together perform better than elastic methods alone [25].

2.5. Single-step and Multiple-Step Ahead Forecast Comparison

The models and methods discussed have a main difference from our aim; they are focusing on one-step-ahead forecasting instead of multi-step forecasting. In retail sector, multiple-step ahead forecasts that can be updated/corrected with additional information is a necessity. For multiple-step ahead forecasts, there are two mainly

used approaches; iterated use of the one-step ahead forecasts (IMS) and direct multi-step estimation (DMS). Chevillon and Hendry review the debate of when each method is more applicable [14]. They state that misspecification of the error process result in DMS being more accurate. They simulate experiments and conclude that non-parametric DMS result in gains "when economic variables exhibit varying trends or are subject to cyclical patterns". This methodology uses individual models for individual lead times and result in more robust forecasts compared to IMS.

For Direct multi-step estimation, Alon et. al. used neural networks for forecasting retail sales [2]. They used a single time series to compare the exponential smoothing, regression, ARIMA and Neural Network models and conclude Neural Networks perform best for multi-period forecasting.

2.6. Insight Gaining from Time Series Similarities

The similarity measures can be leveraged to gain insights regarding the drivers of sales. The main problem in this sense is being able to find similarities without false positives. There is the possibility of visual inspection of intuitive series and the sales, however this is time consuming when there are a lot of series to be inspected.

Lin et. al. propose a method to identify similarities between series quickly [41]. They draw a band with a predetermined width around the original series, and if the tested time series are in this band, they are deemed similar. A landmark method was proposed by Perng et. al.; where "landmark" patterns are identified in series. The test of similarity includes the position and type of the landmark to be matched, which is in line with the intuitive matching of visual inspection [47].

The main drawback of the few methods are the inability to identify dynamic similarities. In retail sales, the environment is quite dynamic with many different drivers.

Chapter 3

BASE METHOD FOR FORECASTING

In this chapter, we describe in detail the forecasting model that has been developed as part of the KUMPEM project. The model provides predictions at the store-category level, given the planned promotion levels by leveraging the information from multitude of stores. The model has been described in detail and results are evaluated in [33].

In this chapter, we provide a summary of the model, because in chapter 4, we will build on this model and provide improvements through behavioral clustering. We will also use this model as the foundation to calculate residual series that are adjusted for seasonal and marketing effects, and provide insights by using Dynamic Time Warping distances to socio-economic series.

3.1. Overview of the Model

The sale forecasting model consists of two layers. The first layer models the store-category sales with seasonality and marketing variables using pooled regression. The second layer provides multi-step ahead forecasts for the residuals of layer 1, using the store-category residuals and the trend-cyclical component of similar stores.

The first layer adjusts the series by eliminating the effect of seasonality and planned marketing; leaving the residuals. Then these residual series are extrapolated in layer

two. The store-category level forecasts are aggregated to higher level, instead of forecasting each aggregation level individually.

The concept behind the model is using a mixture of causal models and extrapolation models. The first layer predicts the time series where there are the seasonal patterns and effects of economic indicators. These series signify the *level,* of the sales. This is the causal part. The resulting series need adjustment to account for trend-cyclical effects.

For the trend-cyclical effects, new series need to be derived. The original series have the levels resulting from marketing expenses, seasonal patterns, and trend-cyclical components. The predictions of the first layer account for the levels of marketing expenses and seasonal patterns. As a result, if the first layer drivers is taken out of the core series, the trend-cyclical effects remain. This "taking out" process brought up the concept of using residuals. The difference between actual series and the predicted series are the residuals, and they represent the trend-cyclical component that we want to model.

3.2. Data Used

The data is provided by the leading retailer in Turkey. The store-category level net sales, marketing expenses, and the format-category level inflation data between January 2007 and December 2011 are given. We have three different main data sets used for forecasting the sales of retail stores across the country. These data sets are store-category level marketing expenses and sales, seasonality and calendar variables and store properties. All of the necessary data are merged into a time series dataset for each individual store-category combination.

3.2.1 Store-Category Level Sales and Marketing Expenses

The first data set is the actual monthly time series for the store-category level net sales, the CRM and Marketing expenses and the monthly price level in the category. The price level indicates the fluctuations and overall pricing of the categories. The CRM and Marketing promotion decisions are the factors that the store managers can control to boost their sales; all the price reduction, 2-for-1 and other sale campaigns' effects are observed with these variables. CRM expenses and Marketing expenses are recorded as two different variables. This is the result of a policy decision from the retailer; if the promotions include a price reduction and/or require the buying of another product to work, they are included in the CRM expenses while all the other overall price reductions are included in the Marketing Expenses. The most important difference is that CRM expenses are added into the net sales when the calculations are made, so those levels directly affect the net sales amount and thus our forecasts.

- "Store Number" is the code number of the stores for identification. There are a total of 971 stores. For the final analysis, there are 336 active stores that are eligible. The eligibility condition is having enough past data of the store to be able to model, i.e. the number of monthly data available.
- "Category" is the number code of the product category. Not all of the stores have the same categories, and some of the categories are too irregular to forecast and/or not necessary to forecast since the marginal amount of sales of the category are too small relative to the total sales of the Store in that time period. A total of seven categories are forecasted. The main aspects of the categories are:
    - Category 1: The highest percentage contributor to the total sales. This category is consistent, meaning it has low seasonal variability, and a consistent trend.

- Category 3: The most consistent category with small linear trends and no seasonality
- Category 4: Contains some of the most seasonal series and this category is highly prone to marketing effects
- Category 5: A category with inherent linear trends and small seasonal trends
- Category 6: This category is expected to be highly affected by the economical and marketing effects that are discussed in detail in the next section
- Category 7: A small sale amount category with no apparent seasonality or trend factors
- Category 8: The category with the lowest levels and magnitude of movement which make it hard to identify patterns.

- The time period ranges from t=1, January 2007: to t=60, December 2011.
- The net sales for a time period is the sum of the sales and the CRM amounts. This is a company policy, and the net sales need to be forecasted in this manner. The data set variables are given in Table 1.

Table 1: The details of the variables in the first data set

Store	The code number for each individual store
Category	The category in each store
T (time period)	The time period of the sales
Net sales	Sum of Sales and the CRM expenses
CRM	CRM campaign expenses
Marketing	Marketing Campaign expenses

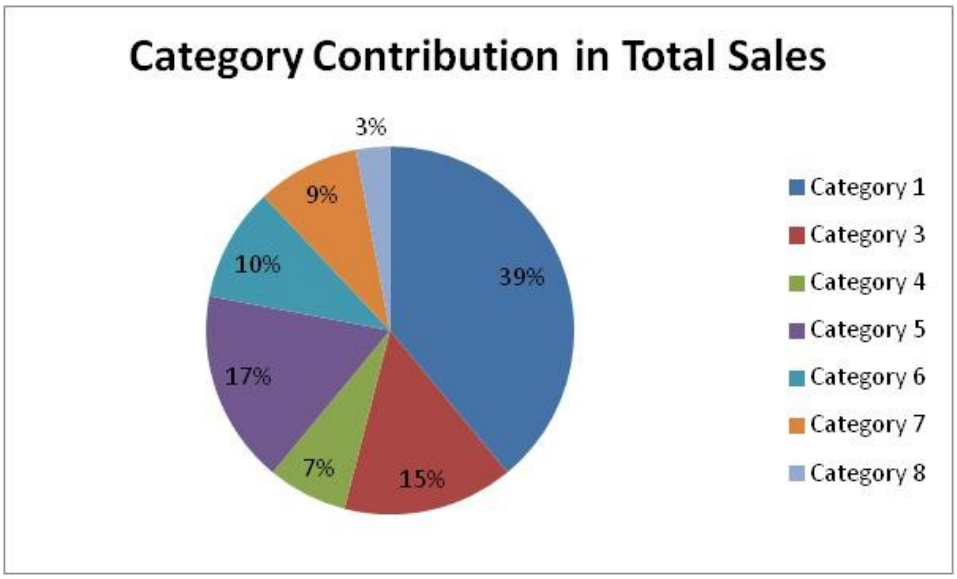


Figure 1: The contribution of different categories in the total sales, averaged over months. (336 Stores)

3.2.2 Seasonality and Calendar Variables

The second data set contains the seasonal, monthly, weekly and special days regarding the individual time periods. The data set has the binary variables of each month and the percentages of total vacation days, Ramazan month, Ramazan holiday and Kurban holiday days in the corresponding month. The sales are affected by the vacations and holidays. Their percentage in a month are used to eliminate the need to put the month length in the analysis individually.

Month variables are binary variables; if t is the corresponding month then that variable is 1, otherwise 0. The variable details are given in Table 2.

Table 2: The explanation of the variables in the second data set

T (time period)	The time period, same as the time period in data set 1
Month variables	Binary variables determining the months, 12 in total (January-December)
Vacation, Ramazan month, Ramazan Holiday, Kurban holiday variables	Percentage of the special days over the month lengths

3.2.3 Store Properties

The third data set contains the individual store properties. The social and economical factors of each store, the format, sub format, the region of each individual store, opening and closing dates and the repair dates are present in this data set. The Store code defines the store and is the same as the store code in data set 1. The Format, sub format, address, region, profitability group, opening and closing dates and the repair dates are store specific data that is given by the company.

The formats should be investigated in further detail as they are used in calculations with the models developed. There are 4 formats and these formats are divided further into 8 sub formats. The format properties are;

- **Format Y:** The main format of the retailer is format Y. This has the most stores out of 336 active ones with 192 stores. This format is divided into 3 sub formats according to their sizes and sales volumes.
- **Format Z:** Format Z is the secondary format of our retailer. They are again divided into 3 sub formats according to their sizes.131 of the 336 stores belong to this format.

- **Format X:** This format has the highest service quality and prices. There are no subdivisions for this format, as it is considered an exclusive format with 7 store out of 336.
- **Format W:** Actually the part of the regular sub formatting theme of the retailer, they are the largest stores, generally including sub contracting companies or recreative activities included in the stores. Due to their size, their numbers are limited and 6 of the 336 stores are format W.

There are a total of 6 geographical regions that each store is member of. The number of stores in each region is different, with R and S regions being the main regions with the most stores with 89 and 161 stores respectively.

We merged these data sets for the analyses. The first and second data sets are merged by the time period and finally the third data set is merged by the Store code.

The data set is partitioned into training and test periods to be able to use previously unobserved data to test the models. The training period has to be large enough that the models would be as accurate as possible, while the test period should again be relatively large to best test the model. The first 4 years; January 2007-December 2010; are used as the training period, and year 5; January 2011-December 2011; as the test period.

3.3. Seasonality and Marketing Effects Model – Layer 1

Monthly seasonality is a big contributor to the sales fluctuations, along with vacation days, pre-determined special dates, national and religious holiday. Another factor is the marketing and inflation variables. Marketing has a direct impact on retail sales and knowing how much the sales would be affected depending on the amount of money invested in marketing is invaluable. The first layer regression model takes the

seasonal effects, marketing effects and inflation effects into account and models the natural logarithm of store-category sales, denoted by $\ln(y_t)$, where t is the time period. The regression model uses seasonality and calendar variables, shown as a vector $\boldsymbol{H}_t$, and marketing variables $\boldsymbol{M}_t$. Both marketing variables and the net sales are used as natural logarithms as established in marketing literature. Finally there is the residual term ε_t. The base log-linear formulation of store-category level sales is given (1).

$$\ln(y_t) = \alpha + \boldsymbol{\beta} * \boldsymbol{H}_t + \boldsymbol{\gamma} * \boldsymbol{M}_t + \varepsilon_t \quad (1)$$

However, the ratio of number of observations to the number of variables is low for the store category level. The individual seasonal parameters have only one observation per year. Marketing variables may show single high peaks in otherwise steady time series, making those points difficult to model and forecast. Thus store-category level models provide unreliable model parameters.

To circumvent this problem, the store-category time series are pooled to provide more observations per regression model. Properly pooling series, e.g. series with a common trend parameter, would decrease the variance due to improved sample size; however it also introduces a bias due to the inherent differences in the series that are pooled. There are different types of pooling including 1) correlational co-movement groups, 2) clustering locations based on their characteristics and 3) pooling from expert judgment [21]. Store segmentation through expert judgment is already present with different formats and profitability groups. These groups are expected to act in a similar fashion and thus may be used for pooling.

In the pooled model, each individual model assigns the same set of marketing response parameters to all stores in the pool. We expect the customer response to the marketing variables to differ based on customer profitability group decided by expert judgment, product category and store formats. These groupings are in line with the price elasticity factors determined by Bijmolt et. al. [9]. They state that product

categories and market properties are major factors contributing to price elasticity. Therefore, profitability groups and store formats for market properties and product category are used for pooling.

Marketing and economical effects are pooled, however the seasonal variables have a more diverse effects on individual stores. As a result we use seasonal parameters for individual stores with indicator variables. The indicator variables consider both the seasonal effect and which store that effect was observed in. This way, individual store variability regarding seasonal variables is taken into consideration.

The pooled model is given in equation (2). This model is similar to the equation (1) with indices defined for the determined pools. The fixed parameter α_{ij} adjusts for the difference between individual store-category sale series in each pool. The vector $\boldsymbol{\gamma}_j$ is category specific and contains the marketing effect parameters common to the pooled stores for each category. Differences between the store seasonality and calendar effects among the pooled stores are accounted for by the store-category specific parameter vector $\boldsymbol{\beta}_{ij}$.

$$\ln(y_{ijt}) = \alpha_{ij} + \boldsymbol{\beta}_{ij} * \boldsymbol{H}_t + \boldsymbol{\gamma}_j * \boldsymbol{M}_{ijt} + \varepsilon_{ijt} \qquad (2)$$

For estimating the parameters in the model, ordinary least squares (OLS) approach is used. The errors ε_{ijt} are not assumed to be independent; however as long as the independent variables are exogenous the parameters can be estimated consistently with OLS [28]. Consistency in the estimation of parameters state that as more observations are given, the estimated parameter converges to the "true" value of the parameter. The actual dependency between variables does not affect the consistency of estimation [28]. The pooled models increase the number of observations for each model, which should help improve the reliability of the estimators.

3.4. Layer 2 Model

The residuals of the first layer model ε_{ijt} are representative of store-category sales that are adjusted for seasonality, marketing and inflation effects. These residual time series are extrapolated with lead time and they are pooled by business hierarchy for the regression models in layer two.

Two options have been evaluated regarding inputs to the residual extrapolation in the second layer.

3.4.1 Own residuals only

The second layer provides forecasts for time period t_{0+l} using information available at a specific time t_0. The information used is store-category's own residuals for these models. 12 models are estimated representing up to 12 month ahead forecasts, all using the same information available at time t_0.

$$\varepsilon_{ij\,t_0+l} = \theta_0 + \theta_{l1}\varepsilon_{ijt_0} + \theta_{l2}\varepsilon_{ij\,t_0-1} + \theta_{l3}D2\left(\varepsilon_{ijt_0}\right) + \theta_{l4}D3\left(\varepsilon_{ijt_0}\right) + \theta_{l5}D6\left(\varepsilon_{ijt_0}\right) + \theta_{l6}\varepsilon_{ij\,t_0+l-12} + v_{ijt_0+l} \quad (3)$$

Equation (3) shows the base version of the second layer model. It models the category-store specific residuals of equation (2) at timet_{0+l}, using the information that is available as of timet_0. It uses the last two observed residuals, the last residual of the predicted month, and the estimates of the residual trend with the last four, six and twelve observations. The trend estimators are represented with $Dk(\varepsilon_t)$. The trend estimators derived for representation of the short, mid and long term changes, to provide the model with more parameters representative of the overall movement of the series. $Dk(\varepsilon_t)$ function is defined as follows.

$$Dk(\varepsilon_t) = \frac{\sum_{m=0}^{k-1} \varepsilon_{t-m}}{k} - \frac{\sum_{m=0}^{k-1} \varepsilon_{t-k-m}}{k} \quad (4)$$

In equation 4, the k represents the length of the trend indicator i.e. k=2 states the four-month trend indicator, k=3 states the three-month trend indicator and k=6 states the twelve-month indicator. Note that each category-store has a different function to extrapolate the residual for each lead time *l*, meaning each lead model uses the same variables but with different coefficients. Shorter lead times are expected to depend more on closer observations and the short term trend estimator D2, while longer lead time models are expected to put more emphasis on the long term trend estimator D6. The irregular component of the store-category time series is denoted by v_{ijt} .

3.4.2 Information borrowing

The residual series from the first model are adjusted for seasonality and marketing effects but there are still other effects: there may be large scale economical issues affecting the nation or company as a whole, individual store or format specific changes, regional effects, category related changes, or customer segmentation effects; to name a few.

Using dimensional residual series are proposed. Different dimensions provide aggregated residuals series. This wider look may provide insights on larger scale economic and company policy issues mentioned. The dimensional series are the residuals series that are taken from different levels of aggregations, for example the format-category dimensional residuals would represent the trend-cyclical effects in format-category aggregation level.

The average residuals of stores with common characteristics are expected to reflect the trend and cyclical effects of the stores that are akin. Averaging should reduce the variability resulting from individual stores.

In addition to store residuals, the average residuals of the stores that share a common characteristic *d* with each store are used in the second layer. The final model can be seen in (5). Here $\varepsilon_{\bar{d}jt}$ is the dimensional residual of category j at time *t;* this category average is across stores that share characteristic *d* with the store; e.g. same format.

$$\varepsilon_{ij\,t_0+l} = \theta_{jl0} + \theta_{jl1}\varepsilon_{ijt_0} + \theta_{jl2}\varepsilon_{ij\,t_0-1} + \theta_{jl3}D2\left(\varepsilon_{ijt_0}\right) + \theta_{jl4}D3\left(\varepsilon_{ijt_0}\right) + \theta_{jl5}D6\left(\varepsilon_{ijt_0}\right) + \theta_{l6}\varepsilon_{ijt_0+l-12} + \sum_d \varphi_{djl1}\varepsilon_{\bar{d}jt_0} + \varphi_{djl2}\varepsilon_{\bar{d}jt_0-1} + \varphi_{djl3}D2\left(\varepsilon_{\bar{d}jt_0}\right) + \varphi_{djl4}D3(\varepsilon_{\bar{d}jt_0}) + \varphi_{djl5}D6(\varepsilon_{\bar{d}jt_0}) + \varphi_{djl6}\varepsilon_{\bar{d}j\,t_0+l-12} + v_{ij\,t_0+l}$$

(5)

3.5. Accuracy of Base Regression Method

The first layer considers the seasonal and marketing effects and the second layer extrapolates the residuals to take the trend-cyclical components into consideration.

The benchmark methods facilitate forecast accuracy comparison. The main benchmark method is Winter's exponential smoothing which is used by the retailer for forecasting. In this benchmark, aggregate level series are forecasted individually; and thus the sum of forecasts that make up the region do not match the forecast for the region. The comparison between the accuracies of first layer and both of the layers together are also provided.

One important aspect of the retail forecasts is the aggregate dimensional forecasts. Store, format, category and region level forecast are necessary for both financial

decisions and incentive calculations. The aggregate level forecast accuracy is needed and provided for both the benchmark methods and the proposed method. The store-category level forecast accuracy for all the 12 leads are given in Table 1. Mean Absolute Percentage Errors and Mean Percentage Errors are provided. MAPE signifies the absolute difference between the prediction and the series, so the overall accuracy of the predictions. MPE signifies if the model constantly over or under-predicts. The formulas for MAPE and MPE are given in (6)and (7).

$$MAPE = \frac{\sum_{i \in Predictions} abs(\frac{Actual_i - Prediction_i}{Actual_i})}{n} \quad (6)$$

$$MPE = \frac{\sum_{i \in Predictions} \frac{Actual_i - Prediction_i}{Actual_i}}{n} \quad (7)$$

For significance analysis, we want to check whether the improvement is significant. A hypothesis test is done to calculate the significance. First of all, we calculate the difference between MAPE values, as given in (8).

$$MAPE\ Difference_{ij} = Method\ 1\ MAPE_{ij} - Method\ 2\ MAPE_{ij} \quad (8)$$

$$i \in Stores, j \in Categories$$

Then, hypothesis testing is done to find whether the mean of MAPE differences is significantly different from 0. Test parameters are;

$$H_o: \mu = 0$$

$$H_a: \mu \neq 0$$

The calculated t-values for each lead, each method and MEAN of the leads is given, along with the p-values for the two-tailed distribution for %95 confidence interval.

The national level values can not be calculated as there is only one observation, a single series, in the utmost aggregation level.

Table 3: Store-Category level forecasts' MAPE values

Store Category MAPE				
Lead	Winter's Exponential Smoothing	Layer 1	Layer 2 with Own Residuals	Layer 2 with Dimensional Residuals
1	13.85%	17.09%	13.08%	12.49%
2	14.66%	18.74%	14.99%	14.37%
3	16.84%	19.66%	16.50%	15.83%
4	22.26%	22.91%	20.17%	18.51%
5	19.64%	20.82%	18.23%	17.17%
6	20.42%	19.14%	16.90%	16.11%
7	21.73%	18.55%	16.92%	16.28%
8	24.18%	21.18%	19.50%	18.69%
9	26.50%	22.32%	20.47%	20.17%
10	29.19%	23.92%	22.45%	21.51%
11	27.41%	22.68%	20.87%	20.24%
12	27.99%	24.24%	23.24%	22.32%
MEAN	22.05%	20.94%	18.61%	17.81%

Table 4: Store-Category level forecasts' MPE values

Store Category MPE				
Lead	Winter's Exponential Smoothing	Layer 1	Layer 2 with Own Residuals	Layer 2 with Dimensional Residuals
1	1.14%	-5.19%	-3.65%	-4.37%
2	-3.21%	-7.41%	-5.84%	-5.50%
3	0.04%	0.20%	0.81%	0.71%
4	-4.11%	-7.27%	-5.98%	-4.32%
5	-0.89%	-3.83%	-2.66%	-2.85%
6	0.54%	-3.35%	-2.26%	-3.25%
7	4.06%	-1.62%	-1.20%	-2.80%
8	5.71%	-6.04%	-5.32%	-5.59%
9	-7.75%	-7.14%	-6.06%	-7.44%
10	0.06%	-5.40%	-4.71%	-6.31%
11	-0.15%	-3.20%	-2.03%	-3.38%
12	14.28%	12.62%	12.13%	8.65%
MEAN	0.81%	-3.14%	-2.23%	-3.04%

Table 5: Store-Category level MAPEs' differences' t-values (n=2352)

Store-Category MAPE Difference t-values			
Lead	**Winters-Layer 1**	**Layer 1-Layer 2 w/ Own Resid.**	**Layer 2 w/ Own Resid. - Layer 2 w/ Dimensional Resid.**
1	3.59	21.25	4.49
2	7.46	20.10	2.15
3	10.06	20.74	4.16
4	12.21	18.44	10.18
5	14.97	18.39	6.58
6	21.27	17.13	4.28
7	17.29	12.96	3.95
8	21.54	12.54	4.25
9	22.52	11.43	-
10	21.67	11.12	4.41
11	23.74	11.09	1.60
12	23.14	10.82	4.48
MEAN	59.04	53.59	13.83

t-distribution value of 0.05 p-value (two-tailed): 1.96112353

Table 6: Store-Category level MAPEs' differences' p-values (n=2352)

Store-Category MAPE Difference p-values			
Lead	**Winters-Layer 1**	**Layer 1-Layer 2 w/ Own Resid.**	**Layer 2 w/ Own Resid. - Layer 2 w/ Dimensional Resid.**
1	0.0003	0.0000	0.0000
2	0.0000	0.0000	0.0319
3	0.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0001
8	0.0000	0.0000	0.0000
9	0.0000	0.0000	-
10	0.0000	0.0000	0.0000
11	0.0000	0.0000	0.1101
12	0.0000	0.0000	0.0000
MEAN	0.00	0.00	0.00

Tables 3-6 compare the MAPE, MPE, t-values and s-values for store-category level forecasts. Rows indicate the leads and columns indicate alternate methods tested. Observations from Table 3;

- At store-category level forecasts, our two-layer model with dimensional residuals included in the second layer results in the highest accuracy. The comparison of MAPE values across different lead times are given along the rows of Table 3.
- By investigating the accuracy of different leads, we conclude that as the lead time increases the accuracy decreases. However this decrease has a higher rate in exponential smoothing predictions than the proposed model. This lower decay speed property also provides far more accurate results as the lead time increases.
- Layer 1 result in less accurate forecasts than the exponential smoothing in the first few leads, and this proves the necessity and benefits of layer 2.
- From Table 4, it is observed that exponential smoothing has the least bias in store-category level, and the proposed method is over-estimating the sales with the exception of lead 12.
- Table 5 and 6 indicate the t-values and p-values for the difference between methods. The improvements are significant for all the leads except lead 11 between using dimensional residuals in layer 2 and only using own residuals.

These observations are for store-category level forecasts; aggregated sales may result in different observations. MAPE, MPE, t-values and p-values of store, format and national level sales are provided in tables 7-16.

Table 7: Store level forecasts' MAPE values

Store MAPE				
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**
1	8.44%	9.88%	7.56%	7.23%
2	8.37%	10.78%	8.44%	8.09%
3	10.79%	12.78%	10.69%	9.93%
4	14.58%	14.71%	12.95%	12.08%
5	12.95%	13.00%	11.46%	10.85%
6	12.52%	10.96%	9.68%	9.29%
7	13.95%	11.81%	10.79%	10.34%
8	16.34%	12.81%	12.03%	11.46%
9	17.12%	14.64%	13.49%	12.76%
10	19.34%	15.27%	14.55%	13.85%
11	17.15%	14.77%	13.94%	13.43%
12	19.49%	18.57%	17.83%	16.07%
MEAN	14.25%	13.33%	11.95%	11.28%

Table 8: Store level forecasts MPE values

Store MPE				
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**
1	3.49%	-3.14%	-2.33%	-2.85%
2	-0.87%	-5.16%	-4.37%	-4.48%
3	2.88%	3.08%	3.09%	2.46%
4	-0.86%	-4.46%	-3.66%	-2.13%
5	2.19%	-1.15%	-0.64%	-1.15%
6	2.08%	-2.07%	-1.54%	-1.91%
7	5.31%	-0.25%	-0.14%	-1.28%
8	7.53%	-4.46%	-3.80%	-2.87%
9	-5.85%	-3.84%	-2.98%	-3.36%
10	2.74%	-1.44%	-0.92%	-2.33%
11	0.46%	-1.08%	-0.29%	-1.03%
12	13.69%	14.01%	13.70%	11.39%
MEAN	2.73%	-0.83%	-0.32%	-0.79%

Table 9: Store level MAPEs' differences t-values (n=336)

Store MAPE Difference t-values			
Lead	**Winters-Layer 1**	**Layer 1-Layer 2 w/ Own Resid.**	**Layer 2 w/ Own Resid. - Layer 2 w/ Dimensional Resid.**
1	-	9.30	1.19
2	-	10.50	-
3	-	11.18	3.16
4	-	7.48	1.97
5	-	7.42	1.26
6	2.08	6.52	-
7	3.58	5.58	0.54
8	4.70	3.76	2.51
9	2.84	4.72	1.92
10	4.03	3.68	2.08
11	2.55	2.99	1.26
12	2.40	7.82	10.65
MEAN	4.30	21.95	6.76

t-distribution value of 0.05 p-value (two-tailed): 1.96704937

Table 10: Store level MAPEs' differences p-values (n=336)

Store MAPE Difference p-values			
Lead	**Winters-Layer 1**	**Layer 1-Layer 2 w/ Own Resid.**	**Layer 2 w/ Own Resid. - Layer 2 w/ Dimensional Resid.**
1	-	0.00	0.23
2	-	0.00	-
3	-	0.00	0.00
4	-	0.00	0.05
5	-	0.00	0.21
6	0.04	0.00	-
7	0.00	0.00	0.59
8	0.00	0.00	0.01
9	0.00	0.00	0.06
10	0.00	0.00	0.04
11	0.01	0.00	0.21
12	0.02	0.00	0.00
MEAN	0.00	0.00	0.00

Table 11: Format level forecasts MAPE values

Format MAPE				
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**
1	14.21%	3.23%	2.77%	2.72%
2	13.39%	4.81%	4.76%	5.14%
3	15.21%	4.44%	4.36%	4.18%
4	15.44%	3.12%	3.44%	2.37%
5	17.84%	2.72%	2.82%	2.13%
6	15.49%	5.36%	4.92%	3.66%
7	20.97%	2.52%	2.39%	1.87%
8	22.78%	5.34%	4.86%	3.78%
9	10.71%	5.15%	4.68%	3.21%
10	17.24%	5.45%	5.51%	4.21%
11	15.20%	9.71%	8.71%	7.74%
12	21.52%	15.97%	15.44%	12.71%
MEAN	16.67%	5.65%	5.39%	4.48%

Table 12: Format level forecasts MPE values

Format MPE				
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**
1	14.21%	-2.64%	-2.68%	-1.98%
2	10.09%	-4.81%	-4.76%	-5.14%
3	13.53%	3.80%	3.16%	2.89%
4	15.39%	0.31%	0.15%	0.76%
5	15.42%	0.33%	0.31%	-0.05%
6	13.88%	-1.84%	-1.63%	-1.87%
7	16.77%	-0.40%	-0.51%	-1.25%
8	15.67%	-4.79%	-4.32%	-2.80%
9	5.77%	-2.78%	-2.34%	-1.64%
10	16.68%	5.45%	5.51%	4.21%
11	5.00%	-5.72%	-4.50%	-4.48%
12	21.47%	15.97%	15.44%	12.71%
MEAN	13.66%	0.24%	0.32%	0.11%

Table 13: Format level MAPEs' differences t-values (n=4)

Format MAPE Difference t-values			
Lead	**Winters-Layer 1**	**Layer 1-Layer 2 w/ Own Resid.**	**Layer 2 w/ Own Resid. - Layer 2 w/ Dimensional Resid.**
1	3.59	21.25	4.49
2	7.46	20.10	2.15
3	10.06	20.74	4.16
4	12.21	18.44	10.18
5	14.97	18.39	6.58
6	21.27	17.13	4.28
7	17.29	12.96	3.95
8	21.54	12.54	4.25
9	22.52	11.43	-
10	21.67	11.12	4.41
11	23.74	11.09	1.60
12	23.14	10.82	4.48
MEAN	11.58	3.58	0.97

t-distribution value of 0.05 p-value (two-tailed): 2.77644511

Table 14: Format level MAPEs' differences p-values (n=4)

Format MAPE Difference p-values			
Lead	**Winters-Layer 1**	**Layer 1-Layer 2 w/ Own Resid.**	**Layer 2 w/ Own Resid. - Layer 2 w/ Dimensional Resid.**
1	0.04	0.00	0.02
2	0.00	0.00	0.12
3	0.00	0.00	0.03
4	0.00	0.00	0.00
5	0.00	0.00	0.01
6	0.00	0.00	0.02
7	0.00	0.00	0.03
8	0.00	0.00	0.02
9	0.00	0.00	-
10	0.00	0.00	0.02
11	0.00	0.00	0.21
12	0.00	0.00	0.02
MEAN	0.00	0.00	0.34

Table 15: National level forecasts MAPE values

National MAPE				
Lead	Winter's Exponential Smoothing	Layer 1	Layer 2 with Own Residuals	Layer 2 with Dimensional Residuals
1	17.77%	1.32%	1.25%	1.23%
2	14.27%	3.28%	3.06%	3.20%
3	17.49%	5.89%	5.21%	4.34%
4	17.67%	1.33%	1.69%	2.21%
5	17.46%	1.56%	1.67%	1.11%
6	14.69%	1.53%	1.25%	1.51%
7	18.23%	0.99%	0.90%	0.04%
8	18.88%	3.40%	2.96%	1.78%
9	6.83%	2.83%	2.23%	2.07%
10	17.68%	4.57%	4.62%	3.53%
11	12.09%	0.97%	1.78%	1.11%
12	20.77%	14.48%	13.86%	11.58%
MEAN	16.15%	3.51%	3.37%	2.81%

Table 16: National level forecasts MPE values

National MPE				
Lead	Winter's Exponential Smoothing	Layer 1	Layer 2 with Own Residuals	Layer 2 with Dimensional Residuals
1	17.77%	-1.32%	-1.25%	-1.23%
2	14.27%	-3.28%	-3.06%	-3.20%
3	17.49%	5.89%	5.21%	4.34%
4	17.67%	1.33%	1.69%	2.21%
5	17.46%	1.56%	1.67%	1.11%
6	14.69%	-1.53%	-1.25%	-1.51%
7	18.23%	0.99%	0.90%	-0.04%
8	18.88%	-3.40%	-2.96%	-1.78%
9	6.83%	-2.83%	-2.23%	-2.07%
10	17.68%	4.57%	4.62%	3.53%
11	12.09%	0.97%	1.78%	1.11%
12	20.77%	14.48%	13.86%	11.58%
MEAN	16.15%	1.45%	1.58%	1.17%

The aggregated forecasts' MAPE values given in Tables 7, 11 and 15 show the accuracy of our model in higher dimensions.

- As the aggregation level increases; proposed model provides much better forecasts than Winter's Exponential smoothing. This result is despite the fact that the proposed models aggregate the lower level forecasts while Winter's

Exponential smoothing is used to forecast aggregated series within themselves.

- There is a high spike of MAPE values of the proposed model in lead 12. It is presumed that this spike is related to an effect or incident related to the specific time period forecasted; i.e. t=60, December 2011; that is difficult to find let alone incorporate in the model. This spike is more apparent as the aggregation level increases.
- From the MPE values given in 8,12 and 16, it is seen that as the aggregation level increases, the bias of the proposed models decrease while bias of exponential smoothing increases. This is most probably due to the fact that exponential smoothing forecasts the aggregated series while the proposed method aggregates the individual forecasts to higher levels
- From tables 9, 10, 13 and 14 we see that the improvements stay significant as the aggregation level increases, however the p-values increase as the aggregation level increases. This is possibly due to the sample size decreasing with increase in aggregation level

These tables have shown that the proposed model performs better than Winter's exponential smoothing for 12 months ahead predictions. Aggregated base method forecast accuracy is also better than aggregated exponential smoothing forecasts. Finally the improvements are significant, although p-value increases as the aggregation level increases.

Chapter 4

IMPROVING REGRESSION ACCURACY THROUGH BEHAVIORAL CLUSTERING

In this chapter, we describe the proposed method to improve the forecasting model, and the intuition behind it. We implement behavioral clustering to pool the data instead of using the business hierarchy clustering. To quantify the similarity of the series, we used Dynamic Time Warping.

4.1. Motivation

The second layer of the model described in Chapter 3 uses the local store-category residuals, the trend cyclical components of similar stores to extrapolate the residuals and to avoid over fitting in the second layer, pools the observations according to category, sub format and profitability groups instead of estimating individual store-category level models. The main reasons for selecting this specific grouping are convenience and ease of communication and interpretation in business terms. Series that are in the same category, sub format and profitability group are expected to be similar in their reactions to environmental effects.

We propose that there may be a better grouping of stores with less in-group variability than the business hierarchy. Even though the stores belong to the same hierarchical groups, there may be exceptional stores that have different trend-cyclical components. We expect that instead of using the default properties of the stores, a

grouping according to the behavior similarity of the stores will improve the accuracy of the forecasts by decreasing the in-group variability. We use the past residual time series from the first layer to group the stores which improves the quality of the inputs into the regression. If the grouped stores are more similar, the estimation of the model is easier and more accurate.

There are some drawback and costs of implementing behavioral pooling as well. The model should be an interpretable model to understand the impact of known drivers, i.e. marketing expenses, seasonal parameters and economical indicators. When business hierarchy is implemented, the interpretation of the parameters is much easier to define in business terms. It is more logical and business friendly when a specific format is addressed. The behaviorally similar stores and groups may be of different characteristics formidably different from the business point of view.

To counteract the problem of interpretability, the behavioral pooling methods are implemented only in the second layer model. The first layer takes into account the known variables and they are still grouped according to business hierarchy. The second layer do not need to be as transparent as it is an extrapolation method without outside effects. There are no outside variables to interpret in the second layer. Thus the cost of interpretability is reduced to only technicality in layer two. This is the main reason behind applying the behavioral pooling solely on the second layer.

Another reason for applying behavioral pooling only on the second layer is that proposed method for similarity measure is Dynamic Time Warping, which has a tendency to match noise-related oscillations. Pooling the residuals of the first layer provides a more reliable similarity measure. If the input variables of the first layer such as marketing expenses' effects are not removed, that may lead to behavioral pools revolving around the business decisions instead of reactions to those decisions.

The concept of behaviorally pooling both of the layers and creating models solely revolving around the concept of behavioral pooling is a possible future research venue.

4.2. Methods and Options for Time Series Clustering

For clustering time series, there are 2 main factors to determine. The first one is the similarity measure. The second factor is the clustering method.

The usual method of similarity between two series is the Euclidian distances. The formula for Euclidian distances are given in (9)

$$d(x, y) = \sqrt[n]{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (9)$$

However Euclidian distances result in some loss of characteristics in the compared series because it formulates the total distance between two series. As a results, highly similar parts of the series can be neutralized by another highly dissimilar part. This is disadvantageous as similar sub patterns in the series are important to identify as well. Euclidian distance also compares only the matching time periods between series and do not consider adjacent time periods. DTW on the other hand is able to find similar patterns across varying time periods so the similar parts can be matched even if they are observed in different time periods. This way similar patterns are not missed. DTW is able to take into consideration the lagged and out-of-sync series' similarities.

We have decided to use DTW as a distance measure for similarity. The reason we use DTW is that in retail sales, the time component is a defining factor in the movements of individual time series. Each store has different reaction times and/or speeds with the same inputs.

The second factor to determine for the clustering of the series is the clustering method. K-means clustering is a method possible for the clustering of the series. k-means algorithm uses the mean distance of observations to the center point. The pseudo code for k-means is given in figure 2.

Inputs:
$I = \{i_1, ..., i_k\}$ (Instances to be clustered)
n (Number of clusters)
Outputs:
$C = \{c_1, ..., c_n\}$ (cluster centroids)
$m : I \rightarrow C$ (cluster membership)

procedure **KMeans**
Set C to initial value (e.g. random selection of I)
For each $i_j \in I$
$m(i_j) = \underset{k \in \{1..n\}}{\text{argmin}}\, distance(i_j, c_k)$
End
While m has changed
For each $j \in \{1..n\}$
Recompute i_j as the centroid of $\{i | m(i) = j\}$
End
For each $i_j \in P$
$m(i_j) = \underset{k \in \{1..n\}}{\text{argmin}}\, distance(i_j, c_k)$
End
End
return C
End

Figure 2: Pseudo code for k-means clustering algorithm [45]

However, for time series clustering with DTW as a distance measure, k-means does not provide good solutions [46]. The reason behind this is that time series averaging method implemented (DTW) provide miscalculations when means of the distances are calculated. As an example to these miscalculations; if two series are highly similar

in a specific short span of time but otherwise are not so similar, Euclidian distance would not be able to capture that similarity period as it is neutralized by the long periods of less similarity. To counteract this problem, k-medoids clustering are proposed and we have implemented k-medoids clustering as well. K-medoids is better than k-means clustering when DTW is used as the distance measure. There are two reasons for this;

1) k-means averages the series and finds the cluster centers at the center of the clusters. The series are averaged to find the centers which we want to prevent with DTW in the first place
2) DTW distances may be asymmetrical. This means identifying medians for the cluster centers is more suitable instead of averaging the distances; which would result in loss of information.

The pseudo code for k-medoids is given in figure 3.

---

1. Initialize: randomly select k of the n data points as the medoids
2. Associate each data point to the closest medoid
3. For each medoid m
    1. For each non-medoid data point o
        1. Swap m and o and compute the total cost of the configuration
4. Select the configuration with the lowest cost.
5. repeat steps 2 to 4 until there is no change in the medoid.

Figure 3: The pseudo Code for k-medoids clustering algorithm [55]

4.3. Dynamic Time Warping

A time series is a series of data with a time index for indentifying the time of each observation. In Dynamic Time Warping, two time series are matched together. The main difference of Dynamic Time Warping (DTW) from other similarity/matching methods is that DTW is able to "warp" the series in time. Warping means to elongate or shorten a part or all of the series in time, thus enabling discovery of similar patterns regardless of time synchronization or pattern speed. This property makes DTW able to discover slower or faster moving patterns as well as lagged series.

4.3.1 Dynamic Time Warping Algorithm Basics and Formulation

Dynamic Time Warping tries to match and fit two time series together. One of these series is called a *reference* series, and the other one is a *query* series. The query series are warped to fit the reference series [27].

$X=(x_1,...,x_N)$ is the query series.

$Y=(y_1,...,y_M)$ is the reference series.

It is also assumed a nonnegative dissimilarity function is defined between each element x_i and y_j:

$$d(i,j)=f(x_i,y_j)>=0 \quad (10)$$

The function f is generally defined as the Euclidian distance, however some alternatives are also possible. One thing to note that will be described in later stages is that Normalization of the series are a beneficial part of this distance calculating formula. Normalization eliminates the magnitudal differences between series, making

the distance calculated from this formula more reliable as it solely focuses on pattern similarity.

We want to warp and match the query series in such a way that it is most similar to the reference series. The core of this warping is the "warping curve", which is calculated with the formula;

$\emptyset(k) = \left(\emptyset_x(k), \emptyset_y(k)\right)$ with

$\emptyset_x(k) \in \{1..N\}$,

$\emptyset_y(k) \in \{1..M\}$ (11)

In this formulation, "k" is the number of observations in the series. The individual functions $\emptyset_x(k)$ and $\emptyset_y(k)$reassign each time indices of X and Y respectively, to formulate the warping. As a result $\emptyset_x(k)$ and $\emptyset_y(k)$ show to which point an individual observation is matched to: for example $\emptyset_x(4) = 6$ would state the fourth observation of the series X is assigned to the 6th member of the series Y. Given these values, the warping distance is calculated by;

$$d_\emptyset(X,Y) = \sum_{k=1}^{T} d\left(\emptyset_x(k), \emptyset_y(k)\right) * \frac{m_\emptyset(k)}{M_\emptyset} \quad (12)$$

$m_\emptyset(k)$ is per-step weighing constant. $M_\emptyset$ is the normalization constant, which makes the warping paths of the matching series comparable. $m_\emptyset(k)$ and $M_\emptyset$ are used to specify the matched members of the series so the distance function can properly calculate the distance between those members.

Different constraints are generally imposed on $\emptyset_x(k)$ and $\emptyset_y(k)$to ensure logical warpings. Most common one is the monotoniousity constraints which ensures the time ordering of the series are preserved;

$$\phi_x(k+1) \geq \phi_x(k)$$
$$\phi_y(k+1) \geq \phi_y(k) \qquad (13)$$

The objective of the DTW algorithm is to find the warping path that brings the series as close together as possible. The formula of the objective function is:

$$D(X,Y) = \min_{\phi} d_{\phi}(X,Y) \qquad (14)$$

The resulting "alignment" series is the defining result of the DTW. Some graphical examples of this alignment are:

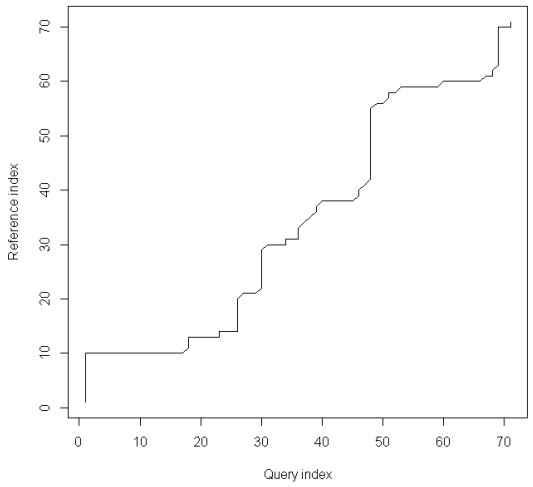


Figure 4: An example alignment graph for DTW

From these alignments, we are able to warp the query. We are also able to get a sense of similarity between R and Q from the graph of the alignment; each match between query and the reference can be identified. For example, in figure 4, we see that the first 10 elements of the referenc series are matched with the first element of the query series. If we use the DTW algorithm to match to identical series, the alignment is a

diagonal line without any deviation. So the closer the alignment is to a diagonal line, the more similar the series are.

4.3.2 Options in DTW

There are different options when applying DTW. Adjusting these options is necessary to derive accurate similarities.

One of the most important options in DTW is specifying whether the beginning and the end of the series are going to be fixed when matching. By default, DTW matches the first and last elements of both series to their respective counterparts and warps the elements in between. We want the possibility of defining lagged similarities and fixing the beginnings and ends would limit this opportunity.

Warping windows are binding corridors that limit how much the warping alignment can deviate from the diagonal line. The benefit of using these windows is unlikely warpings or warpings where a large proportion of the query is matched to a small proportion of the reference can be prevented. There are 3 different warping windows applicable in the R environment where we implemented DTW. These windows are;

- **Sakoe Chiba:** This method creates an alignment line that matches elements in the series with corresponding time indices together. A 45 degree diagonal alignment line is created, and the window size gives a confidence band around this line. This method is inefficient when the series compared are of different lengths, but in our experiments this is not the case.
- **Slanted Band:** The slanted band calculates a "perfect alignment" by creating a diagonal alignment line between the first and last elements of query and reference. The difference of this window from Sakoe Chiba is that; when the series are not of identical length, this band is not 45 degrees. Slanted band

proactively warps the shorter series to the same length then adjusts the series. The window size is a confidence band with width w. This method is useful when the series compared are of different length. However in our experiments the series are of equal length, thus rendering this advantage obsolete. After initial experimenting, Sakoe Chiba and Slanted Band proved to result in the same alignments.

- **Itakura Parallelogram:** This method uses pre-defined Itakura Parallelogram as a band for the alignment. The window size is the farthest point of the band; it is narrower as the beginning and end points are fixed. This method is not applicable in our experiments because of the fact that we want open ended alignments.

From these options, we have used open beginning and open ended alignments. This way, lagged similarities at the beginning and at the end of the series are also accounted for. For the alignment window, we used Sakoe-Chiba window with window size of ±2. This width gives each observation 5 different options to be matched with.

Aside from the options of DTW, we have decided on not using the residuals from the test period of layer 1 regression; they are used for the model building and thus are not representative of the actual fit of the model to the training data. Also series are smoothed with centered moving average of 5 periods. The reason behind this is that series exhibit erratic behavior and DTW is prone to matching series with high oscillation even when there is not a great similarity. Finally, the series are normalized to mean 0 and standard deviation 1, to lower the magnitudal difference between series. This is proven to improve the accuracy of similarity measures. The DTW distances between all of the series are calculated and a "dissimilarity matrix" is achieved, containing the DTW distance of each series to all the other series. One thing to note is that this dissimilarity Matrix may not be symmetrical.

4.4. Clustering with DTW

We used DTW as the dissimilarity measure and k-medoids as the clustering method. The number of clusters should be determined. The current business hierarchical grouping results in 27 groups of stores.

In determining the clusters, two properties are important. The first one is the average in-cluster dissimilarity, which states how homogenous the clusters are. The formula for dissimilarity is given in (15).

$$\frac{\sum_{i=1}^{k} d(x_m, x_i)}{n} \tag{15}$$

Where X_m is the medoid of the cluster, X_i are the series belonging to the cluster. $d(X_m,X_i)$ states the DTW distance between the cluster medoid and the member series. k is the number of series belonging to the cluster. The lower this value is, the more homogenous the clusters are.

The second factor is separation, which states the distance between clusters. The formula for separation is given in (16).

$$\text{Min}(d(x_i, x_j)) \ , \ \forall\, x_i \in Cluster\ k, x_j \notin Cluster\ k \tag{16}$$

The function d in (16) is the DTW distance between two points. The minimum distance between member of one cluster and all the members of other clusters is the separation of a cluster. An example separation is given in figure 5; the blue line represents the separation of the red cluster from the green one.

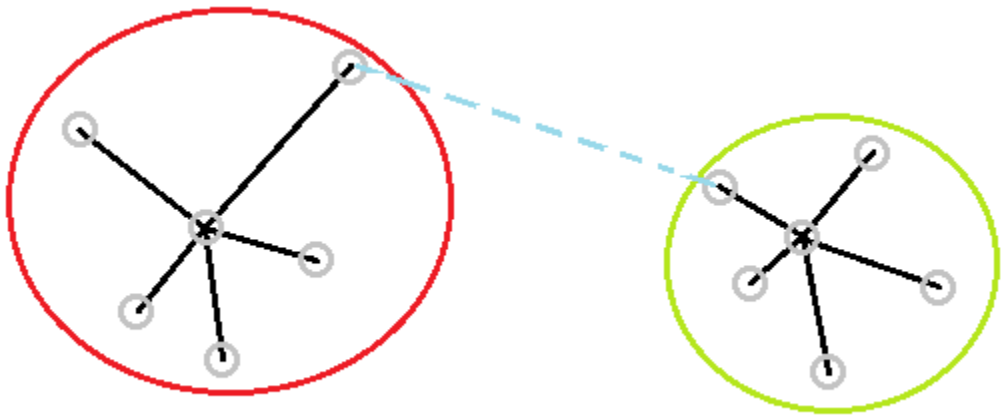


Figure 5: The separation visualization between the red and green clusters

The higher separation is, the more distinguished clusters are. As a result we want average in-cluster dissimilarity to be low and separation of clusters to be high.

4.5. Neural Networks

4.5.1 Motivation

Neural networks (NNs) are a prediction or classification method. NNs use an observation and works it through model to create a predicted result. Then the model parameters are adjusted by using the margin of error of the model for the observation. This process is done for each observation, where the NNs are taught the most accurate model achievable from the observations.

The NNs use nodes with different weights. There are layers of nodes and different number of nodes available in each layer. An observation is put in the network from the input nodes. Then they are multiplied by each node in each layer in succession until the output layer is reached.

There may be multiple inputs and also multiple outputs as well. Different inputs give a wider observation pool to learn from, and multiple outputs can be used in the classification methods to assign each observation to the best cluster.

The reason Neural Networks are implemented is two-fold. First of all, it is proven that non-linear models perform better than linear models for analogous series modeling [16]. Neural networks are non-linear method and should outperform base regression method. The second reason also describes the usage of neural networks; Chu and Zhang also stated that Neural Networks that use seasonally adjusted series outperform all of the tested linear and non-linear methods. In our case, the residuals of the first layer is actually seasonally adjusted series. So using Neural Networks instead of behaviorally pooled regression have the potential of increasing the accuracy of forecasts.

4.5.2 Options of Neural Networks

There are different options and properties of a neural network. First, the number of layers and nodes need to be given. Next, the number of networks crated should be given. NNs start with random seeds and multipliers, so each resulting network is different even with the same observations. As a result increasing the number of networks trained may yield more accurate networks. Also, a "threshold" can be defined as a stopping condition of the network. Threshold tells the network to stop iterating when a certain error upper limit is reached. Finally and most importantly, the inputs need to be defined.

A standard network is formulated A single hidden layer with 5 nodes are implemented. All of the connections between the nodes are present and can not be removed or added; so the tested model is not a self-organizing NN.

Assuming NNs would outperform regression methods, the behavioral clustering methods should be implemented as well to observe the most increase and create a level comparison. For each behaviorally similar store cluster and product category, an individual model is created. As the inputs, own residuals and behavioral cluster average residuals of the first layer are put in the NN.

There are two different parameters present in the R environment that should be adjusted for better accuracy. The first one is "repetition" which indicates how many networks are going to be built for each instance. Each instance is different due to the fact that the initial weights of the nodes are random. Ten repetitions for each instance is selected; the network with least error is kept. The second is the "threshold" level, which signifies a threshold point for ending the iteration of the network. When threshold level error is reached, the iteration stops. This value should not be too low as then the models would not converge; however providing it is too large, the models would be inaccurate. We proposed a threshold level of 0.01.

Aside from the NN options, for the testing options we have divided the data set into 3 categories; training, validation and test sets. Training is determined to be the first 3 years of data, validation is the 4^{th} year of the data and testing period is the last year of the data which is consistent with the regression calculations.

The pooling we determined in the regression part is applied here as well. We have 7 categories, 27 clusters and for each lead we create different networks. We have 189 networks for each 12 leads, 2268 in total. When the number of repetitions is taken into consideration, 22680 networks are created and most accurate (least error) models for the training set are kept for each category, cluster and lead. The inputs we have given the neural networks are past 1 year of store-category residuals and cluster residuals.

Chapter 5

APPLICATION OF PROPOSED IMPROVEMENTS AND RESULTS COMPARISON

In this section, we will apply the behavioral clustering method we have described in chapter 4 to a real world data set and evaluate the impact of our method on accuracy. We use the data sets previously described, in chapter 3, and then the results of our methods on the accuracy of the models.

To achieve the best properties and cluster size, we have tested different number of cluster sizes and we provide the graphs for average in-cluster dissimilarity and separation. These graphs are given in figures 7 and 8. Red lines signify the default business hierarchical clustering's dissimilarity and separation values.

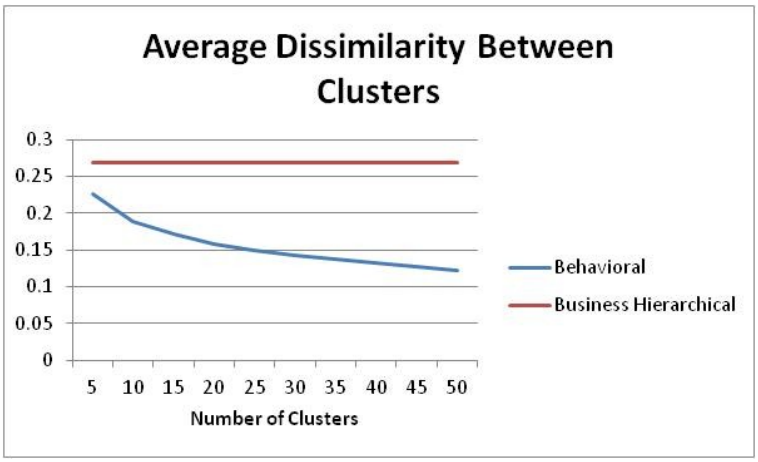


Figure 6: Average In-Cluster Dissimilarity comparison between different number of clusters tested and business hierarchical clustering

- It is seen from figure 6 that the dissimilarity is lower with behavioral clustering than the currently used business hierarchical clusters.
- Cluster dissimilarity decreases at a decaying rate as the number of clusters increases. This reaction is expected; a point of good in-cluster dissimilarity is achieved in between 25 and 30 clusters.

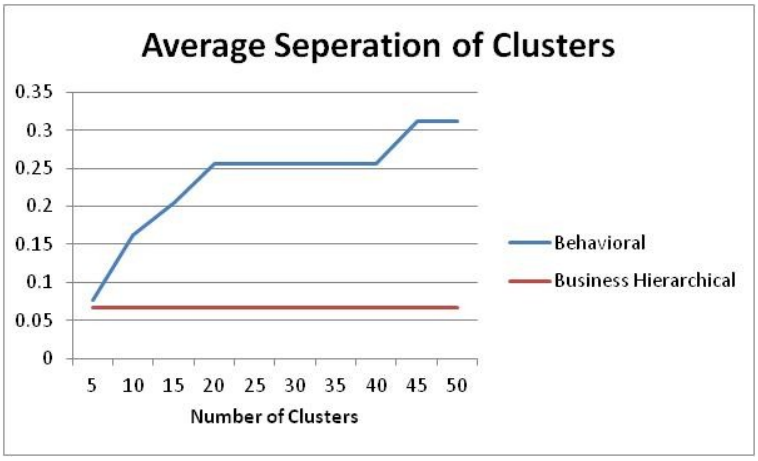


Figure 7: Average cluster separation comparison between different number of clusters tested

- From figure 7, it is seen that behavioral clustering always has higher cluster separation than the currenly used business hierarchical pooling.
- Cluster separation has breaking points between number of clusters. Between 20-40 clusters, the separation does not increase. After that point, there is a low marginal increase (from ~0.23 to ~0.31).

Our aim is finding the best balance between average dissimilarity and cluster separation. In the tested number of clusters, the best balance of average in-cluster dissimilarity and cluster separation is achieved in 27 clusters with DTW and test period not included. This cluster number is also consistent with the behavioral clustering, thus provides us with a direct comparison opportunity between hierarchical and behavioral clustering.

5.1. Results of Clustering through DTW and k-medoids

As we have the same number of clusters in both hierarchical and behavioral clustering, we have a comparison opportunity. In figures 9 and 10 we have given the stacked bar graphs of in-cluster store characteristics; the characteristics shown are the ones used in hierarchical clustering.

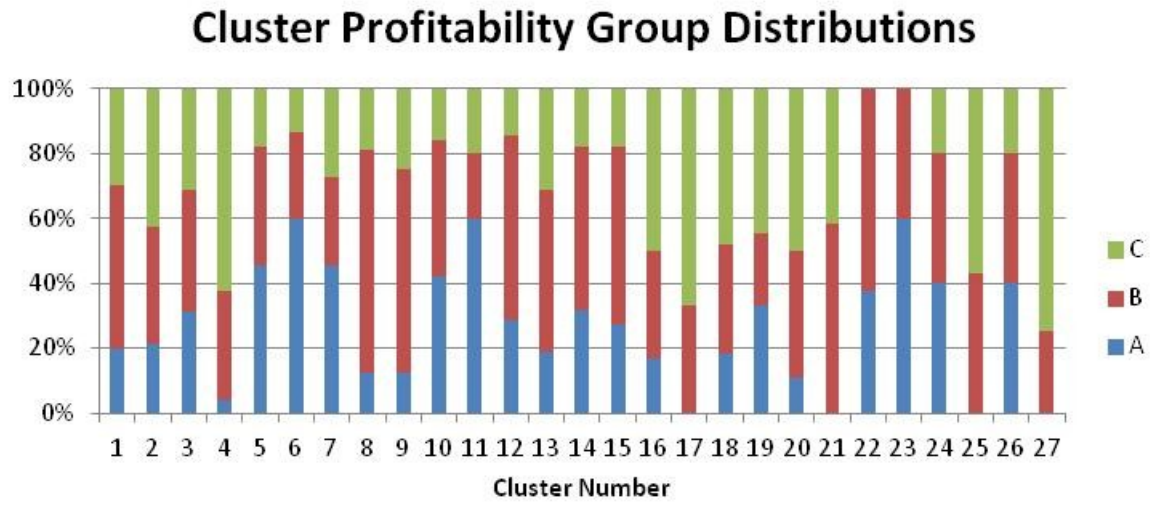


Figure 8: Distribution of profitability groups in behavioral clusters

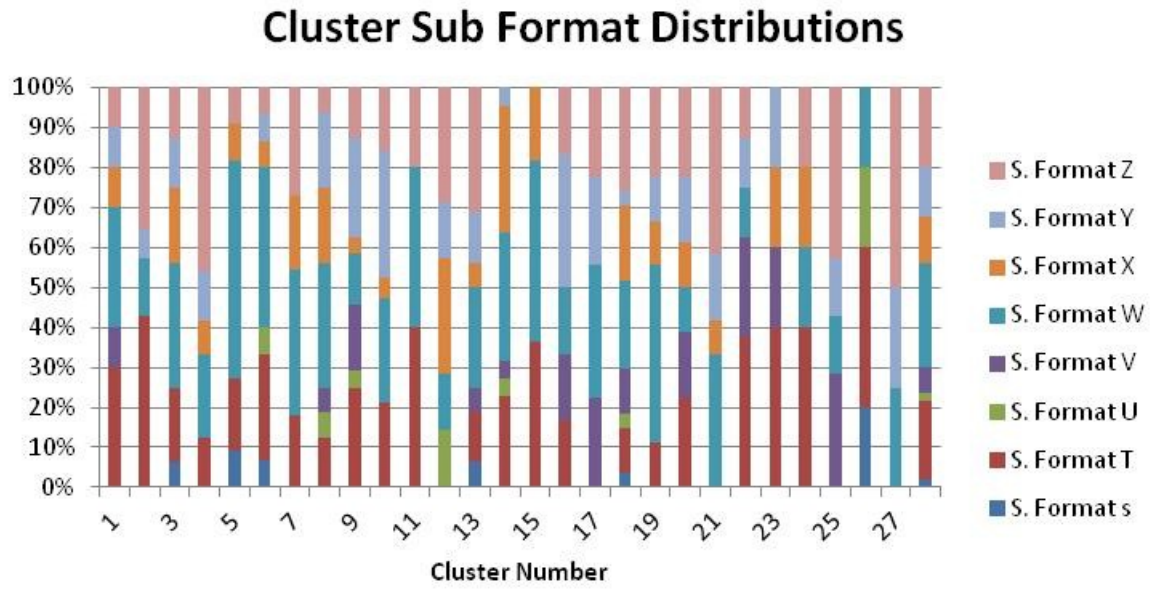


Figure 9: Distribution of sub formats in behavioral clusters.

From figures 8 and 9, we see that behavioral clustering results in highly different groupings than hierarchical clusters. In the hierarchical cluster, these graphs would show uni-characteristic clusters. However in the behavioral clustering this is not the case. These graphs validate our suspicions about stores acting differently from each other even when they possess the same business characteristics.

To visually prove that using these cluster options and methods actually result in useful behavioral clustering, we have provided some example series in different clusters. These comparisons of time series can be found in figure 10.

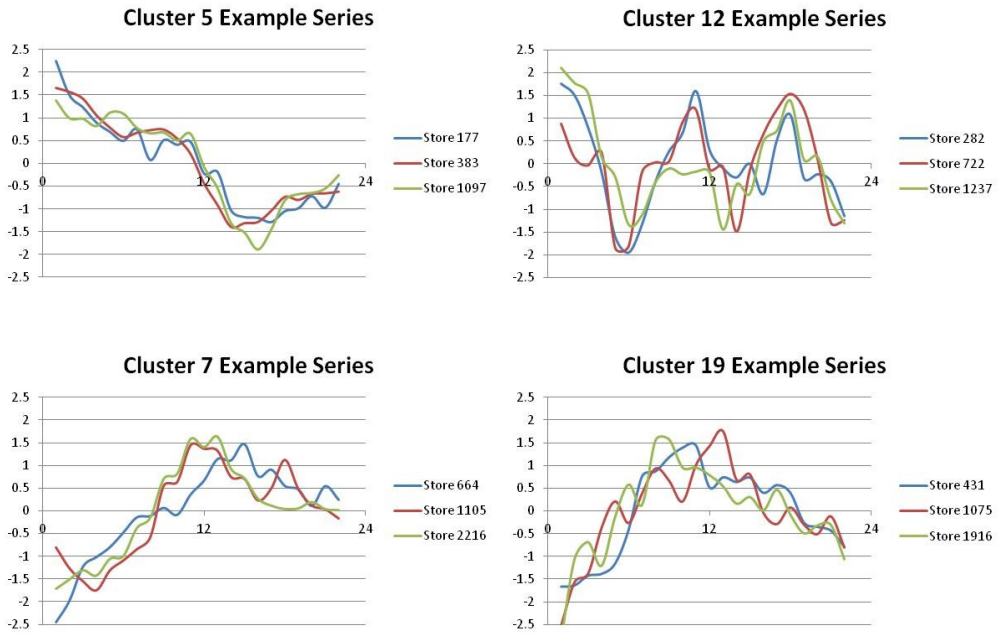


Figure 10: Comparison of example series in different clusters

From figure 10, we see that the clusters do behave differently from each other while the series belonging to the same cluster are similar. Different trend patterns (increasing or decreasing) are distinguished, as well as seasonal patterns (e.g. Cluster 12). One important thing to note is that different rates of increases are also

distinguished; as seen in the comparison between Cluster 7 and Cluster 19. This states that the time warping properties are actually handy in distinguishing series even though the overall pattern is similar. This is an important property and highly necessary for retail environment.

5.2. Evaluation of Forecasting Accuracy for Alternative Ways of Using Behavioral Clustering

After we have determined which cluster each store belongs, there are different ways to use these clusters with our original regression method to increase the accuracy. We can; a) change the pooling groups to the clusters determined, b) add the cluster trend-cyclical components as inputs to the regression, c) implement both. We have tested each option, the comparison of them is explained.

We have generated an experiment to decide how to implement behavioral clusters. The blueprint for the results table is given in Table 17. The values are the mean MAPE values of different models tested. The horizontal axis of the table shows the input variables used in the model. We tested using different combinations of own residuals, dimensional (format-product category, product category, store residuals) and the behavioral cluster residuals. The crosses (X) indicate which of these series are used as input.

The vertical axis indicates the segmentation, or pooling, done for that experiment. We either use the business hierarchical, characteristic oriented segments of the behavioral segments we have identified.

The Mean MAPE values at the final row and column indicate horizontal and vertical average of MAPE values, respectively. Finally the bottom part of the table shows the

benefits of adding the behavioral cluster when using only own residuals and when using dimensional residuals. An example graph is given in Table 17.

Table 17: Blueprint of residual presentation table.

		Input Series				
	Own residuals	X	X	X	X	
	Dimensional Average Residuals		X		X	
	Cluster Average Residuals			X	X	MAPE
Segmentation Groups	Property Segments	A	B	C	D	(A+B+C+D) / 4
	Behavioral Segmens	E	F	G	H	(E+F+G+H) / 4
	Mean MAPE	(A+E) / 2	(B+F) / 2	(C+G) / 2	(D+H) / 2	

	Adding Cluster residuals to Own	Ading Cluster residuals to Own and Dimensional
Improvement in MAPE	(A+E) / 2 - (C+G) / 2	(B+F) / 2 - (D+H) / 2

For example in Table 17, the cell C would be the MAPE of the model which uses own residuals and cluster average residuals as inputs and has business hierarchical pooling. It can be compared to cell G, which uses the same inputs but has behavioral pooling. The vertical MAPE values can be compared to understand whether property segments or behavioral segments have better accuracy. Finally the sub-table indicates how much improvement adding cluster residuals bear.

The input series at the top give us which input series are used in each experiment. The pooling segmentation stated on the left show whether business hierarchy or behavioral clustering is used to estimate a model.

Finally the separate part of the table show the marginal improvement when the cluster residuals are added and behavioral clusters are adopted, sequentially. We have tested

the options, and the comparison of different aggregation levels of accuracies are given in tables 18, 19 and 20.

To determine how to use the behavioral pooling, specific comparisons between cells are made. To decide on the benefit of behavioral segmentation over business hierarchical segmentation, t-values and p-values of difference between A-E, B-F, C-G and D-H are calculated. These values signify whether t-values for the comparison of methods are given in Table 21 and p-values are given in 22 for comparison of the methods. The significance testing will be done on as stated in chapter 3.

Table 18: Store level MAPE comparison between different cluster using options

		Input Series				
	Own residuals	X	X	X	X	
	Dimensional Average Residuals		X		X	
	Cluster Average Residuals			X	X	**Mean MAPE**
Segmentation Groups	**Property Segments**	11.95%	11.28%	11.55%	11.06%	**11.46%**
	Behavioral Segments	11.64%	11.12%	11.14%	10.81%	**11.18%**
	Mean MAPE	11.79%	11.20%	11.35%	10.93%	

	Adding Cluster Residuals to Own Residuals	**Adding Cluster Residuals to Own and Dimensional Residuals**
Improvement in MAPE	0.45%	0.27%

Table 19: Format level MAPE comparison between different cluster using options

		Input Series				
	Own residuals	X	X	X	X	
	Dimensional Average Residuals		X		X	
	Cluster Average Residuals			X	X	Mean MAPE
Segmentation Groups	Property Segments	5.39%	4.48%	4.69%	4.08%	**4.66%**
	Behavioral Segments	5.15%	4.67%	4.33%	3.91%	**4.51%**
	Mean MAPE	5.27%	4.57%	4.51%	4.00%	

	Adding Cluster Residuals to Own Residuals	Adding Cluster Residuals to Own and Dimensional Residuals
Improvement in MAPE	0.76%	0.58%

Table 20: National level MAPE comparison between different cluster using options

		Input Series				
	Own residuals	X	X	X	X	
	Dimensional Average Residuals		X		X	
	Cluster Average Residuals			X	X	Mean MAPE
Segmentation Groups	Property Segments	3.37%	2.81%	3.23%	2.69%	**3.02%**
	Behavioral Segmens	3.27%	2.81%	2.52%	2.12%	**2.68%**
	Mean MAPE	3.32%	2.81%	2.87%	2.41%	

	Adding Cluster residuals to Own	Ading Cluster residuals to Own and Dimensional
Improvement in MAPE	0.45%	0.40%

Table 21: t-values of the compared pairs and aggregation levels

Compared Pair	Store	Format	National
A-E	2.56172	0.3552	0.3552
B-F	14.7067	3.3305	3.3305
C-G	5.15436	0.98886	0.98886
D-H	3.02413	0.56069	0.56069

Store MAPE t-distribution value of 0.05 p-value (two-tailed): 1.96704937

Format MAPE t-distribution value of 0.05 p-value (two-tailed): 2.77644511

National MAPE t-distribution value of 0.05 p-value (two-tailed): 2.200985159

Table 22: p-values of the compared pairs and aggregation levels

Compared Pair	Store	Format	National
A-E	0.010452	0.724029	0.729157
B-F	9.86E-48	0.001694	0.006705
C-G	2.67E-07	0.327795	0.343985
D-H	0.002509	0.577672	0.586248

From the analysis of the MAPE values, t-values and p-values, some conclusions are made.

From tables 18, 19, and 20 we see that behavioral segmentation performs better than business hierarchical pooling.

In the store level, these improvements are statistically significant as seen from tables 21 and 22. However as the aggregation level increases, the significance decreases even though the improvement increases. Most probably, this decrease in significance is related to decrease in the sample size.

The most significant improvement in all of the levels is from implementing behavioral segmentation without using cluster residuals as input.

Now that behavioral segmentation is proved to result in significant improvements, the significance between inputs to the series is calculated. There are three sets of inputs that can be used; 1) Own residuals, 2) Dimensional residuals and 3) Cluster Residuals. For the comparison of the inputs, the cells E,F,G and H will be compared against one another for significance tests. t-values and p-values of these comparisons are given in Tables 23 and 24.

Table 23: t-values of the compared pairs and aggregation levels

Compared Pair	Store	Format	National
E-F	2.47608	0.86108	0.86108
F-G	1.57133	0.43734	0.43734
G-H	12.2113	3.73916	3.73916

Store MAPE t-distribution value of 0.05 p-value (two-tailed): 1.96704937

Format MAPE t-distribution value of 0.05 p-value (two-tailed): 2.77644511

National MAPE t-distribution value of 0.05 p-value (two-tailed): 2.200985159

Table 24: p-values of the compared pairs and aggregation levels

Compared Pair	Store	Format	National
E-F	0.013324	0.393566	0.407572
F-G	0.116184	0.663868	0.670327
G-H	1.06E-33	0.000501	0.003271

- From tables 23 and 24, it is seen that the improvements are significant for store level in all of the comparisons.
- As the aggregation level increases, the p-value of significance increases. The difference between G and H is the most significant in all of the aggregations. The input difference is adding dimensional residuals to own and cluster residuals.

- The decrease in aggregation is most probably due to the reduction in sample size.

Tables 18 through 24 show that adding the cluster residuals and pooling with the clusters should be both applied, as we have previously determined. One important point is when the pooling is behavioral, the improvement gained from adding cluster-level residuals is less as seen from the MAPE values. This means they have diminishing returns when both of the methods are applied; but they still perform better than business-hierarchical clustering. Also the significance of improvement decreases as the aggregation level increases. As a result of these observations, the final method applies behavioral pooling and add cluster residuals as trend-cyclical indicators due to:

- Behavioral segmentation provides better accuracy than business hierarchical segmentation
- Using cluster residuals amongst the dimensional residuals in regression increases accuracy
- Behavioral segmentation along with using cluster residuals as input provide diminishing returns, seen from the comparison part of tables 17-20. However the improvement is still considerable.

5.3. Breakdown of Accuracy Measures by Lead Times

The final comparison between the behaviorally pooled layer 2 regression with cluster residuals, our original method, and our benchmarks for store-category level forecasts are given in tables 25 and 26. These tables add the improved method's accuracy to the tables 3 and 4 given in chapter 3. Also the mean MAPE of the leads correspond to cells A B and H in tables 17-20.

- Layer 2 with own residuals = cell A
- Layer 2 with dimensional residuals = cell B
- Layer 2 with behavioral pooling = cell H

Table 25: Store-Category level forecasts MAPE values

Store-Category MAPE Values					
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**	**Layer 2 with Behavioral Pooling**
1	13.85%	17.09%	13.08%	12.49%	12.11%
2	14.66%	18.74%	14.99%	14.37%	13.57%
3	16.84%	19.66%	16.50%	15.83%	15.22%
4	22.26%	22.91%	20.17%	18.51%	18.10%
5	19.64%	20.82%	18.23%	17.17%	16.78%
6	20.42%	19.14%	16.90%	16.11%	16.11%
7	21.73%	18.55%	16.92%	16.28%	16.53%
8	24.18%	21.18%	19.50%	18.69%	18.80%
9	26.50%	22.32%	20.47%	20.17%	19.52%
10	29.19%	23.92%	22.45%	21.51%	21.18%
11	27.41%	22.68%	20.87%	20.24%	19.74%
12	27.99%	24.24%	23.24%	22.32%	22.44%
MEAN	22.05%	20.94%	18.61%	17.81%	17.51%

Table 26: Store-Category level forecasts MPE values

Store-Category MPE Values					
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**	**Layer 2 with Behavioral Pooling**
1	1.14%	-5.19%	-3.65%	-4.37%	-2.75%
2	-3.21%	-7.41%	-5.84%	-5.50%	-3.47%
3	0.04%	0.20%	0.81%	0.71%	0.58%
4	-4.11%	-7.27%	-5.98%	-4.32%	-4.46%
5	-0.89%	-3.83%	-2.66%	-2.85%	-2.34%
6	0.54%	-3.35%	-2.26%	-3.25%	-2.52%
7	4.06%	-1.62%	-1.20%	-2.80%	-2.35%
8	5.71%	-6.04%	-5.32%	-5.59%	-5.04%
9	-7.75%	-7.14%	-6.06%	-7.44%	-5.53%
10	0.06%	-5.40%	-4.71%	-6.31%	-5.78%
11	-0.15%	-3.20%	-2.03%	-3.38%	-2.91%
12	14.28%	12.62%	12.13%	8.65%	5.67%
MEAN	0.81%	-3.14%	-2.23%	-3.04%	-2.57%

From Tables 25 and 26, it is seen that;

- At store-category level behavioral pooling provides an improvement, however a small one.
- At some leads behavioral pooling is better (e.g. lead 3,4) while in others hierarchical pooling works better (e.g. lead 7).
- The bias of the forecasts decrease when behavioral pooling is implemented.
- Behavioral pooling is consistently better than business hierarchical pooling; in 9 out of 12 leads, behavioral pooling performs better than business hierarchical pooling.

The main improvement is seen in the aggregated dimensional forecasts, given in tables 27-32.

Table 27: Store level forecasts MAPE values

Store MAPE Values					
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**	**Layer 2 with Behavioral Pooling**
1	8.44%	9.88%	7.56%	7.23%	6.95%
2	8.37%	10.78%	8.44%	8.09%	7.15%
3	10.79%	12.78%	10.69%	9.93%	9.17%
4	14.58%	14.71%	12.95%	12.08%	11.14%
5	12.95%	13.00%	11.46%	10.85%	10.41%
6	12.52%	10.96%	9.68%	9.29%	9.11%
7	13.95%	11.81%	10.79%	10.34%	10.28%
8	16.34%	12.81%	12.03%	11.46%	11.74%
9	17.12%	14.64%	13.49%	12.76%	12.58%
10	19.34%	15.27%	14.55%	13.85%	13.06%
11	17.15%	14.77%	13.94%	13.43%	13.06%
12	19.49%	18.57%	17.83%	16.07%	15.03%
MEAN	14.25%	13.33%	11.95%	11.28%	10.81%

Table 28: Store level forecasts MPE values

Store MPE Values					
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**	**Layer 2 with Behavioral Pooling**
1	3.49%	-3.14%	-2.33%	-2.85%	-1.17%
2	-0.87%	-5.16%	-4.37%	-4.48%	-2.15%
3	2.88%	3.08%	3.09%	2.46%	2.21%
4	-0.86%	-4.46%	-3.66%	-2.13%	-2.51%
5	2.19%	-1.15%	-0.64%	-1.15%	-0.19%
6	2.08%	-2.07%	-1.54%	-1.91%	-1.33%
7	5.31%	-0.25%	-0.14%	-1.28%	-0.60%
8	7.53%	-4.46%	-3.80%	-2.87%	-2.65%
9	-5.85%	-3.84%	-2.98%	-3.36%	-2.03%
10	2.74%	-1.44%	-0.92%	-2.33%	-1.94%
11	0.46%	-1.08%	-0.29%	-1.03%	-0.89%
12	13.69%	14.01%	13.70%	11.39%	8.39%
MEAN	2.73%	-0.83%	-0.32%	-0.79%	-0.41%

Table 29: Format level forecasts MAPE values

Format MAPE Values					
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**	**Layer 2 with Behavioral Pooling**
1	14.21%	3.23%	2.77%	2.72%	1.89%
2	13.39%	4.81%	4.76%	5.14%	2.45%
3	15.21%	4.44%	4.36%	4.18%	2.32%
4	15.44%	3.12%	3.44%	2.37%	1.39%
5	17.84%	2.72%	2.82%	2.13%	1.36%
6	15.49%	5.36%	4.92%	3.66%	4.03%
7	20.97%	2.52%	2.39%	1.87%	2.44%
8	22.78%	5.34%	4.86%	3.78%	4.45%
9	10.71%	5.15%	4.68%	3.21%	3.31%
10	17.24%	5.45%	5.51%	4.21%	4.86%
11	15.20%	9.71%	8.71%	7.74%	8.63%
12	21.52%	15.97%	15.44%	12.71%	9.83%
MEAN	16.67%	5.65%	5.39%	4.48%	3.91%

Table 30: Format level forecasts MPE values

Format MPE Values					
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**	**Layer 2 with Behavioral Pooling**
1	14.21%	-2.64%	-2.68%	-1.98%	-0.96%
2	10.09%	-4.81%	-4.76%	-5.14%	-2.45%
3	13.53%	3.80%	3.16%	2.89%	2.32%
4	15.39%	0.31%	0.15%	0.76%	1.24%
5	15.42%	0.33%	0.31%	-0.05%	0.61%
6	13.88%	-1.84%	-1.63%	-1.87%	-1.49%
7	16.77%	-0.40%	-0.51%	-1.25%	-1.28%
8	15.67%	-4.79%	-4.32%	-2.80%	-3.15%
9	5.77%	-2.78%	-2.34%	-1.64%	-1.50%
10	16.68%	5.45%	5.51%	4.21%	4.86%
11	5.00%	-5.72%	-4.50%	-4.48%	-5.26%
12	21.47%	15.97%	15.44%	12.71%	9.83%
MEAN	13.66%	0.24%	0.32%	0.11%	0.23%

Table 31: National level forecasts MAPE values

National MAPE Values					
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**	**Layer 2 with Behavioral Pooling**
1	17.77%	1.32%	1.25%	1.23%	0.30%
2	14.27%	3.28%	3.06%	3.20%	1.26%
3	17.49%	5.89%	5.21%	4.34%	3.87%
4	17.67%	1.33%	1.69%	2.21%	1.71%
5	17.46%	1.56%	1.67%	1.11%	1.54%
6	14.69%	1.53%	1.25%	1.51%	1.48%
7	18.23%	0.99%	0.90%	0.04%	0.14%
8	18.88%	3.40%	2.96%	1.78%	1.97%
9	6.83%	2.83%	2.23%	2.07%	1.64%
10	17.68%	4.57%	4.62%	3.53%	3.19%
11	12.09%	0.97%	1.78%	1.11%	0.40%
12	20.77%	14.48%	13.86%	11.58%	7.99%
MEAN	16.15%	3.51%	3.37%	2.81%	2.12%

Table 32: National level forecasts MPE values

National MPE Values					
Lead	**Winter's Exponential Smoothing**	**Layer 1**	**Layer 2 with Own Residuals**	**Layer 2 with Dimensional Residuals**	**Layer 2 with Behavioral Pooling**
1	17.77%	-1.32%	-1.25%	-1.23%	-0.30%
2	14.27%	-3.28%	-3.06%	-3.20%	-1.26%
3	17.49%	5.89%	5.21%	4.34%	3.87%
4	17.67%	1.33%	1.69%	2.21%	1.71%
5	17.46%	1.56%	1.67%	1.11%	1.54%
6	14.69%	-1.53%	-1.25%	-1.51%	-1.48%
7	18.23%	0.99%	0.90%	-0.04%	0.14%
8	18.88%	-3.40%	-2.96%	-1.78%	-1.97%
9	6.83%	-2.83%	-2.23%	-2.07%	-1.64%
10	17.68%	4.57%	4.62%	3.53%	3.19%
11	12.09%	0.97%	1.78%	1.11%	0.40%
12	20.77%	14.48%	13.86%	11.58%	7.99%
MEAN	16.15%	1.45%	1.58%	1.17%	1.02%

Some results are observed from Tables 27-32.

- As the aggregation level increases, behavioral pooling provides increasingly more accurate forecasts.
- Possibly the most important observation is that the improvement is the most in lead 12, the difficult to predict time period. This shows behavioral pooling is able to better take into account the movements and model the expected change in the reaction to such an event, than hierarchical clustering.
- The bias of the forecasts decrease in all of the aggregation levels with behavioral pooling; apparent from the MPE values.

By looking at this finding and the increase in the accuracy, we determined behavioral pooling and behavioral cluster residuals are necessary and implemented to the second layer model.

5.4. Neural Network Results

We trained the networks with 10 repetitions over the training data set for each cluster, product category and lead time. For leads 3, 7 and 9 the networks could not converge to a value.

An example network is given in figure 11. Note that the coefficients are not visible in graphical representation due to the number of variables, however they are reachable within the environment.

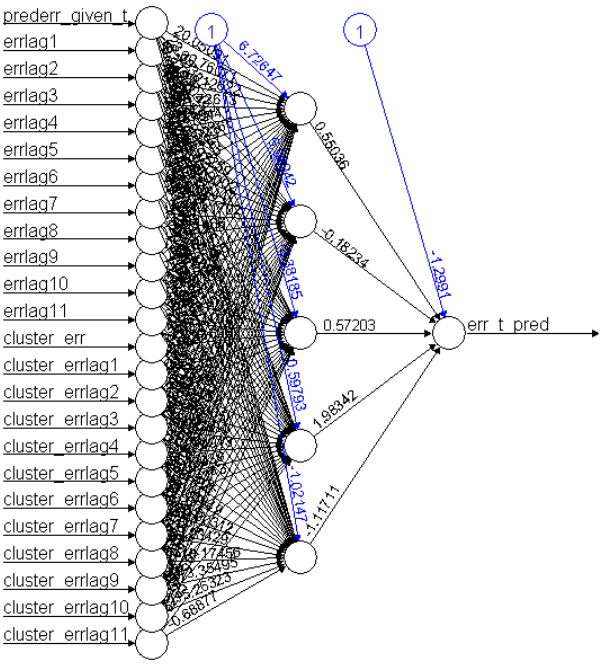


Figure 11: An example Neural Network for lead 10, cluster 15 and lead 3.

We have calculated the mean and median APE's and mean PE's of the store-category level sales' predictions as accuracy measures. The aggregated residuals are not calculated. We compared the store-category prediction accuracy to our accuracies from the behavioral pooled regression. The results are given in Table 33.

Table 33. Mean APE, median APE and mean PE of the predictions for behavioral pooling and neural networks.

Store-Category MAPE Values		
Lead	Layer 2 with Dimensional Residuals	Neural Network
1	13.04%	17.09%
2	14.78%	18.74%
3	16.33%	
4	19.89%	22.90%
5	18.05%	20.82%
6	16.80%	19.14%
7	16.78%	
8	19.31%	21.18%
9	20.26%	
10	22.09%	23.92%
11	20.77%	22.68%
12	23.10%	24.32%
MEAN *	18.65%	21.20%

Store-Category Median MAPE Values		
Lead	Layer 2 with Dimensional Residuals	Neural Network
1	9.53%	12.36%
2	10.36%	12.88%
3	11.26%	
4	11.90%	13.82%
5	12.03%	14.15%
6	11.92%	13.76%
7	11.72%	
8	12.21%	13.70%
9	13.39%	
10	12.69%	14.06%
11	13.54%	15.15%
12	17.27%	18.55%
MEAN *	12.38%	14.27%

Store-Category MPE Values		
Lead	Layer 2 with Dimensional Residuals	Neural Network
1	-3.50%	-5.19%
2	-5.44%	-7.41%
3	1.18%	
4	-5.64%	-7.27%
5	-2.41%	-3.83%
6	-2.11%	-3.35%
7	-1.05%	
8	-5.08%	-6.04%
9	-5.91%	
10	-4.60%	-5.40%
11	-2.30%	-3.20%
12	11.74%	12.97%
MEAN *	-2.15%	-3.19%

We have deduced some observations from Table 33.

- Neural Networks could not converge to a model in leads 3, 7 and 9. The means are calculated from the remaining leads for both methods for comparability.
- From the accuracy measures, we see that standard neural networks perform worse than behaviorally pooled regression, however the margin is not great.
- Smarter NN systems that have optimized parameters and number of nodes may provide better forecasts. The optimal NN setup and the resulting accuracy of forecasts are possible future research venues.

Chapter 6

GAINING INSIGHTS WITH DYNAMIC TIME WARPING

There are trends and patterns present in the residual of store level sales. The second layer tries to model these patterns using past residuals. However these patterns may indicate other similarities and hidden drivers of a specific stores' sales. Potential drivers of sales, such as consumer satisfaction indices, are identified as candidate series, but they are too numerous to manually compare with each residual series. An automated method for discovering similarities is necessary.

We have explained our method of using Dynamic Time Warping to find similarities between stores in chapter 4. The robustness of Dynamic Time Warping can also be used to discover similarities with other socio-economic indicators and gain insights about potential drivers of sales in specific stores, formats or categories as well. These insights are invaluable for decision makers as indicators can be identified thus providing more preparation opportunities for proper reaction to incidents. For comparison, other similarity measures are described first to use as benchmarks.

6.1. Benchmark Methods for Similarity Identification

6.1.1 Correlation

Correlation is a method to identify dependence between two time series or two data sets. By calculating a correlation coefficient through various methods, the level of dependence between data sets is quantified. We use the Pearson correlation coefficient which takes a value between -1 and 1. -1 signifies the series are inversely proportional and 1 signifies the series being proportional. Unrelated series have a correlation coefficient of 0. The correlation coefficient used for our measures is the Pearson Correlation Coefficient, which is calculated with the formula given in (17).

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{17}$$

X and Y represent the residual and candidate series.

We want to use the correlation coefficients between two series as a similarity measure between the said series. If the correlation is significant and different from 0, this method proposes a relation between two data sets, which can be interpreted as similarity for our time series.

6.1.2 Cross Correlation

Cross Correlation is a correlation between lagged series. The main advantage of cross correlation is it's ability to determine similarities that may not have been apparent in correlation due to a shift in time. Also finding lagged dependencies may prove useful in determining leading series in clusters. However this method, although more flexible than correlation, is still less flexible than DTW as it is unable to warp the series; but only shifts them.

The similarity measure deduced from cross correlation is the same as correlation with only the addition of at which lag the highest correlation is found is also kept.

6.1.3 Derivative DTW

Derivative Dynamic Time Warping is first proposed by Keogh and Pazzani [37]. This method manipulates the series and derives a new series from the original time series. The new series are concerned more with the movement patterns of the series. The magnitudal differences and the size of the oscillations are removed and the obvious, trend-cyclical components remain.

The manipulation of the series involve using the last and the following time periods' level to calculate the trend for the selected three points in time. The formula is given in (18).

$$D_x[q] = \frac{(q_i - q_{i-1}) + (q_{i+1} - q_{i-1})/2}{2} \quad (18)$$

The series are manipulated according to formula (18) and the newly derived series are put into the DTW formulation. The rest of the procedure is the same as DTW procedure with the same options.

6.2. Similarity Benchmark Determination Method for DTW

The DTW distances are hard to interpret as a similarity measure on their own. There are multiple series that are compared, and each query-reference combination result in different minimum distances. Switching the reference and query series may result in different DTW distances as query is warped to fit the reference series; the dissimilarity matrix may be asymmetrical. Finally, the length of the series affects the

DTW distances. Therefore, we need a way of identifying when query series is statistically more similar to a reference series than a stationary series.

To counteract this problem, a Monte Carlo simulation experiment is proposed. Each series need an individual benchmark as each exhibit different behavior. Stationary series have no trend-cyclical components; we want to use the DTW distance between stationary series and store residual series as the benchmark. However a single DTW distance with a single stationary series is unreliable. Therefore we calculate the distances with 100 stationary series and use the 15^{th} percentile of this distribution. as the benchmark for similarity.

6.3. Proposed Method for Identifying Potential Drivers of Sales with DTW and DDTW

The first step of the proposed similarity identification method is identifying the reference and queries of the DTW method. The query series are the format, category, region and profitability group level residuals; in this case a total of 20 series are used. The reference series are socio-economic indicators. They are taken from Istanbul Stock Market Exchange, Turkey Central Bank statistics, Inflation levels. There are a total of 55 series as the reference.

Both the reference and query series are smoothed to eliminate the noise present that may interfere with the trend-cyclical similarities we want to find. Centered moving average with time periods are implemented. Afterwards the series are normalized to mean 0 and standard deviation 1. Magnitudal differences may interfere with the DTW distances and normalization improves the representativeness of DTW distances [41].

One hundred stationary series are created from normal distribution with mean 0 and standard deviation 1, the same as the normalization magnitude. Afterwards, DTW distance between each of these residual series and these 100 stationary series are

calculated. Finally the 15th percentile of these distances for each residual series are calculate and used for similarity benchmark. The meaning of this benchmark is; if the DTW distance between a query and reference is less than the benchmark for the reference, the query is more similar to the reference series than a stationary series.

Finally the DTW distance between each query and reference is calculated. If this distance is less than the calculated similarity benchmark for the residual series, than they are more similar then a stationary series

6.4. Comparison of the Methods Proposed for Similarity

We now want to investigate whether DTW or Cross correlation is better in identifying actual similarities and whether the proposed series are actually similar for each method.

We have created a list of socio-economic indicators that may potentially be drivers of residual sales. There are 55 such series. For comparison and indicators, 20 dimensional residual series are created from aggregation of the first layer residuals. Afterwards, there following steps are used to compare the identified the similarities.

- Similarities between residual series and socio-economic indicators are determined for correlation, cross-correlation and DTW.
- The proposed similarities are graded from 1 to 5 through visual inspection; 1 being not similar and 5 being highly similar.
- A merged list of all the proposed similarities by are created along with their grades.
- Four measures for the accuracy of the similarity measure are calculated.
    - The percentage of series that have a score of 4 or 5 amongst all the proposed series by that method.

- Within the pool of all the proposed series, percentage of how many of the 4 or 5 scored ones are proposed by each method.
- The average score of each method's proposed similarities
- Average number of series identified.

The summary statistics of the methods are given in Table 34.

Table 34: Comparison of the determined accuracy indicators between different similarity identification methods

	% of Series Proposed from All That Are Found Similar	% of Identified with Score 4/5	Average Score	Number Identified
DTW	86%	66%	3.56	1.05
DDTW	60%	45%	3.80	0.25
Correlation	40%	49%	2.44	5.1
Cross Correlation	23%	13%	1.97	3.35

Some observations are made from Table 34.

- DTW outperforms correlation and cross correlation in identifying true similarities with %86 accuracy
- The average score of DTW is large, meaning there is a small number of false positives.
- Correlation proposes a lot of similarities, however the quality of these proposels is not high, seen from the average score.
- Derivative DTW method proposes a really small number of series, and thus it has limited potential in identifying all of the similarities; it is prone to missing some of the similar series.

Proposed similarity measure through DTW is better than correlation, cross correlation, and Derivative DTW. DTW is used with the explained methodology for similarity identification.

6.5. Insights Gained from Similarity Analysis Using DTW

The least intuitive but the most interesting similarities are present for the Format X. This format aims for the upper class citizens with a good income. The similarities proposed are the Istanbul Stock Exchange price rate, Oil Prices, Investment Rate, Propensity to buy automobiles, and their related indicators. These examples show that the segmentation for this format is done expertly, and the customers in this format react accordingly to their profitability group.

Other similarities that are worth mentioning are the investment amounts for different segments in the country and different branches. As a background information, Turkey is divided into 6 different groups with different incentives for possible investors. A detailed list is available, but the important significance is that if a similarity is captured between investments in a investment segment and a branch is found, that branch has a lot of cities belonging to that investment segment. For example, Branch R includes cities belonging to investment segment 3, or Branch S includes cities belonging to segments 3 and 4. These similarities mean that when investments to a specific region increases, the stores benefit from them as well as the region itself. This insight will be invaluable in assessing future consumer behavior and profit fluxuation.

To visualize the similarities that are proposed, example graphs of three important similarities are given. For each example similarity, the matching graph of DTW, the alignment plot and visual overlay of the smoothed, normalized series are given. These graphs are seen in figures 12,13 and 14.

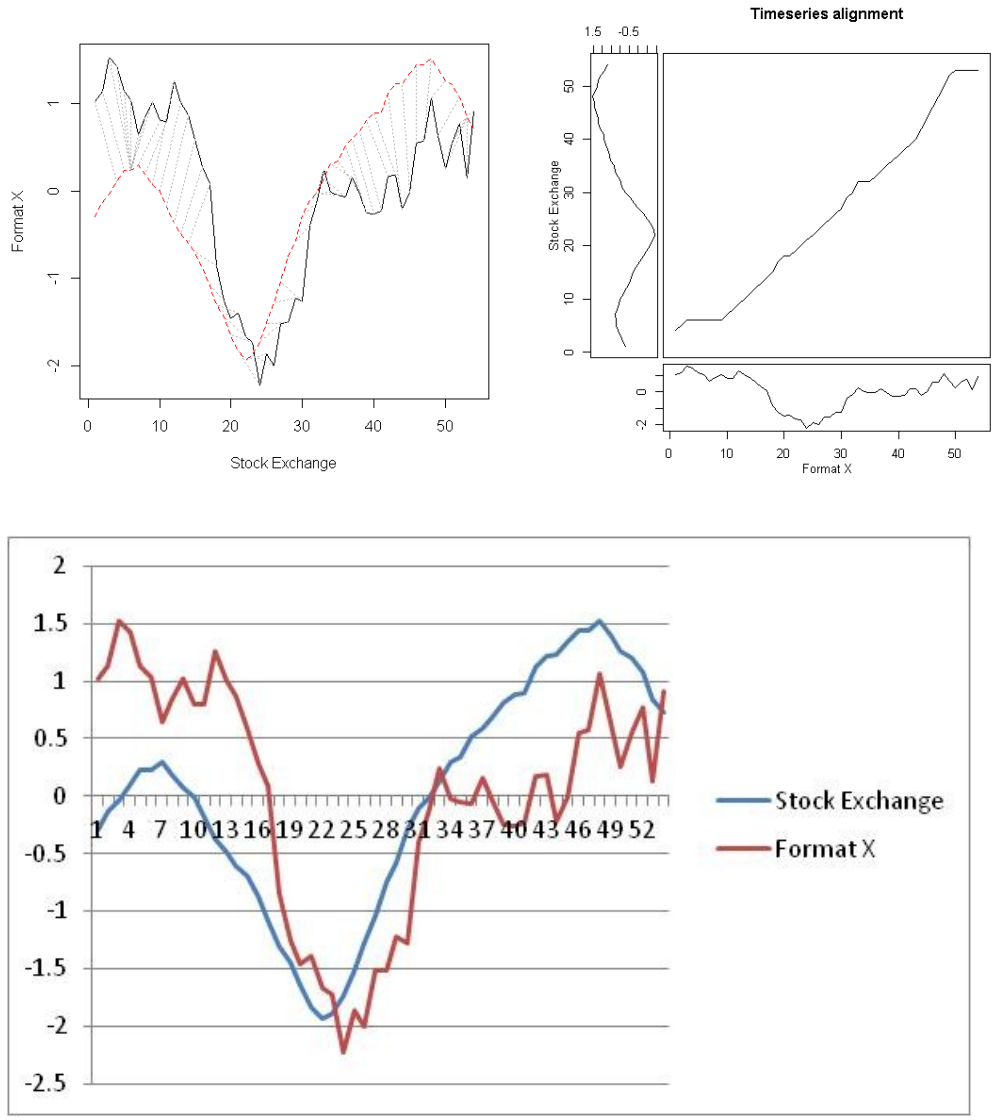
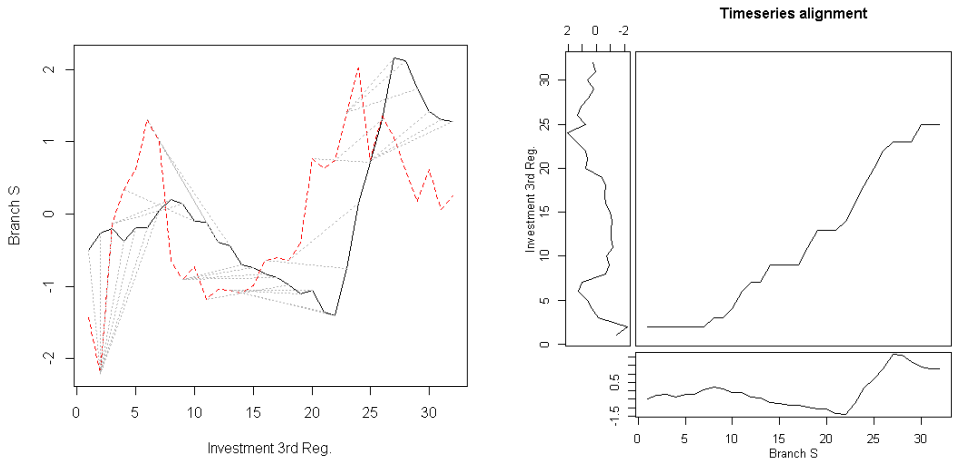


Figure 12: Comparison of the found similarity between Stock Exchange and Format X



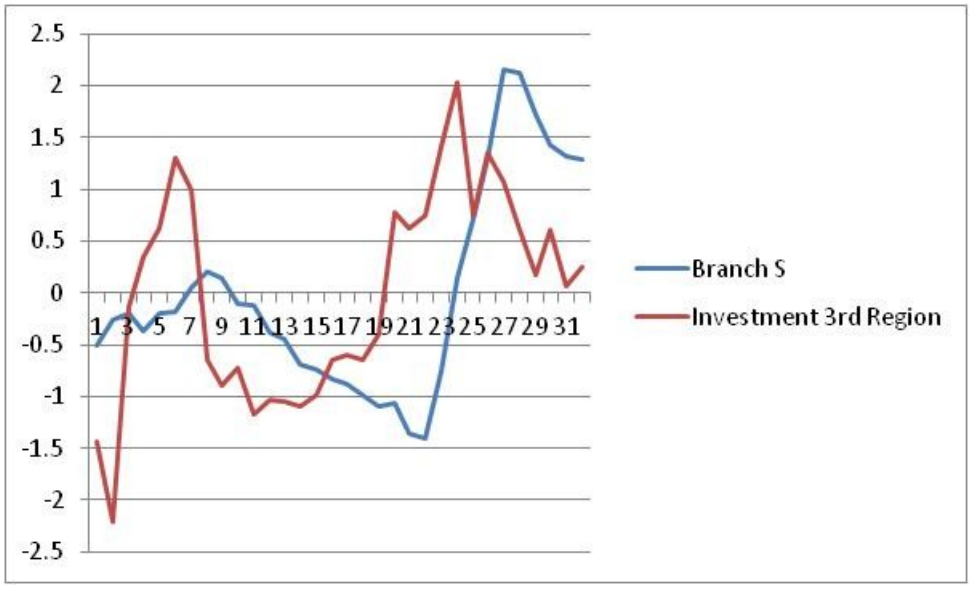
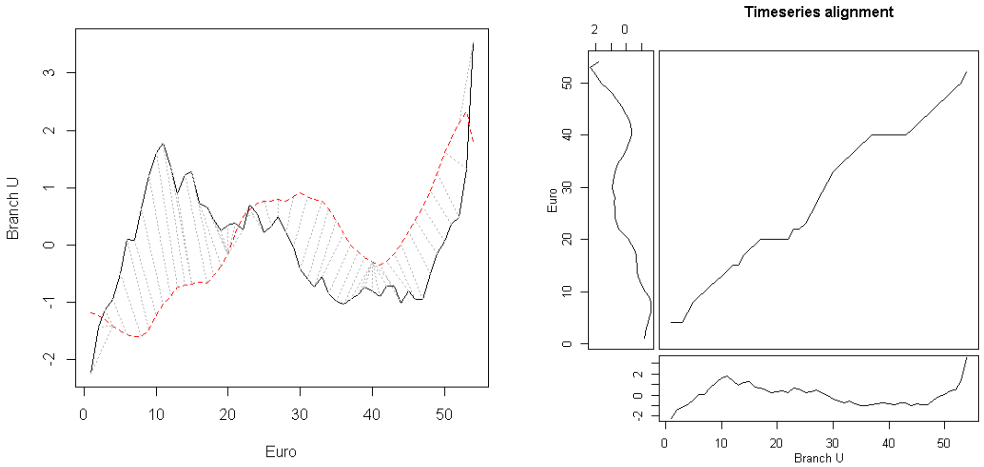


Figure 13: Comparison of the found similarity between Investment in 3rd Region and Branch S



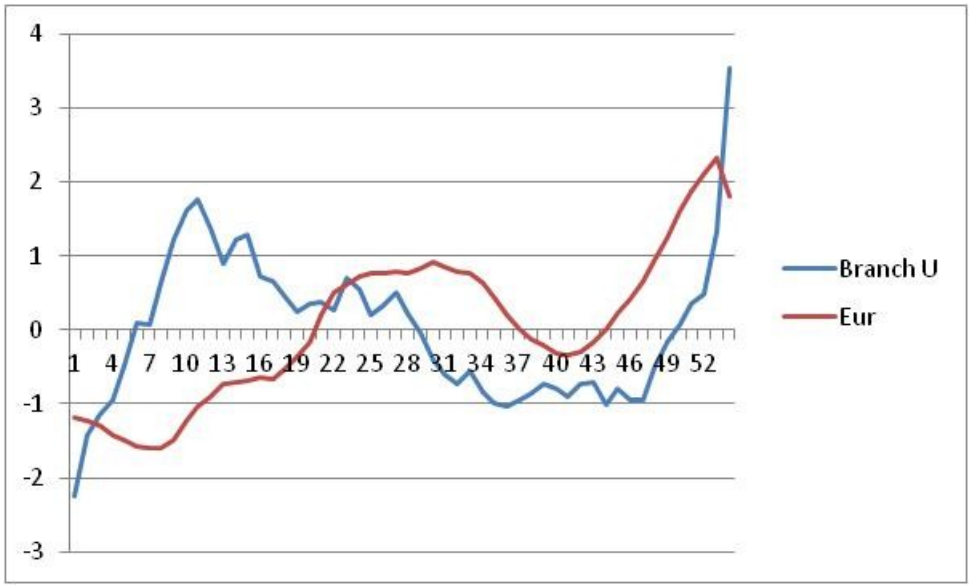


Figure 14: Comparison of the found similarity between Euro and Branch U

Chapter 7

CONCLUSION

In this thesis, we have proposed using behavioral pooling to improve the base regression method proposed for forecasting of retail sales by Gür Ali and Pınar [33]. We used Dynamic Time Warping for similarity measure in behavioral clustering of time series.

We have also developed an automated similarity identification method for insight gaining into retail sales. We developed a Monte Carlo experiment with Dynamic Time Warping distances to determine benchmark and tested against 55 predetermined socio-economic tindicator time series to develop insights.

In chapter 3, we described in detail the data used and the base regression method proposed by Gür Ali and Pınar [33]. We have shown that

- As the aggregation level increases; base method provides much better forecasts than Winter's Exponential smoothing. This result is despite the fact that the proposed models aggregate the lower level forecasts while Winter's Exponential smoothing is used to forecast aggregated series within themselves.
- There is a high spike of MAPE values of the proposed model in lead 12. It is presumed that this spike is related to an effect or incident related to the specific time period forecasted; i.e. t=60, December 2011; that is difficult to

find let alone incorporate in the model. This spike is more apparent as the aggregation level increases.

- As the aggregation level increases, the bias of the proposed method decrease while bias of exponential smoothing increases. This is most probably due to the fact that exponential smoothing forecasts the aggregated series while the proposed method aggregates the individual forecasts to higher levels
- Accuracy improvements stay significant as the aggregation level increases, however the p-values increase as the aggregation level increases. This is possibly due to the sample size decreasing with increase in aggregation level

In chapter 4 we described in detail the methods used for behavioral pooling, how it may be implemented into the models, Dynamic Time Warping methodology, and Neural Networks. In chapter 5 we have provide the analysis and the results of the improvements and methods suggested. We have concluded;

- Behavioral clustering always have higher cluster separation and lower in-cluster dissimilarity than the currenly used business hierarchical pooling fort he time series at hand.
- Behavioral segmentation improves the accuracy of the forecasts, and the accuracy improvement is statistically significant at the store level predictions.
- Using cluster residuals as input improves the forecasting accuracy and this increase is statistically significant in the store-level.
- As aggregation level increases, the improvements in accuracy persists however the statistical significance decrease. The most probable reason fort his decrease is the reduction in sample size as aggregation level increases.
- Behavioral segmentation along with using cluster residuals as input provide diminishing returns, seen from the comparison part of tables 17-20. However the improvement is still considerable.

We then break down the overall accuracy into different lead times and we made some observations from the lead-time specific MAPE and MPE values

- At store-category level behavioral pooling provides an improvement, however a small one.
- The bias of the forecasts decrease when behavioral pooling is implemented.
- Behavioral pooling is consistently better than business hierarchical pooling; in 9 out of 12 leads, behavioral pooling performs better than business hierarchical pooling.
- As the aggregation level increases, behavioral pooling provides increasingly more accurate forecasts.
- Possibly the most important observation is that the improvement is the most in lead 12, the difficult to predict time period. This shows behavioral pooling is able to better take into account the movements and model the expected change in the reaction to such an event, than hierarchical clustering.
- The bias of the forecasts decrease in all of the aggregation levels with behavioral pooling; apparent from the MPE values.

Finally for the forecasting method, we have provided the results of Neural Networks in the accuracy of the forecasts. From these results we deduce;

- Standard neural networks perform worse than behaviorally pooled regression, however the margin is not great.
- Smarter NN systems that have optimized parameters and number of nodes may provide better forecasts. The optimal NN setup and the resulting accuracy of forecasts are possible future research venues.

In chapter 6, we provide a method for similarty determination, applicable to insight generation for retail sales. We provide alternative methods and give a cmparison of the methods to determine which is more beneficial in similarity determination. We conclude that;

- DTW outperforms correlation and cross correlation in identifying true similarities with %86 accuracy
- With DTW, there is a small number of false positives.
- Correlation proposes a lot of similarities, however the quality of these proposels is not high, seen from the average score.
- Derivative DTW method proposes a really small number of series, and thus it has limited potential in identifying all of the similarities; it is prone to missing some of the similar series.

From this analysis we decided to use DTW as a similarity measure, and we applied the method to gain insights about the retail sales.

As a summary, we have provided an improvement with behavioral pooling through DTW, and have proved it performs better than business hierarchical pooling and Neural Networks. We have developed a method for insight generation which proves to be accurate and intuitive in similarty determination and insight gaining.

BIBLIOGRAPHY

[1] Allen, P. G., & Fildes, R. (2001). Econometric forecasting In J. S. Armstrong (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners (pp. 303-362). Boston: Kluwer

[2] Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods. Journal of Retailing and Consumer Services, 8(3), 147-156.

[3] Armstrong, J. Scott. "Findings from evidence-based forecasting: Methods for reducing forecast error." International Journal of Forecasting 22.3 (2006): 583-598.

[4] Armstrong, J. Scott. "Forecasting by extrapolation: Conclusions from 25 years of research." Interfaces 14.6 (1984): 52-66.

[5] Armstrong, J. Scott, and Fred Collopy. "Error measures for generalizing about forecasting methods: Empirical comparisons." International Journal of Forecasting 8.1 (1992): 69-80.

[6] Assent, Ira, et al. "The TS-tree: efficient time series search and retrieval."*Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. ACM, 2008.

[7] Bagnall, Anthony, et al. "Transformation Based Ensembles for Time Series Classification." *SDM*. 2012.

[8] Berndt, D., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. Paper presented at the KDD workshop.

[9] Bijmolt, T. H. A., Van Heerde, H. J., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. Journal of Marketing Research, 42(2), 141-156. doi: DOI 10.1509/jmkr.42.2.141.62296

[10] Bunn, D. W., & Vassilopoulos, A. I. (1999). Comparison of seasonal estimation methods in multi-item short-term forecasting. International Journal of Forecasting, 15(4), 431-443. doi: http://dx.doi.org/10.1016/S0169-2070(99)00005-9

[11] Chang, Pei-Chann, et al. "Evolving Neural Network with Dynamic Time Warping and Piecewise Linear Representation System for Stock Trading Decision Making." *Computer Science and Information Engineering, 2009 WRI World Congress on*. Vol. 5. IEEE, 2009.

[12] Chen, H., & Boylan, J. E. (2008). Empirical evidence on individual, group and shrinkage seasonal indices. International Journal of Forecasting, 24(3), 525-534. doi: http://dx.doi.org/10.1016/j.ijforecast.2008.02.005

[13] Chen, F. L., and T. Y. Ou. "Gray relation analysis and multilayer functional link network sales forecasting model for perishable food in convenience store."*Expert Systems with Applications* 36.3 (2009): 7054-7063.

[14] Chevillon, G., & Hendry, D. F. (2005). Non-parametric direct multi-step estimation for forecasting economic processes. International Journal of Forecasting, 21(2), 201-218. doi: http://dx.doi.org/10.1016/j.ijforecast.2004.08.004

[15] Chiu, Bill, Eamonn Keogh, and Stefano Lonardi. "Probabilistic discovery of time series motifs." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003.

[16] Chu, C.-W., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of Production Economics, 86(3), 217-231. doi: http://dx.doi.org/10.1016/S0925-5273(03)00068-9

[17] Corberán-Vallet, A., Bermúdez, J. D., & Vercher, E. (2011). Forecasting correlated time series with exponential smoothing models. International Journal of Forecasting, 27(2), 252-265. doi: http://dx.doi.org/10.1016/j.ijforecast.2010.06.003

[18] Dalhart, G. (1974). Class seasonality—A new approach. Paper presented at the American Production and Control Society Conference Proceedings.

[19] Dekker, M., van Donselaar, K., & Ouwehand, P. (2004). How to use aggregation and combined forecasting to improve seasonal demand forecasts. International Journal of Production Economics, 90(2), 151-167. doi: http://dx.doi.org/10.1016/j.ijpe.2004.02.004

[20] Ding, Hui, et al. "Querying and mining of time series data: experimental comparison of representations and distance measures." *Proceedings of the VLDB Endowment* 1.2 (2008): 1542-1552.

[21] Duncan, G. T., Gorr, W., & Szczypula, J. (2001). Forecasting analogous time series. INTERNATIONAL SERIES IN OPERATIONS RESEARCH AND MANAGEMENT SCIENCE, 195-214.

[22] Erdem, T. (1996). A dynamic analysis of market structure based on panel data. Marketing Science, 15(4), 359-378.

[23] Frees, E. W., & Miller, T. W. (2004). Sales forecasting using longitudinal data models. International Journal of Forecasting, 20(1), 99-114. doi: http://dx.doi.org/10.1016/S0169-2070(03)00005-0

[24] Fu, Ada Wai-Chee, et al. "Scaling and time warping in time series querying."*The VLDB Journal—The International Journal on Very Large Data Bases* 17.4 (2008): 899-921.

[25] Fu, Tak-chung. "A review on time series data mining." *Engineering Applications of Artificial Intelligence* 24.1 (2011): 164-181.

[26] Gao, Yue-Fang, et al. "A neural-network-based forecasting algorithm for retail industry." *Machine Learning and Cybernetics, 2009 International Conference on*. Vol. 2. IEEE, 2009.

[27] Giorgino, Toni. "Computing and visualizing dynamic time warping alignments in R: the dtw package." Journal of Statistical Software 31.7 (2009): 1-24.

[28] Greene, W. H. (2008). Econometric analysis (6th ed.). Upper Saddle River, N.J.: Prentice Hall.

[29] Guadagni, P. M., & Little, J. D. (1983). A logit model of brand choice calibrated on scanner data. Marketing Science, 2(3), 203-238.

[30] Gür Ali, Ö., (2013). Driver Moderator Method for Retail Sales Prediction. International Journal of Information Technology and Decision Making

[31] Gür Ali, Ö., & Yaman, K. (2013). Selecting rows and columns for training support vector regression models with large retail datasets. European Journal of Operational Research, 226(3), 471-480. doi: http://dx.doi.org/10.1016/j.ejor.2012.11.013

[32] GürAli, Ö., Sayin, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. Expert Systems with Applications, 36(10), 12340-12348.

[33] Gür Ali, Ö., Pınar, E. (2013) Multi-Step-Ahead Retail Sales Forecasts and Insights about Impact of Price, Promotion and External Drivers. In Review for International Journal of Forecasting.

[34] Hoch, S. J., Kim, B. D., Montgomery, A. L., & Rossi, P. E. (1995). Determinants of Store-Level Price Elasticity. Journal of Marketing Research, 32(1), 17-29. doi: Doi 10.2307/3152107

[35] Jeong, Young-Seon, Myong K. Jeong, and Olufemi A. Omitaomu. "Weighted dynamic time warping for time series classification." *Pattern Recognition* 44.9 (2011): 2231-2240.

[36] Keogh, Eamonn, and Shruti Kasetty. "On the need for time series data mining benchmarks: a survey and empirical demonstration." Data Mining and Knowledge Discovery 7.4 (2003): 349-371.

[37] Keogh, Eamonn J., and Michael J. Pazzani. "Derivative dynamic time warping."*the 1st SIAM Int. Conf. on Data Mining (SDM-2001), Chicago, IL, USA*. 2001.

[38] Keogh, Eamonn J., and Michael J. Pazzani. "Scaling up dynamic time warping for datamining applications." *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000.

[39] Lazarevic, Aleksandar, et al. "Clustering-regression-ordering steps for knowledge discovery in spatial databases." *Neural Networks, 1999. IJCNN'99. International Joint Conference on*. Vol. 4. IEEE, 1999.

[40] Lee, Myung-Soo, B. Elango, and Steven P. Schnaars. "The accuracy of the Conference Board's buying plans index: A comparison of judgmental vs. extrapolation forecasting methods." *International Journal of Forecasting* 13.1 (1997): 127-135.

[41] Lin, Rake& Agrawal King-lp, and Harpreet S. Sawhney Kyuseok Shim. "Fast similarity search in the presence of noise, scaling, and translation in time-series databases." VLDB, 1995.

[42] Lu, C.-J., & Wang, Y.-W. (2010). Combining independent component analysis and growing hierarchical self-organizing maps with support vector regression in product demand forecasting. International Journal of Production Economics, 128(2), 603-613.

[43] Mcintyre, S. H., Achabal, D. D., & Miller, C. M. (1993). Applying Case-Based Reasoning to Forecasting Retail Sales. Journal of Retailing, 69(4), 372-398. doi: Doi 10.1016/0022-4359(93)90014-A

[44] Mentzer, J. T., & Bienstock, C. C. (1998). Sales forecasting management : understanding the techniques, systems, and management of the sales forecasting process. Thousand Oaks: Sage Publications.

[45] Mohammed Waleed Kadous (2002). http://www.cse.unsw.edu.au/~waleed/phd/html/node1.html

[46] Niennattrakul, Vit, and Chotirat Ann Ratanamahatana. "On clustering multimedia time series data using k-means and dynamic time

warping."*Multimedia and Ubiquitous Engineering, 2007. MUE'07. International Conference on*. IEEE, 2007.

[47] Perng, C-S., et al. "Landmarks: a new model for similarity-based pattern querying in time series databases." *Data Engineering, 2000. Proceedings. 16th International Conference on*. IEEE, 2000.

[48] Rakthanmanon, Thanawin, et al. "Searching and mining trillions of time series subsequences under dynamic time warping." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.

[49] Ratanamahatana, Chotirat Ann, et al. "Mining time series data." Data Mining and Knowledge Discovery Handbook. Springer US, 2010. 1049-1077.

[50] Ratanamahatana, C. A., & Keogh, E. (2005). Three Myths about Dynamic Time Warping Data Mining. Proceedings of the Fifth Siam International Conference on Data Mining, 506-510.

[51] Sakoe, Hiroaki, and Seibi Chiba. "Dynamic programming algorithm optimization for spoken word recognition." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 26.1 (1978): 43-49.

[52] Scott Armstrong, J., and Fred Collopy. "Causal forces: Structuring knowledge for time-series extrapolation." *Journal of Forecasting* 12.2 (1993): 103-115.

[53] Collopy, Fred, and J. Scott Armstrong. "Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations." *Management Science* 38.10 (1992): 1394-1414.

[54] Tang, Liang, Tao Li, and Larisa Shwartz. "Discovering lag intervals for temporal dependencies." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.

[55] Theodoridis, Sergios, et al. *Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach*. Access Online via Elsevier, 2010.

[56] Torgo, Luís, and Joaquim Pinto da Costa. "Clustered partial linear regression."*Machine Learning: ECML 2000*. Springer Berlin Heidelberg, 2000. 426-436.

[57] Toyoda, Machiko, and Yasushi Sakurai. "Discovery of cross-similarity in data streams." *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*. IEEE, 2010.

[58] Vintsyuk, T. K. "Speech discrimination by dynamic programming." *Cybernetics and Systems Analysis* 4.1 (1968): 52-57.

[59] Withycombe, R. (1989). Forecasting with combined seasonal indices. International Journal of Forecasting 5(4), 547-552.

[60] Wooldridge, J. M. (2009). Introductory econometrics : a modern approach (4th ed.). Mason, OH: South Western, Cengage Learning.

[61] Zotteri, G., & Kalchschmidt, M. (2007). A model for selecting the appropriate level of aggregation in forecasting processes. International Journal of Production Economics, 108(1–2), 74-83. doi: http://dx.doi.org/10.1016/j.ijpe.2006.12.030