VRIJE UNIVERSITEIT AMSTERDAM

RESEARCH PAPER

Hierarchical Agglomerative Clustering for Product Sales Forecasting

Author:
R.E. VAN RUITENBEEK

Supervisor:
Prof. G. KOOLE

January 22, 2019

Hierarchical Agglomerative Clustering for Product Sales Forecasting

Author:
R.E. VAN RUITENBEEK

Supervisor:
Prof. G. KOOLE

A research paper submitted in fulfillment of the requirements for the degree of MSc Business Analytics

Vrije Universiteit Amsterdam
Faculty of Science
Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam

January 22, 2019

Preface

This research paper has been written to fulfill the requirements for the master Business Analytics at the Vrije Universiteit Amsterdam. This research paper has been conducted from June 2018 till January 2019 for the amount of 6 EC in the course module: *Research Paper Buiness Analytics*.

This research paper investigates the relative effectiveness of cluster aggregated sales components, on the daily forecasting accuracy, for outdoor sport articles. Its draws the comparison over a test period of one year for approximately 3000 different products.

I would like to thank my university supervisor, Prof. G. KOOLE, for the support and guidance during the process towards fulfillment of the requirements concerning this research paper. The guidance and support, provided by Prof. G. KOOLE, have been key to develop this research paper to my best abilities.

Lastly, I would like to thank the host organization Pon Holdings BV for facilitating the data required to conduct this research successfully.

VRIJE UNIVERSITEIT AMSTERDAM
Faculty of Science
Business Analytics

Abstract

This research paper investigates the relative effectiveness of cluster aggregated sales components, on the daily forecasting accuracy of outdoor sport articles. A comparison is made between forecasting using an individual product and forecasting using aggregated time series. The effect of aggregation is examined by use of predefined product groups, from the business, and product groups constructed with hierarchical agglomerative clustering. A case study is performed with over 3000 unique products, showing that forecasting can benefit from clustering depending on the nature of time series.

Keywords: forecasting, clustering, aggregation, time series

Contents

List of Figures

List of Tables

List of Abbreviations

APE	**A**bsolute **P**ercentage **E**rror
B2B	**B**usiness to **B**usiness
B2C	**B**usiness to **C**onsumer
BU	**B**ottom-**U**p
DTW	**D**ynamic **T**ime **W**arping
EOQ	**E**conomical **O**rder **Q**uantity
EPQ	**E**conomical **P**roduction **Q**uantity
FT	**F**ourier **T**ransformation
GA-DTW	**G**lobal **A**lignment **D**ynamic **T**ime **W**arping
GLM	**G**eneralizsed **L**inear **M**odels
HAC	**H**ierarchical **A**gglomerative **C**lustering
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
MAE	**M**ean **A**bsolute **E**rror
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
MASE	**M**ean **A**bsolute **S**caled **E**rror
MLE	**M**aximum-**L**ikelihood **E**stimator
MSE	**M**ean **S**quared **E**rror
MUSS	**MU**lti-**S**cale **S**moothing
OHE	**O**ne-**H**ot-**E**ncoding
RMSE	**R**oot **M**ean **S**quared **E**rror
SKU	**S**tock-**K**eeping **U**nit
SVM	**S**upport **V**ector **M**achine
TD	**T**op-**D**own
WAPE	**W**eighted **A**bsolute **P**ercentage **E**rror

Chapter 1

Management Summary

This research identifies the influence of cluster aggregated sales data on individual SKU forecasts. It proposes a framework to cluster products, based on historical sales, to enhance the forecasting performance on a daily level. Furthermore, it performs a detailed data analysis on sales data and shows the relative improvement on the forecast accuracy, obtained from cluster influence. Lastly, it establishes the connection between the forecast results and its application in the business, by identifying how inventory management should deal with fluctuations and uncertainty obtained in the forecast.

Research question: How can time series clustering enhance daily forecasting?

This research shows that forecasting can benefit from predefined product groups from the business. However, it shows that using these product groups for the wrong time series can dramatically increase the forecasting error. This research shows that hierarchical agglomerative clustering generally outperforms the predefined product groups. Furthermore, it shows that the benefits of the clustering approach are larger, compared with the predefined groups. Additionally,the increase in the forecasting error is limited for the time series where a detailed approach should be preferred.

In general, a forecast on an individual product is preferred when the sales frequency is rather high or the variation is rather low. The clustering approach is preferred for a low sales frequency or a high variation within the time series. Therefore, it can be concluded that clustering can enhance the accuracy for daily forecasting for specific time series. In all cases, the predefined product groups are depreciated in the general line. Applying the cluster approach when the detailed approach is preferred, will result in an increased error and should therefore be threaded with care.

When no attention is payed to the type of time series, both the detailed approach and the clustering approach are performing rather similar. It is therefore advised to treat each time series accordingly to its characteristics, to apply either the detailed approach or the clustering approach.

Chapter 2

Introduction

For decades, companies have been trying to cut costs while increase efficiency in the business process. One of the methods used in achieving this efficiency is forecasting. Forecasting is used excessively to keep track of financial goals, estimate future sales, predict employee requirements and more.

So far, a large proportion of researchers have studied different types of forecasting models, to identify the strengths and weaknesses of each model. However, most research has focused on forecasting for an individual product, neglecting relationships between products. Zotteri and Kalchschmidt (2007) and Zotteri, Kalchschmidt, and Caniato (2005) argue that aggregation of multiple products could be used to improve the forecasting accuracy on a product level.

This paper tries to identify if forecasting on a product level can benefit from product aggregation. It examines this by use of a clustering technique to identify similarity between time series. Furthermore, it compares these clusters with a regular forecast approach to identify the potential value of aggregation within a cluster. Additionally, this paper adds the comparison with predefined product groups. Therefore, this paper tries to answer the following question:

How can time series clustering enhance daily forecasting?

This research shows that forecasting on a product level can benefit from sales aggregation. It shows that the use of predefined product groups, from the business, should be treated with care when used to perform the aggregation. Furthermore, it shows that the use of clustering techniques, before applying aggregation, can result in a forecast accuracy improvement. Although the research shows that this improvement can be achieved, it also shows that not all time series can benefit from this approach. This research tries to draw a distinction between the time series on which this approach is or is not applicable.

This paper starts with a literature study, identifying previous research on the effect of time series aggregation for forecasting. Besides this, it examines past research in measuring and identifying time series similarity. The main part of this research consists of a case study where a detailed data analysis is performed in section 4. Section 5 elaborates on the framework used to test the effect of time series clustering and explains the models and methods used to conduct this research. Lastly, the results are presented in section 6, followed by a conclusion and a call for further research in section 7.

Chapter 3

Literature Review

The different types of companies where forecasting plays an important role is almost inexhaustible, making it a widely investigated area by both academicians and practitioners. The interdisciplinary nature of forecasting makes it possible to apply the same forecasting technique in a variety of business perspectives, ranging from advanced budget forecasting, to the prediction of illness. Many different models have been adopted throughout the past decades, where the first models were implemented without the use of computational power. Nowadays, more advanced models are applied with interference from the field of computer science, to improve accuracy and to handle the ever increasing amount of data.

Although the field of forecasting keeps expanding, relatively little research has been done on relationships between time series (Zotteri and Kalchschmidt, 2007). Forecasting models are mainly focusing on an individual SKU, neglecting potentially valuable relationships with other products. A large proportion of papers argue that aggregation of time series improves the ability to estimate trend and seasonality (Zotteri, Kalchschmidt, and Caniato, 2005; Zotteri and Kalchschmidt, 2007; Rostami-Tabar et al., 2015). Furthermore, Babai, Ali, and Kourentzes (2012) and T. Tabar (2013) argue that aggregation could increase the frequency and quantity which reduces the number of zero observations in slow-moving products.

Orcutt, Watts, and Edwards (1968), Barnea and Lakonishok (1980) and Fliedner (1999) showed that the preference of demand aggregation over forecasting on single SKU levels, depends strongly on the correlation between the aggregated time series. Zotteri and Kalchschmidt (2007) suggests to use aggregation for product with short historical information, or after unexpected changes in the environment such as promotions. A detailed forecast on SKU level can then be used when enough historical information has been collected. Although their research focuses on the aggregation between stores, they argue that their findings could be applied to aggregation of SKUs. Even though no specific time series length is given, it gives rise to the idea that aggregation would be most beneficial for recently introduced products.

3.1 Aggregation

Two types of time series aggregation are frequently studied within the literature (Babai, Ali, and Kourentzes, 2012). The most common aggregation type is temporal aggregation, transforming a high frequency time series into a lower frequency by non-overlapping aggregation segments. Kourentzes, Petropoulos, and Trapero (2014) showed that temporal aggregation on different levels can improve the

overall forecasting accuracy. They proposed a framework showing that combining exponential smoothing at different aggregation levels, allows for more complex demand patterns. The key is the ability to model different seasonal and trend patterns separately over each aggregation level. The proposed framework aggregates each individual time series to k aggregated time series and applies exponential smoothing separately on each of the aggregation levels. The separate levels are eventually disaggregated by an averaging function, resulting in one final forecast.

The second aggregation method mentioned by Babai, Ali, and Kourentzes (2012) is cross-sectional aggregation, focusing on the relationships between time series for different SKUs. The effectiveness and practical application of this method has frequently been discussed in the literature. Rostami-Tabar et al. (2015) distinguishes and compares the top-down (TD) and bottom-up (BU) approach for cross-sectional aggregation. Commonly refereed to as, aggregated forecasting for TD and sub-aggregated forecasting for BU. We will go with the TD and BU convention to avoid name confusion with temporal aggregation. Rostami-Tabar et al. (2015) found that both types of cross-sectional aggregation achieved performance benefits for non-stationary and stationary time series, compared with single SKU level forecasting. They found that non-stationary time series showed the highest accuracy improvement. They found that the BU approach outperforms TD when cross-sectional correlations are negative, or relatively low and positive. Furthermore, they argued that TD is preferred when correlations are relatively high.

Gross and Sohl (1990) empirically compared different disaggregation methods in combination with multiple forecasting methods. They used the TD approach to examine the reduction in accuracy in relation with the time savings by forecast aggregation. They concluded that the disaggregation was applicable in two out of the three product lines, used in the empirical research.

So far, the majority of studies on cross-sectional aggregation focused on aggregation within predefined product families. Most companies use SKU mappings toward groups or families in order to generate higher level forecasts (Chen and Boylan, 2008). This aggregation process is highly affected by the predefined groups which need to be generated by the business. Furthermore, it assumes that the SKUs within each group are following similar sales patterns, which might not always be the case. Similarity in product families is mostly based on product features or names and not on the desired sales similarity.

3.2 Time Series Similarity

Automatic clustering of time series based on similarity has not frequently been studied in the literature. So far, most studies focused on using the predefined groups developed from business knowledge, to assess the cross-sectional aggregation effects. However, time series similarity as a broad topic is more commonly examined and therefore explored in this section.

Distance metrics are most widely used for defining similarities between multiple time series. The most commonly known metrices are the Manhattan distance and the Euclidean distance which are both generalised under the Minkowski distance (Jain, Murty, and Flynn, 1999). Jain, Murty, and Flynn (1999) pointed out that the Minkowski distance metrics gives large-scaled features the tendency to dominate and they argued that this metric works best in compact and isolated clusters. Bergen et al. (2005) used the Mahalanobis distance metrics to classify

land cover which overcomes the main disadvantage of the Minkowski metrics by accounting for non-stationarity of variance (Lhermitte et al., 2011).

In the process of time series similarity, many authors used the Fourier Transformation (FT) before applying a distance metrics. Azzali and Menenti (2000) used the FT in combination with a proposed distance metrics based on the Euclidean distance. Use of the Euclidean distance results in sensitivity to amplitude scaling, time scaling and time translation controversial to the other metrics (Lhermitte et al., 2011). Furthermore, Evans and Geerken (2006) proposed a more shape based similarity measure after applying the FT.

Troncoso, Arias, and Riquelme (2015) presented a Multi-scale smoothing kernel (MUSS) for measuring time series similarity. The major advantage of this kernel is that it accounts for minor shifts in time as well as misalignments by focusing on similarities in shape rather than absolute values. They examined the proposed kernel against a linear kernel and the DTW kernel, in combination with a SVM classifier for several datasets. The MUSS kernel showed similar results in terms of accuracy compared with the well applied GA-DTW kernel and outperformed the linear kernel. However, the MUSS kernel showed to be much faster than the GA-DTW kernel making it more applicable in real word situations. A drawback of the MUSS kernel, is that it does not account for difference in time series length, which arises for sales data. Furthermore, both DTW and MUSS are computationally expensive in contrast with the correlation metrics or the Minkowski distances.

Another time series clustering, based on fuzzy sets, has been proposed by Shou-Hsiung, Shyi-Ming, and Wen-Shan (2016). They showed by use of empirical research that the proposed forecast method results in a higher accuracy compared with previous fuzzy clustering methods.

One of the most straight forward options is the use of correlation to define similarity. Correlation coefficients such as the Pearson correlation are independent of scale. Another measure closely related to Pearson correlation is the cosine similarity which defines vector differences in terms of the angle.

Chapter 4

Data

The data used in this empirical research has been made available by a daughter company of Pon Holding, which requested to remain anonymous. Upon request, any information regarding the companies nature has been left out and product codes have been hashed. The company in question mainly sells to retail stores (B2B), whereas a small proportion is sold directly to the customer by means of a web-shop (B2C). The two-sided market results in a strong diversity within the sold quantity per customer, as larger batches can be purchased by the B2B market. The sold products are all centered around the same outdoor sport, resulting in highly seasonal data.

4.1 Data Description

The original dataset consists of 8278 unique products or product configurations[1], with in total approximately four million unique sales records. The corresponding sales records are recorded on a daily level where the product number, the date and the quantity are registered. A data sample is shown in appendix A. The data contains some rows (0.38 percent) with missing values in either the date or the product number.

The data has been collected over the period from 2009-03-02 to 2017-10-19. Furthermore, eight observations are deviating from this period, all being registered on 1899-12-31 with a quantity of one. These eight observations are removed, as expected to be errors. All products are introduced between the start and end date of the dataset where the first sold date is considered to be the introduction date. Product introductions frequently deviate from the regular sales pattern, being significantly higher or lower than the remaining time series. These introduction effects are further discussed in section 4.5.

Due to the occurrences of introductions and products leaving the portfolio, the time series length varies strongly throughout the data. A histogram of the length in days of the product history[2] is shown in figure 4.1. There is a remarkable number of products (453) without any sales, resulting in a history of zero days.

[1] A product configuration is defined as a different colour or size.

[2] Difference between the first sold date and the last sold date.

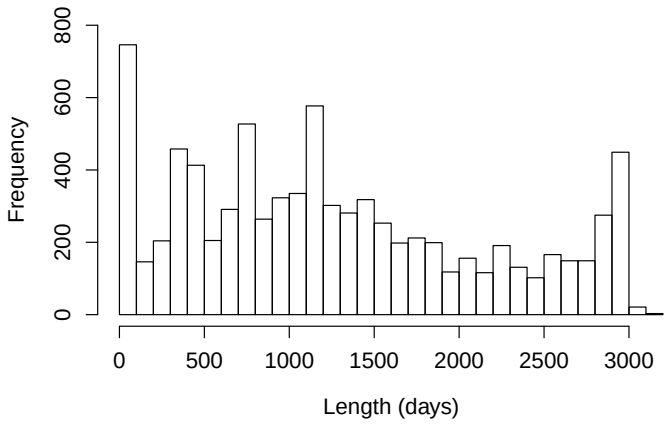

FIGURE 4.1: Time series length in days for all 8278 products (bin size 100).

The average sold quantity deviates strongly between the products where some are slow moving and others are fast moving. Figure 4.2 shows the coefficient of variation ($CV = \frac{\sigma}{\mu}$) for each time series in relation with the fraction of days where the product has no sales. The behaviour of figure 4.2 shows that an increase in the zero fraction results in general in a larger coefficient of variation. This indicates that an increase in the zero fraction results in a large σ compared to μ.

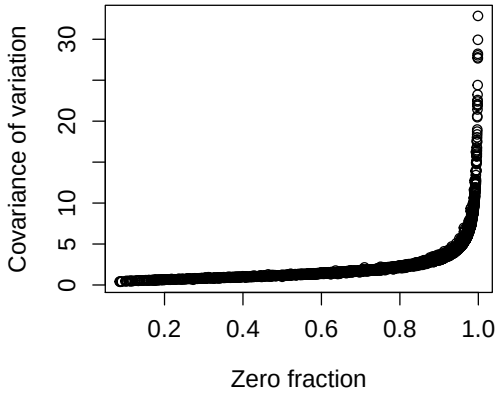

FIGURE 4.2: Relationship between the fraction of zero observations (no sales on a day) and the coefficient of variation.

4.2 Data Preprocessing

Not all products and sales records are fit for forecasting models and therefore, data preparation is required. A large proportion of products is excluded from this research, where figure 4.3 shows the number of excluded products per cause.

FIGURE 4.3: Excluded number of products per cause. Total 5395 products excluded, resulting in 2892 products for this research.

The sales records are prepared by removing any negative sales occurrences. Negative sales occur as a result of returns which cannot be linked to the actual sales date. Therefore, excluding these negative sales will prevent from reduced sales at the return date which should, in fact, be subtracted at the sales date.

The explained 0.38 percent of sales records without either a product number, sold date, or quantity are removed from the data. Furthermore, any outliers with large sales quantities are kept in the data, as these are highly affecting the inventory management and therefore crucial to incorporate in the forecast.

Research by Zotteri and Kalchschmidt (2007) shows that aggregation is beneficial for shorter time series and of less added value when more data becomes available. Therefore, relatively short time series of at least 2 years will be kept, to ensure at least one year of train data and one year of test data.

4.3 Distributions

There is a large difference in frequency between time series. Figure 4.4 (left) shows a fast moving item, having a baseline above zero and a clear monthly seasonality. Contrary, figure 4.4 (right) shows a slow moving item with many zero observations but a rather similar seasonality.

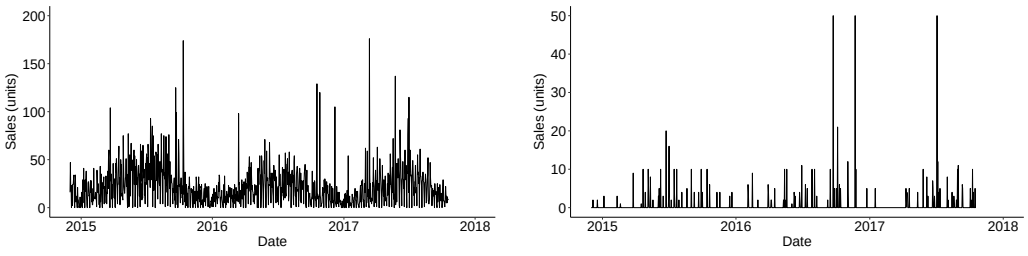

FIGURE 4.4: Fast moving product with clear seasonality (left) and slow moving product with less clear seasonality (right).

The different sales patterns are reflecting in the distributions of the sales count. Figure 4.5 shows the distribution of both products and its log-transformed[3] sales. The distribution of the fast moving product shows an exponential distribution where the log transformation transforms this to a distribution not significantly deviating from normal, when neglecting the observations at zero. The slow moving product shows a similar pattern whereas the proportion of zero observations is much higher.

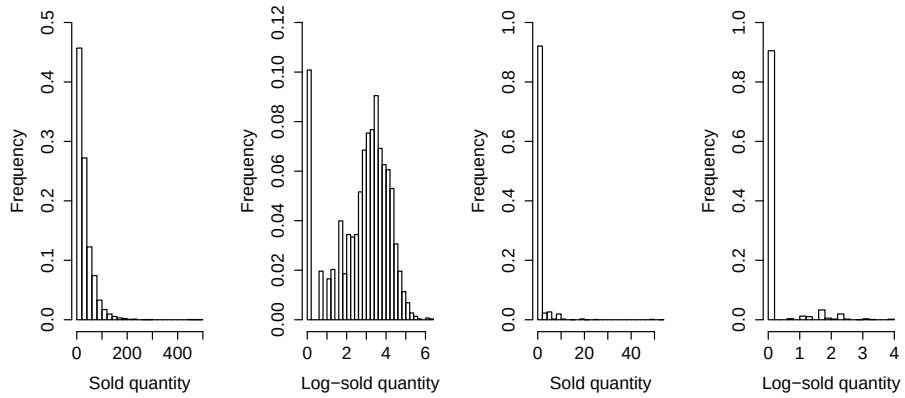

FIGURE 4.5: Probability density plot for a fast moving product (left) and a slow moving product (right).

4.4 Periodicity

Figure 4.4 indicates strong evidence for an intra-year seasonality which is shown to be significant in figure 4.6.

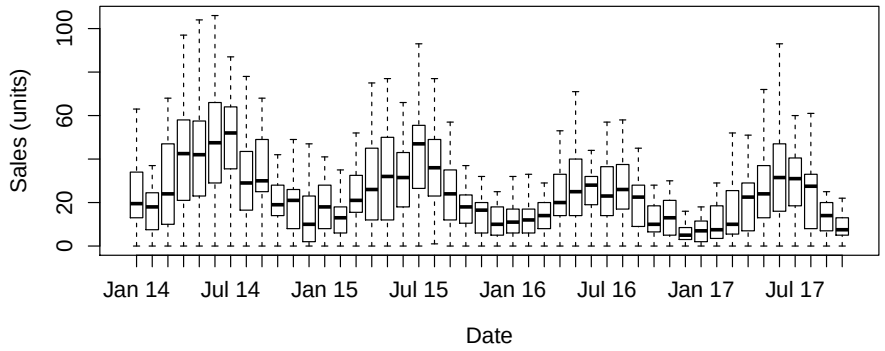

FIGURE 4.6: Sales effect per year and month for one product.

[3]Log transformation is applied by $Log(sales + 1)$ to overcome non existence of $log(0)$.

The demand pattern throughout the week is shown in figure 4.7, indicating a decreasing demand trend from Monday to Sunday. The quantities are based on the time series from figure 4.4, but is representative for the majority of products.

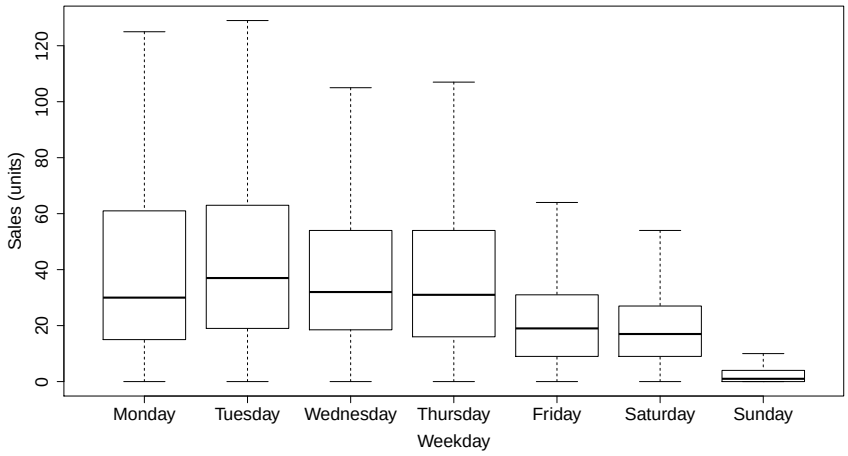

FIGURE 4.7: Sales difference per weekday.

4.5 Introductions

Product introductions occur throughout the year where a yearly reoccurring peak is observed in the months February, July and September. Product introduction are of high interest due to its deviation from the remainder of the time series. This deviation can be twofold, where the sales is either significantly lower (figure 4.8 left) or significantly higher (figure 4.8 right) than the remainder.

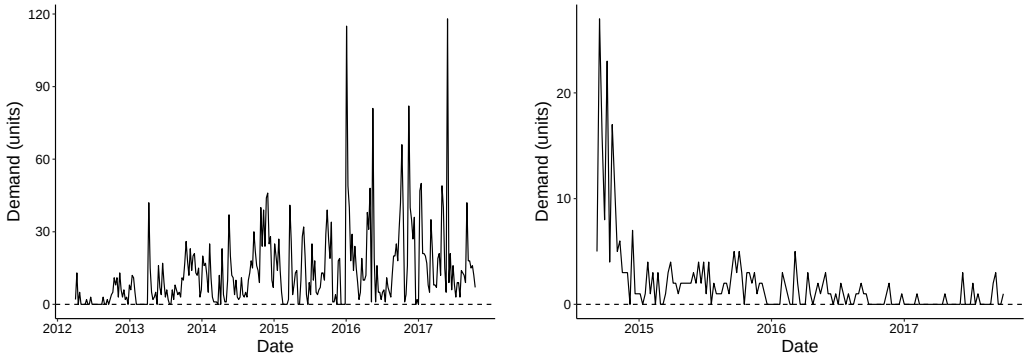

FIGURE 4.8: Product introduction patterns with slow start (left) and peak sales (right).

4.6 Holidays

Some businesses are closed on national holidays which largely affects the sales on these days. Figure 4.9 shows the difference between a regular day (first boxplot) and the seventeen dutch holidays. Additionally, the combination between the holiday and the weekday has a strong effect on the sales quantity, as well as the days surrounding the holiday. To illustrate, a holiday on Monday has a larger effect on the surrounding days then a holiday on Sunday. The sold quantity on Sunday is on average lower than Monday which results in a small added sales quantity to the surrounding days of Sunday.

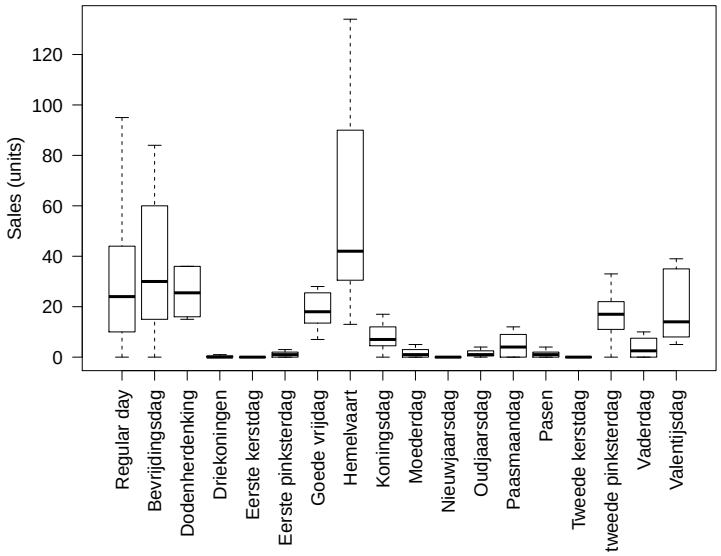

FIGURE 4.9: Holiday effect on the sales quantity.

4.7 Multiplicative Relationship

Figure 4.10 indicates a clear multiplicative relationship between trend and seasonality, which is present for the majority of products.

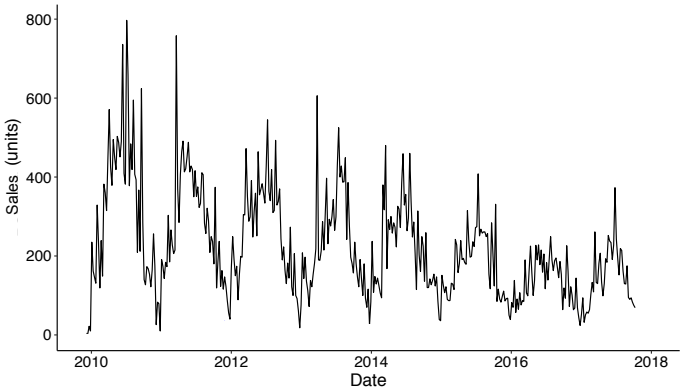

FIGURE 4.10: Multiplicative relationship between trend and seasonality.

Chapter 5

Methodology

This section describes the methodology for testing the effectiveness of product clustering on a daily forecast. Section 5.1 elaborates on the three methods used to draw this comparison. The forecasting models used in this comparison are explained in section 5.4, whereas the clustering approach is explained in section 5.2. The performance of the three methods will be evaluated with use of the WAPE, MASE and MAE as explained in section 5.7. Lastly, section 5.8 combines theory and practice by making the translation from forecasting results towards inventory management.

5.1 Framework

This empirical study examines the effectiveness of adding clustered sales information, to enhance the forecasting accuracy on a SKU based level. This comparison is constructed by developing a detailed forecasting model (I), on SKU level, and comparing its results with a forecast taking existing product group sales into account (II). Lastly, the comparison is made with product groups generated by use of time series clustering, referred to as cluster approach (III). In order to conduct the comparisons, the following three methods are introduced and graphically represented in figure 5.1.

1. **Method I**: *Detailed approach*
   (a) Forecast each SKU separately.
2. **Method II**: *Product group approach*
   (a) Forecast the aggregated demand within each predefined product group.
   (b) Forecast each SKU separately with the additional independent variable: product group forecast.
3. **Method III**: *Product cluster approach*
   (a) Cluster similar products based on historical demand.
   (b) Forecast the aggregated demand within each cluster.
   (c) Forecast each SKU separately with the additional independent variable: cluster forecast.

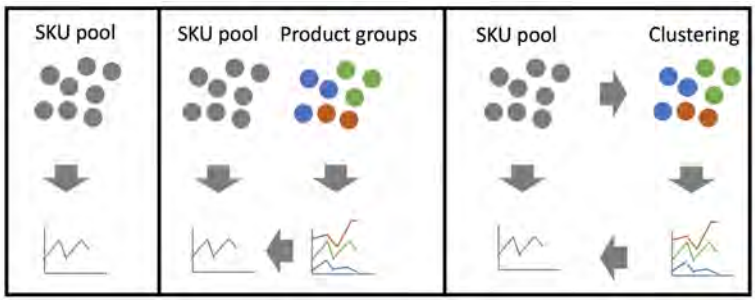

FIGURE 5.1: Graphical representation of Method I (left), Method II (middle) and method III (right).

5.2 Clustering

Flat clustering methods such as *k*-means provide a segmentation of the input data into *k* clusters. Another type of clustering is *Hierarchical Agglomerative Clustering* (HAC) which provides the notion of hierarchy in its constructed clusters. This is of particular interest in this study, as it is assumed that products can be classified hierarchical (e.g. a sweater categorizes under winter clothing which categorizes under clothing). Furthermore, HAC does not require a predefined number of clusters, giving it an advantage over other methods such as *k*-means. In this study the number of clusters is not known upfront and the clustering is computationally expensive due to the large sample size, making HAC the prefered clustering method.

HAC is defined as a bottom up algorithm, merging (agglomerating) pairs of clusters until all branches are merged to one single cluster of size n. The initial state space consists of n clusters containing exactly one observation, where $n-1$ merges are required to obtain the final cluster. The results are mostly provided in the form of a dendogram which shows the hierarchical structure of all $n-1$ merges.

5.2.1 Implementation

HAC is performed in two successive steps: construction of the dissimilarity matrix, which represents the distances between all n products, and the construction of the hierarchical clusters. The dissimilarity matrix requires the measurement of similarity between a pair of time series, which can be done by use of a large variety of metrics as explained in section 3.

Defining the right distance measure between two time series is of major importance for the final clustering. Two problems need to be accounted for in the selected distance measure: differences in scale and difference in length. The first is of importance as the scale differs strongly between pairs of time series, which should not affect the dissimilarity. To illustrate the importance, one can assume a time series A being a multiple of time series B such that $A = a \cdot B$. Obviously, the pattern is exactly the same and the pair of products should therefore be classified as completely similar. The latter is of great influence as the time series length differs strongly throughout the dataset. Misalignment between two time series where one of the two starts earlier or ends later, frequently occurs as most time series are introduced at different points in time. The used distance measure is explained in section 5.2.2.

The second step of HAC is the iterative process of merging clusters. For this, the dissimilarity of two clusters should be obtained where each iteration merges the two clusters with the lowest cluster distance. Calculation of the cluster distance is performed using a linkage function which is explained in section 5.2.3.

A simple and unoptimized implementation of the HAC algorithm is for illustrative purposes shown in Algorithm 1, reprinted from Manning, Raghavan, and Schütze (2008, p. 349).

Algorithm 1 Unoptimized HAC implementation

```
for n ← 1 to N do
    for i ← 1 to N do
        C[n][i] ← sim(d_n, d_i)
        I[n] ← 1 (keeps track of active clusters)
    end
end
A ← []
for k ← 1 to N-1 do
    (i,m) ← arg max_{[(i,m):i≠m∧I[i]=1∧I[m]=1]} C[i][m]
    A.append((i, m))
    for j ← 1 to N do
        C[i][j] ← Sim(i, m, j)
        C[j][i] ← Sim(i, m, j)
    end
    I[m] ← 0 (deactivate cluster)
end
```

5.2.2 Distance Measure

The distance measure defines the distance between each pair of time series which can be represented in an upper triangle matrix, resulting in a dissimilarity matrix. To overcome the previously explained difference in time series length, the intersection of the two time series is used to draw the comparison. The intersection is selected over the union as it gives the option for short time series to benefit from the history of long time series. This is of particular interest when the product of the short time series is introduced later but the sales pattern is rather similar.

To quantify the dissimilarity, a distance metric is required. The most widely known distance metric would be the Minkowski distances. However, this metric gives large scale features the tendency to dominate and is therefore not suited to cope with scale differences between time series (Jain, Murty, and Flynn, 1999). The correlation is also widely applied to quantify similarity between time series. The correlation is however sensitive to outliers which can result in the case where a rather similar time series gets a low correlation due to one outlier in either of the two series. More advanced measures such as MUSS and DTW as explained in section 3 are computationally too expensive to perform on the dataset.

The selected measure is the Cosine similarity which is not largely affected by outliers and has a tremendous computational speedup compared to MUSS and DTW. The following example, with time series **A** and **B** compares the Cosine similarity with the Pearson Correlation to explain the effect of outliers.

$$\mathbf{A} = [1, 2, 3, 30, 5, 6, 7, 8, 9, 10]$$
$$\mathbf{B} = [2, 4, 6, 30, 10, 12, 14, 16, 0, 20]$$

$$\text{Pearson Cor} = 0.051$$
$$\text{Cosine Sim} = 0.603$$

The above shows that the Cosine similarity gives a higher similarity when outliers occur, which is the desired result. The Cosine similarity is defined in equation (5.1) , where the dissimilarity is defined by equation (5.2). Equation (5.2) reverses the interval $[0, 1]$, which results in $dis(\mathbf{A}, \mathbf{A}) = 0$.

$$sim(\mathbf{A}, \mathbf{B}) = cos(\theta) = \frac{\langle \mathbf{A}, \mathbf{B} \rangle}{\|\mathbf{A}\| \, \|\mathbf{B}\|} \tag{5.1}$$

$$dis(\mathbf{A}, \mathbf{B}) = 1 - sim(\mathbf{A}, \mathbf{B}) \tag{5.2}$$

5.2.3 Linkage Function

The linkage function defines the distance calculation between a pair of clusters. The most common linkage functions are: Single linkage, Complete linkage, Average linkage and Centroid linkage which are defined below (Flach, 2018, pp. 254–255).

$$L_{single}(A, B) = \min_{x \in A, y \in B} dis(\mathbf{x}, \mathbf{y}) \tag{5.3}$$

$$L_{complete}(A, B) = \max_{x \in A, y \in B} dis(\mathbf{x}, \mathbf{y}) \tag{5.4}$$

$$L_{average}(A, B) = \frac{\sum_{x \in A, y \in B} dis(\mathbf{x}, \mathbf{y})}{|A| \cdot |B|} \tag{5.5}$$

$$L_{centroid}(A, B) = dis\left(\frac{\sum_{x \in A} \mathbf{x}}{|A|}, \frac{\sum_{y \in B} \mathbf{y}}{|B|}\right) \tag{5.6}$$

One major drawback of Single linkage and Complete linkage is the reduction of cluster quality to a single pair of products: the two most similar products or the two most dissimilar products (Manning, Raghavan, and Schütze, 2008, pp. 350–353). The major drawback of Single linkage is called *chaining* which could result in elongated clusters. Complete linkage on the other hand gives large importance to outliers by measuring between the two most dissimilar products.

Flach (2018, pp. 255–558) argues against Centroid linkage as only Centroid linkage violates the monotonicity constraint, defined in Theorem 5.2.1. From this, the preferred linkage function is the Average linkage as it incorporates all products within a cluster, but does not violate the monotonicity. Incorporating all products should logically be used as this research focuses on the aggregation of sales from the complete cluster.

Theorem 5.2.1 (Monotonocity) *A linkage function satisfies the monotonicity requirement if and only if the following condition holds (Flach, 2018, p. 257):*

If $L(A,B) < L(A,C)$ ***and*** $L(A,B) < L(B,C)$

Then $L(A,B) < L(A \cup B, C)$

5.2.4 Number of Clusters

It is expected that the final forecast accuracy depends on the selected number of clusters. Too large clusters could result in strong time series dissimilarity, whereas too small clusters could loose the stability of the group information. Manning, Raghavan, and Schütze (2008, pp. 348–349) describe the following four methods to determine the optimal number of clusters. This research will make use of the Elbow method to find the optimal number of clusters graphically. This method is selected as it is one of the most widely used methods and the most intuitive in use.

1. **Predefined number of clusters**
   Cutting-off the dendogram at exactly k clusters (k-means approach).
2. **Predefined similarity**
   Cutting-off the dendogram at a similarity s and accepting the resulting k number of clusters.
3. **Largest gab identification (knee method or elbow method)**
   The elbow method is a graphical method which requires to calculate for each number of clusters the sum of the distances within the clusters. The optimal number of clusters can then be defined as the value for which the monotonically decreasing line shows an elbow.
4. **Distortion method**
   λ is described as the penalty value for each additional cluster where the $RSS(K')$ is the residual sum of squares for cluster size k' which can be replaced by other distortion measures.

$$K = \underset{k'}{\operatorname{argmin}}[RSS(K') + \lambda K'] \tag{5.7}$$

5.3 Train and Test Data

The forecast is performed for two weeks ahead, which will be referred to as a batch. Due to the presence of yearly seasonality, the two week test period selected in the year will largely influence the final forecast accuracy. To overcome this problem, a sliding window method is applied where the test data of 364 days (52 weeks) is used to cover the complete year. This method results in 26 batches which need to be forecasted to cover the complete year. This method is graphically presented in figure 5.2. The final forecast accuracy can then be obtained by applying the accuracy measure over the complete year to obtain an average accuracy.

FIGURE 5.2: Sliding window approach for splitting train and test data.

5.4 Forecasting Models

Several implementations of the Generalized Linear Models (GLMs) and a special case of these models, Linear regression, are implemented to produce a daily forecast. First the Linear Regression model is explained in section 5.4.2, after which the broader GLMs are explained. From this class, the Poisson Model is used as well as a stacked model having a Logistic regression implementation followed by a Linear regression model.

5.4.1 Naive Models

Two naive forecasting models are applied to serve as a benchmark to the regression models. These two models are referred to as *Naive last batch* and *Naive last year*.

- Naive last batch
  The forecast for the next batch will equal the current sales. This means that the upcoming two weeks will equal the past two weeks. In terms of days: $F_t = Y_{t-14}$.
- Naive last year
  The forecast for the upcoming two weeks will equal the sales in the same period last year.

5.4.2 Linear Regression

Linear regression has been widely applied, where the relevance in the business has gained much attention due to its convenience to interpret the constructed model. Besides this, the regression model is used in this paper due to its ability to incorporate relatively complex patterns, such as holiday effects, by use of dummy variables.

During this section, the mathematical formulation of Bijma, Jonker, and Van der Vaart (2013) will be used. Linear regression is build on two components. The response variable, often defined as **Y**, and the explanatory variables **X**.

The linear regression model with independent variables $y_1, ..., y_n$ and p-dimensional explanatory variables $(x_{1,1}, ..., x_{1,p}), ..., (x_{n,1}, ..., x_{n,p})$ is mathematically defined in equation (5.8).

$$Y_i = \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i \qquad i = 1, ..., n \tag{5.8}$$
$$\epsilon_i \sim N(0, \sigma^2) \qquad i.i.d$$

The above formulation can be simplified to matrix notation using $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ as vector of dependent variables. Furthermore, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ represent the regression coefficients. The design matrix $\mathbf{X}_{nxp}$ contains the p explanatory variables for n observations.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{5.9}$$
$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I})$$

5.4.3 Generalized Linear Models

GLMs is a generalization of the Linear regression model, which allow for data with a probability density distribution originating from the exponential family (Kedem and Fokianos, 2002, pp. 142–143). The Linear Model satisfies this condition as the Normal distribution can be rewritten into the canonical form and is therefore a special case of the Generalized Linear Models. GLMs enable to create regression models with different distributions such as binomial for classification. Each GLM consists of the following three components as explained by Gunst (2013, pp. 41–42):

1. **Random Component:** Specifies the distribution of Y_i, indicating the distribution of the uncertainty:
$$Y_i \sim f_i \tag{5.10}$$

2. **Systematic Component:** Vector of predictors, with X_i the independent variables and β the weight of each variable:
$$\eta_i = X_i^T \beta \tag{5.11}$$

3. **Link Function:** Specifies the connection between the random and systematic component. An example of a link function is the Sigmoid function which transforms a real number into a binary classification:
$$\eta_i = g(\mu_i) \tag{5.12}$$

The link function can be derived by rewriting the distribution into the canonical form described in equation (5.13) (Kedem and Fokianos, 2002, pp. 4–6). Additionally, the canonical form can be used to show that a distribution originates from the exponential family.

$$f_i(y) = f_i(y, \theta_i) = exp\left(\frac{y\theta_i - b(\theta_i)}{\frac{\phi}{A_i}} + c\left(y, \frac{\phi}{A_i}\right) \right) \tag{5.13}$$

Kedem and Fokianos (2002, pp. 4–6) describe that the parameters of the canonical form can found by:

$$\mathbb{E}(y_i) = \mu_i = b'(\theta_i) \tag{5.14}$$

$$\mathbb{V}(y_i) = b''(\theta_i)\frac{\phi}{A_i} \tag{5.15}$$

Poisson Regression Model

Poisson regression, or log-linear regression is especially of use when dealing with count data. Furthermore, it is of use when the observed values are defined in the positive domain $\mathbb{Z}^+$, having $\mathbb{E}(y_i) > 0$. This method removes the need for a log-transformation which avoids $log(0)$ (Kedem and Fokianos, 2002, pp. 143–149).

The Poisson distribution can be rewritten to the canonical form and is therefore a proven distribution from the exponential family. Rewriting the Poisson distribution to the canonical form (equation (5.16)) will give the link function of the GLM.

$$\begin{aligned} f_i(y_i) &= \frac{e^{-\theta_i}\theta^{y_i}}{y_i!} \\ log(f_i(y_i)) &= y_i \cdot log(\theta_i) - \theta_i - log(y_i!) \\ f_i(y_i) &= exp\Big(\frac{y_i \cdot log(\theta_i) - \theta_i}{1} - log(y_i!)\Big) \end{aligned} \tag{5.16}$$

From the above derivation, in combination with the canonical form, the following parameters can be derived:

$$\begin{aligned} b(\theta_i) &= e^{\theta_i} \\ \phi &= 1 \\ A_i &= 1 \\ c\Big(y, \frac{\phi}{A_i}\Big) &= -log(y_i!) \end{aligned} \tag{5.17}$$

Kedem and Fokianos (2002, pp. 4–6) show that the link function can then be derived by:

$$\begin{aligned} \mu_i &= g^{-1}(b'(\theta)) \\ \mu_i &= g^{-1}(e^{\theta_i}) \\ \theta_i &= log(\mu_i) \end{aligned} \tag{5.18}$$

The three resulting components are shown below (Gunst, 2013, p. 44):

$$\text{Poisson Regression} \begin{cases} Y_i \sim Poisson(\mu_i) \\ \eta_i = X_i^T\beta \\ \eta_i = g(\mu_i) = log(\mu_i) \end{cases} \tag{5.19}$$

Although the link function is $log(\mu_i)$, the Poisson regression should not be confused with Linear regression on the log-transformed (e.g. $Y' = log(Y)$). For the

Linear model on the log-transformed data, the left hand site becomes: $\mathbb{E}(Y'|x) = \mathbb{E}(log(Y)|x)$. From this, Poisson regression is not equivalent with Linear regression on the log-transformed data. Unless Y is fully determined by x, the following holds:

$$\mathbb{E}(log(Y)|x) \neq log(\mathbb{E}(Y|x)) \tag{5.20}$$

Stacked Model

The data analysis showed a large proportion of zero observations occurring in the data. This can be modelled as a succession of two models where the first model predicts if sales occurs at time t (1/0). A second model predicts the quantity if the first model predicts sales. This methodology removes the large number of zero observations from the distribution of the second model, making the estimation of this model less biased. The stacked model implements Logistic regression followed by Linear regression.

The logistic regression model is defined in the space {0,1} and can be formulated by a similar derivation as done for the Poisson regression, using the notation of Gunst (2013, p. 44). The three resulting components are:

$$\text{Logistic Regression} \begin{cases} n_i Y_i \sim Binomial(n_i, \mu_i) \\ \eta_i = X_i^T \beta \\ \eta_i = g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) \end{cases} \tag{5.21}$$

5.5 Feature Engineering

Feature engineering is among the most important steps in the development of a representative regression model. Well designed features have the ability to improve the regression model by allowing for more complex patterns in the data.

5.5.1 Transformations

Section 4.7 explains the existence of a multiplicative relationship between trend and seasonality. This implicates that an increase in trend gives rise to a larger seasonal effect and the other way around. The multiplicative relationship is transformed to an additive relationship by use of a log transformation, displayed in equation 5.22.

$$\begin{aligned} Y_i &= T_i \cdot S_i \cdot \epsilon_i \\ log(Y_i) &= log(T_i) + log(S_i) + log(\epsilon_i) \end{aligned} \tag{5.22}$$

To overcome the undefined $log(0)$, which occur on days having a demand of zero, the $log(y_i)$ is replaced by $log(y_i + 1)$. The predicted values are reversed to the original level by use of equation (5.23), where $\widetilde{F}_i$ is the predicted value on the log-transformed data.

$$F_i = e^{\widetilde{F}_i} - 1 \tag{5.23}$$

5.5.2 Seasonality

Nominal variables such as month numbers do not represent a numerical quantity which makes the absolute value meaningless (Bijma, Jonker, and Van der Vaart, 2013, pp. 261–262). One-hot-encoding (OHE) can be used to replace the numerical quantity to a list of binary dummy variables. Each nominal variable, having S possible states, translates to $S-1$ binary variables. The S^{th} variable follows from a combination of the $S-1$ variables (e.g. all binary variables being 0). Each binary variable is defined by equation (5.24), where M defines the state space having only one categorical value in the case of OHE. u_i defines the value of the original variable (Makridakis, Wheelwright, and Hyndman, 1997, pp. 269–274; Harvey, 1993, pp. 160–163).

$$x_i = 1_M(u_i) = \begin{cases} 1 & \text{if } u_i \in M \\ 0 & \text{if } u_i \notin M \end{cases} \tag{5.24}$$

Section 4 showed a strong intra-week pattern, with a decreasing demand pattern from Monday to Sunday. The use of one single variable containing the day of the week (1-7) would not suffice, due to its non-linear relationship with the quantity as shown in figure 4.7. Therefore, the following six dummy variables are introduced according to equation (5.24).

$$\begin{aligned} D_1 &= 1_{Monday}(u_i) \\ D_2 &= 1_{Tuesday}(u_i) \\ &\vdots \\ D_6 &= 1_{Saturday}(u_i) \end{aligned} \tag{5.25}$$

The same methodology has been applied for the modeling of the month within the year and the week within the year. This results in an additional $(12-1)+(53-1) = 63$ binary variables. These variables are called $M_1, M_2, ...M_{11}$ and $W_1, W_2, ...W_{52}$ respectively.

5.5.3 Introductions

In a majority of cases, the introduction period heavily deviates from the remaining time series as shown in figure 4.8. A binary dummy variable is added according to definition (5.24), where M contains the first year of the time series.

5.5.4 Holiday Effect

The previously conducted data analysis showed the relative holiday effect where some days result in a completely closed business, having demand equal to zero. Contrary, some holidays result in a reduced demand bigger than zero. Lastly, demand before or after a specific holiday can deviate due to the closure of a holiday. This effect is for example strongly visible with new year, resulting in an increased sales on the second of January.

Most holidays are reoccurring events on a fixed date such as new year and liberation day. Other days such as Easter are depending on the specific year, where most other days such as Ascension day can be derived from the Easter date.

Therefore, seventeen holiday dates are incorporated and marked by a dummy variable. For a detailed overview of the used holidays and the corresponding date calculation, consult appendix B.2. Additionally, variables indicating days before or after a holiday are given dummy variables as well, to capture a shift in demand due to business closure.

5.5.5 Lag Variables

Lag variables describe the sales of a previous period or a previous moment in time. The first type is added by use of the sum over a two week period, indicating the total sales during the two weeks. The second indicates the sales at a specific date, for example 28 days ago.

5.6 Parameter Estimation and Feature Selection

Estimation of the parameter vector $\boldsymbol{\beta}$ for both linear regression and Generalised regression, is generally performed using the Maximum-Likelihood estimator (MLE), where $\boldsymbol{X}$ should be of full rank to make it identifiable, that is $Column\ Rank(\boldsymbol{X}) = p$.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y}) \tag{5.26}$$

One of the alternatives is the use of the Least Absolute Shrinkage and Selection Operator, in short LASSO. This method minimizes the sum of squared errors by estimating $\boldsymbol{\beta}$. Furthermore, it performs feature selection alongside the minimization, as it reduces some coefficients towards zero (Bühlmann and Van de Geert, 2011). The advantage obtained from LASSO is a reduced complexity of the model which can reduce the variance without a large increase in bias. Furthermore, it helps to reduce overfitting which is especially of use when the number of variables is large and the number of observations low. The used notation to define this method originate from Bühlmann and Van de Geert (2011).

Lasso makes use of the penalty parameter λ to shrink some estimated values of $\boldsymbol{\beta}$ to zero. For linear regression, estimation of $\boldsymbol{\beta}$ under the shrinkage parameter λ, is performed using equation (5.27).

$$\hat{\boldsymbol{\beta}}(\lambda) = argmin_{\beta}\left(\frac{\sum_{i=1}^{n}(Y_i - (\boldsymbol{X}\boldsymbol{\beta})_i)^2}{n} + \lambda \cdot \sum_{j=1}^{n}|\beta_j|\right) \tag{5.27}$$

The shrinkage parameter λ needs to be estimated, where a smaller value for λ decreases the number of parameters incorporated in the model. From this, the most important parameters are the ones first entering the model as λ increases. Estimation of λ is done by use of k-fold cross-validation on the train data.

5.7 Evaluation Metrics

The selection of the correct forecasting evaluation measure depends largely on the underlying demand type. Some comparisons between the most common evaluation measures will be made in this section, to find the best suited evaluation metrics. We define the following parameters with time step t ranging from 1 to n.

$$\begin{aligned} Y_{i,t} &:= \text{Observation at time } t \text{ for time series } i & t \in [1, ..., n] \\ \hat{Y}_{i,t} &:= \text{Forecast at time } t \text{ for time series } i & t \in [1, ..., n] \\ e_{i,t} &:= \text{Error at time } t \text{ for time series } i & t \in [1, ..., n] \end{aligned}$$

Throughout this research paper, it is assumed that overestimating demand/sales is equivalently important as underestimating the demand. From this, the absolute forecasting error will be used, where the absolute error is defined in equation (5.28).

$$|e_{i,t}| = |Y_{i,t} - \hat{Y}_{i,t}| \tag{5.28}$$

Scale-dependent error measures such as the MAE, the MSE and the RMSE are traditionally widely applied. However, these methods suffer from difference in scale between time series and are therefore only applicable for comparison of different forecasting methods on the same data, rather than comparing performance across time series (Hyndman and Koehler, 2006; Ragnerstam, 2015).

To allow for comparison of performance across time series, the percentage error measures are introduced. These measures are independent of scale and based on the APE (equation (5.29)). The MAPE is one of the most widely known percentage error measure, taking the average over the APE for each time t.

$$APE_{i,t} = |100 \cdot \frac{e_{i,t}}{Y_{i,t}}| \tag{5.29}$$

$$MAPE_i = \frac{1}{n} \sum_{t=1}^{n} APE_{i,t} \tag{5.30}$$

It is explained in chapter 4 that a large proportion of products follows a lumpy demand pattern. The MAPE is not able to deal with zero demand, as the fraction $\frac{e_{i,t}}{Y_{i,t}}$ could result in division by zero if $Y_{i,t} = 0$ or being undefined when $Y_{i,t} = \hat{Y_{i,t}} = 0$. Furthermore, Chase (1995) explained that the MAPE could result in unfair measurement as it is skewed if $Y_{i,t}$ gets close to zero. Therefore, the WAPE should be preferred. The WAPE has the advantage that it accounts for the total demand, which weights extreme values against the actual demand volume (Ragnerstam, 2015). The WAPE is presented in equation (5.31) and will be used to evaluate the performance between SKUs.

$$WAPE_i = \frac{\sum_{t=1}^{n} \left(APE_{i,t} \cdot Y_{i,t} \right)}{\sum_{t=1}^{n} Y_{i,t}} = \frac{\sum_{t=1}^{n} |e_{i,t}|}{\sum_{t=1}^{n} Y_{i,t}} \tag{5.31}$$

The MASE, on the other hand, is also frequently used when the MAPE is not applicable, due to the presence of division by zero (Hyndman and Koehler, 2006; Davydenko and Fildes, 2013). The MASE was proposed by Hyndman and Koehler

(2006), being independent of scale by focusing on a scaled error instead of an absolute error while avoiding division by zero. Another advantage of this method is that it compares directly with one-step ahead forecasting, where MASE < 1 indicates on average smaller errors than the one step ahead forecasting (Hyndman and Koehler, 2006).

However, the drawback of this method is that it puts more weights to time series which are comparatively stable due to its difference between time t and $t - m$ (Ma, Fildes, and Huang, 2016). The MASE is defined in equation (5.32).

$$MASE_i = \frac{1}{n} \sum_{t=1}^{n} \frac{|e_t|}{\frac{1}{n-m} \sum_{t=m+1}^{n} |Y_{i,t} - Y_{i,t-m}|} \tag{5.32}$$

Both the WAPE and the MASE are not commonly used in business perspectives but are strong measures for comparison between SKUs and will therefore be used throughout this paper. To overcome the lack of business understanding, the MAE will be reported in addition to these two measures and is presented in equation (5.33).

$$MAE_i = \frac{1}{n} \sum_{t=1}^{n} |e_{i,t}| \tag{5.33}$$

5.8 Inventory Management

A detailed forecast is mostly not the final goal and primarily serving as an input variable in further decision making. It can, for example, be of financial importance to estimate expected costs or of importance in supply chain management to coordinate production and resources. In this section, we focus on the practical implementation for inventory management.

Tijms (2013, pp. 249–250) elaborates on a wide range of practical inventory models, making a distinction between deterministic inventory models and stochastic inventory models. The first has both the EOQ and the EPQ model where the EOQ model is the most widely known, assuming no lead-time and continuous order moments. The EPQ adds a production time of p to the model, bringing it closer to reality.

The stochastic inventory models mentioned by Tijms (2013, pp. 272–289) are listed below. All models assume stochastic demand and a lead-time L.

1. **(s,Q)**: Quantity Q ordered if inventory drops below s.
2. **(s,Q) With lost demand**: no back-orders.
3. **(R,S)**: After each R time units, inventory is replenished until S.
4. **(R, s, S)**: (R,s) but inventory only replenished if below s.

The listed models require a demand distribution, without the use of any forecasting information. Therefore, we will propose a small adaptation to use the forecast results as a foundation for the inventory management. Forecasting reduces the uncertainty of the demand, which in turn, reduces the required safety stock. Safety stock is the additional stock required to compensate for short term unpredictable fluctuations. Therefore the required stock R at time t can be defined as $R_t = \hat{F}_t + S_t(\alpha)$ where $\hat{F}_t$ is the forecasted amount for time t. $S_t(\alpha)$ is the required safety stock to meet service level α. The difference between our model

and the models listed by Tijms (2013, pp. 272–289) is that we replace the total demand distribution by the distribution of the errors, to calculate the safety stock. Afterwards, the total sales is added to the safety stock to obtain the required inventory. To calculate the order size and order moment, we define the following parameters.

$$\begin{aligned} L &:= \text{Lead time} \\ S_t(\alpha) &:= \text{Required safety stock for service level } \alpha \\ I_t &:= \text{Inventory available at the start of time t} \end{aligned}$$

The required stock at time t is only depending on the uncertainty of the forecast. For $L = 1$, the required safety stock can simply be found by solving $\alpha = 1 - \mathbb{P}(e \leq S)$ for the empirical error distribution.

$$S = P_{(1-\alpha)}(e) \tag{5.34}$$

For $L > 1$, the required stock (R_t) is is the sum of the forecast values over the lead-time plus the α-quantile of the joined error distribution e_L.

$$R_L = \sum_{i=1}^{L} F_i + e_{L,(1-\alpha)} \tag{5.35}$$

The joined error distribution (equation (5.36)), indicates that the error increases by a factor L when taking the convolution over L i.i.d Normally distributed variables:

$$\mathbb{P}(e_L > S) = \mathbb{P}(e_{t+1} + e_{t+2} + ... + e_{t+L} > S) \tag{5.36}$$

For normal distributions, the sum of L distributions (called Z) is defined by:

$$Z \sim N\Big(\sum_i \mu_i \ , \ \sum_i \sigma_i^2 \Big). \tag{5.37}$$

The total required order quantity at time t is then defined by:

$$O_t = I_t - R_t = \sum_{i=t}^{t+L} F_i + e_{L,(1-\alpha)} \tag{5.38}$$

Figure (5.3) shows the need for inventory management on top of point-forecasting. An inventory management based on point-forecasting results in an error of approximately 50 percent. Higher service level requirements result therefore in higher safety stock, where the drawback is that the required safety stock increases exponentially as the service level increases. Companies generally strive to achieve service levels between 95 and 99 percent, resulting in large inventory costs.

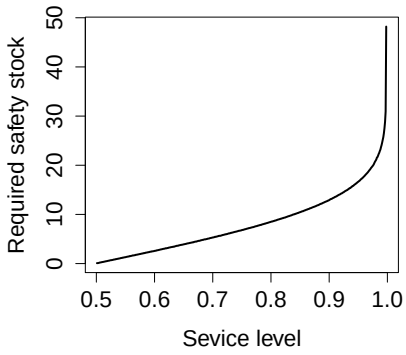

FIGURE 5.3: Influence of service level requirements on the safety stock for $\mu = 0$ and $\sigma = 10$.

Chapter 6

Results

The results are obtained by programming the described forecasting models in Python. The forecasting models are executed on a local machine. The HAC is programmed in Python as well, where a cloud server was used to obtain the distance matrix within a reasonable time frame.

6.1 Forecast Model Performance (Method I)

Figure 6.1 draws the comparison between the different forecasting models for the MAE, WAPE and MASE. It is visible that, although there is no comparison between time series, the MAE is difficult to compare due to its skewed boxplots.

It can be seen by comparing the WAPE and the MASE, that all regression models outperform both benchmark models by far. Lasso linear regression and Logistic linear regression are performing approximately equivalently in terms on the median. Poisson linear regression performs significantly worse, based on the median test ($p < 0.01$), compared with Lasso linear regression and Logistic linear regression.

The logistic linear regression model shows a broader interval of performance for both the WAPE and the MASE, compared with Lasso linear regression. Therefore, the Lasso linear regression seems to perform more stable. It can be remarked that the Naive last year forecast has a MASE of exactly one, which corresponds with the given definition in section 5.7.

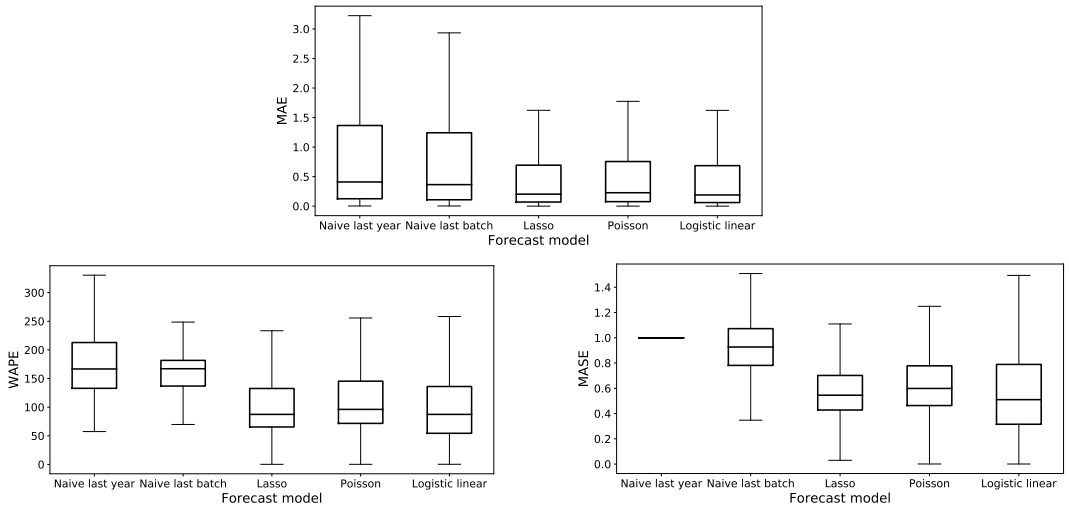

FIGURE 6.1: Comparison between the different forecast models by use of MAE (top), WAPE (left bottom) and MASE (right bottom).

Figure 6.2 shows the forecast result for the test data of exactly one year, for Lasso linear regression on a single product. The figure shows that the weekly sales pattern as well as the yearly seasonality is captured by the model. The year trend is not visible as only the test data is displayed. However, the trend is captured by the model as no systematic error is found in the residuals of the test data.

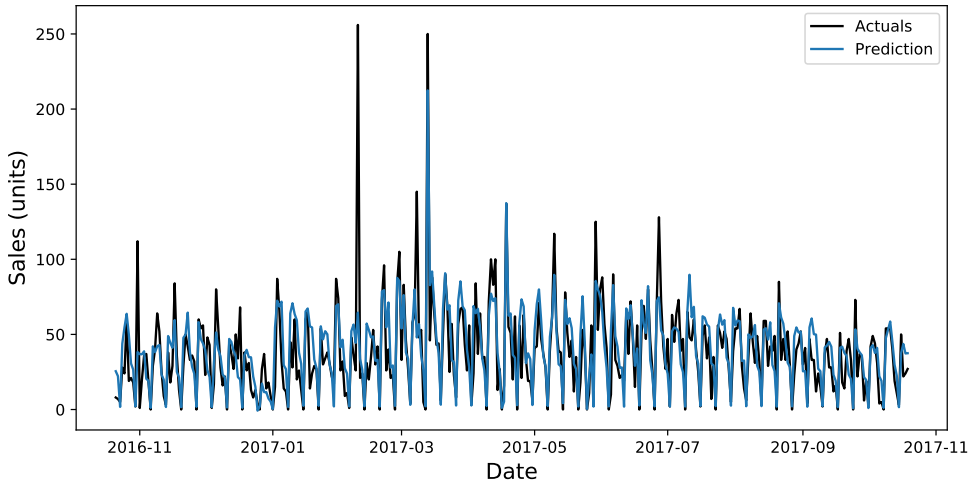

FIGURE 6.2: Forecast compared with the actuals over the test data, using Lasso linear regression and method I for a single product.

6.2 Clustering

Figure 6.3 shows the within sum of squares for the identification of the optimal number of clusters. Figure 6.4 shows the final clustering where the optimal number of clusters ($k = 63$) is used.

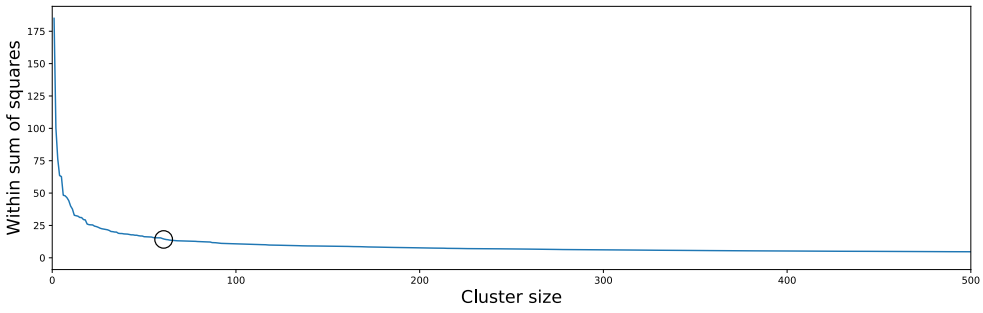

FIGURE 6.3: Within cluster sum of squares with an identified number of clusters of $k = 63$.

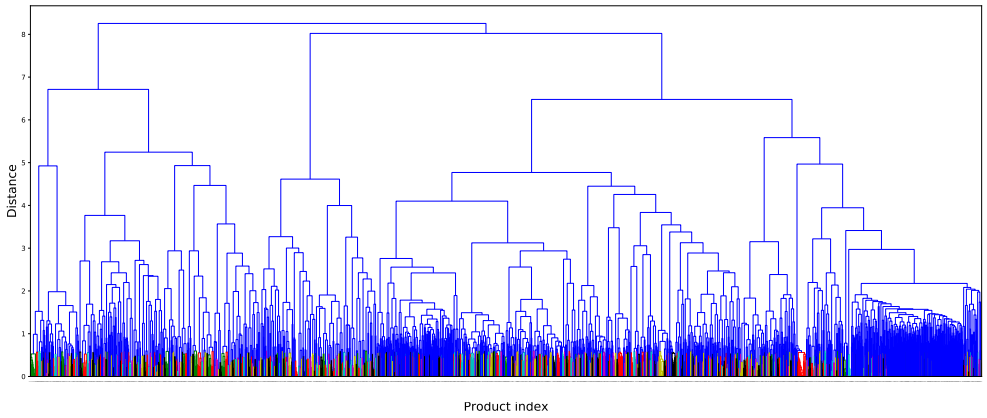

FIGURE 6.4: Dendogram of the final clustering for $n \approx 3000$.

6.3 Forecast Method Comparison

Figure 6.5 compares the three methods by use of the MASE. All three methods make use of Lasso Linear regression for the forecast. The methods are performing quite equally when considering the median over the approximately 3000 products. Method III performs approximately 0.9 percent lower compared with method I in terms of the median. Method II performs approximately 1.4 percent higher than method I. According to the median test, only a significant difference exists between the median of method II and method III ($p = 0.023$).

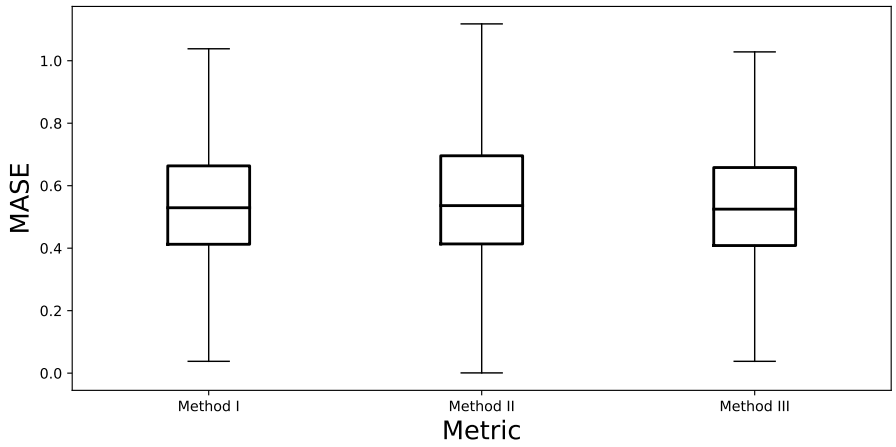

FIGURE 6.5: Boxplot comparison in terms of the MASE between the three methods over $n \approx 3000$.

Figure 6.6-6.8 make a comparison between all three forecasting methods for respectively the zero fraction, the time series variation and the time series length. This comparison is made to compare the performance for the three method accross different time series. The previous boxplot showed that there is only a small difference between the methods, when comparing the median over all time series at once. However, when considering the different types of time series, a more clear performance distinction can be created.

Figure 6.6 shows that method I performs best for time series with a relatively low fraction of zero demand. Additionally, method II has a dramatic increase in the MASE for the same type of time series. Both method II and method III outperform method I for time series with a large fraction of zero observations. This result gives rise to the idea that the use of product groups or clusters is more beneficial, when the time series contain a large number of observations without demand (intermittent demand).

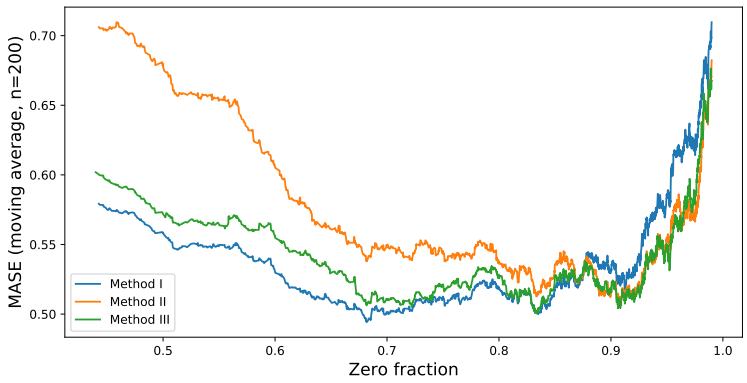

FIGURE 6.6: Influence of the fraction of zero demand within a time series on the MASE for all three methods.

When considering the coefficient of variation to classify the time series, a similar distinction can be made. Figure 6.7 shows that method III outperforms the other methods when the variation exceeds approximately 3.6. Both method II and III outperform method I when the variation grows large and the other way around when the variation is relatively low.

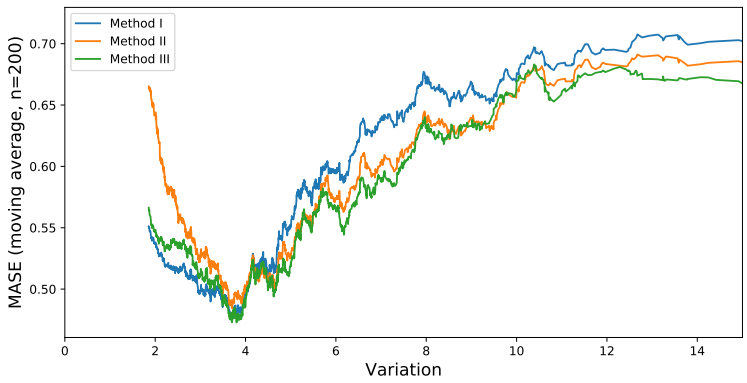

FIGURE 6.7: Influence of the time series variation on the MASE for all three methods.

Figure 6.8 shows a relative similar performance for short time series between the three models. Method I seems to outperform the others on the middle length and method III seems to dominate on the long time series above approximately 2200 days. Counter intuitively is the MASE growing when the history increases. A similar pattern exists when considering the WAPE.

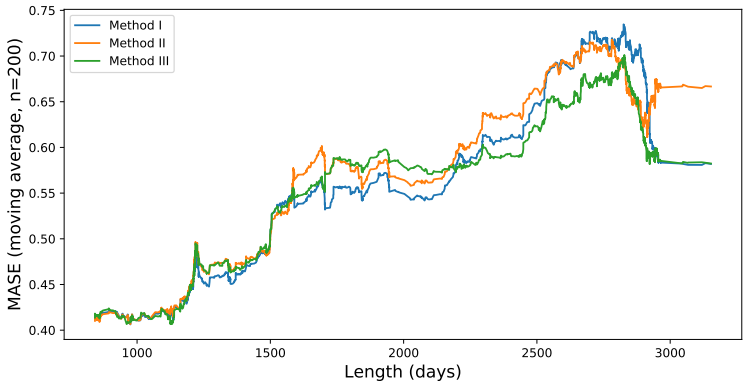

FIGURE 6.8: Influence of the time series length on the MASE for all three methods.

The following tables show a summary of the above discovered trends which can be of guidance when selecting the appropriate forecasting method for a set of time series.

Zero fraction	**<0.85**	**>0.85**
Method	I	III

TABLE 6.1: Preferred method based on time series zero fraction.

Variation	**<3.6**	**>3.6**
Method	I	III

TABLE 6.2: Preferred method based on time series variation.

Length	**<1700**	**1700-2200**	**>2200**
Method	-	I	III

TABLE 6.3: Preferred method based on time series length.

Chapter 7

Conclusion

This study has shown that high variable sales data is rather difficult to forecast, resulting in a high WAPE by default. Comparing the performance of regression techniques with benchmark models such as the naive models, show that the regression techniques are able to derive a comparably good forecast which outperform the benchmark models by far.

The case study suggests that, although the generalization of these results is not yet known, there is no single method to use when considering forecasting. The presented results show that the predefined product groups can be dangerous when a low variation or a low zero fraction is present in the time series. The forecasting accuracy can be negatively influenced when using these groups for the wrong time series. Use of predefined product groups should therefore be carefully considered before being applied.

The results show that forecasting can benefit from forecasts at an aggregated level, when the aggregation is based on a clustering rather than a business grouping. However, this clustering method is no exception on the rule that there is no uniform forecasting method. The results show that cluster based forecasting can outperform the detailed method when the zero fraction is relatively high, the variation is high or the time series has a long history. Although clustering does not always influence the forecasting accuracy positively, it retains the accuracy drop relatively limited, when compared with the predefined product groups. Clustering is therefore a more robust solution then the predefined product groups when applied to the wrong time series.

There is an unexpected correlation between the time series length and the MASE. This behaviour might be explained by the nature of the dataset which contains all types of time series, having long historical series with a small amount of observations, making the MASE grow large.

In practice, one would want to obtain a fast forecast model which might even needs to be executed on a daily basis. The suggested clustering method requires the construction of a n x n upper matrix of dissimilarities between all time series. One can expect this operation to take up too much time when applied in practice as the size of the matrix grows exponentially along the number of time series. The changes in dissimilarity might not change fast for long existing products. Therefore, it might be sufficient to update this matrix less frequently compared with the forecast frequency.

7.1 Further Research

This research considers a short term forecasting horizon which might yield different results from a long term forecasting case study. Forecasting more steps ahead could be of major interest for businesses. Therefore, conducting this research for a larger forecasting horizon could be of interest, yielding potentially different result from the current research.

Besides the used forecasting models, there is still a large variety of potential models which could increase the overall accuracy. Other forecasting models could potentially outperform the currently used model. As this research has its main focus on the three forecasting methods, no full comparison is made between forecasting models. Therefore this research could be repeated with use of different forecasting models to identify if the found results can be generalized.

Although the used dataset contains a large variety of time series, it would be of major interest to repeat this study in other market segments. If the found results can be generalized towards other market segments such as retail markets is not yet clear.

This research did not incorporate time series with short historical information. Identifying if time series with short historical information could benefit from clustering, would be of major importance for companies as new product introductions can frequently occur.

This research identified the best clustering method by use of a more theoretical analysis. Further research could try to incorporate the effect of each cluster method on the forecasting accuracy, by use of a case study. A clearer performance analysis can then be conducted between the cluster methodology (method III) and the detailed forecasting method (method I).

A detailed analysis of the influence from the cluster size on the forecasting accuracy was out of scope for this research but is expected to largely influence the performance of the forecasting method. Therefore, a case study with respect to the cluster size influence on the performance differences between method I and method III would therefore result in a stronger comparison.

Lastly, this research makes some distinguishes between different time series. Further research could focus on a more clear separation in groups, to further investigate the performance per time series type.

Appendix A

Data Sample

Data sample with hashed product codes.

Date	Product	Quantity
2017-09-19	14546	100
2017-09-19	14528	1
2017-09-19	14887	5
2017-09-19	14525	2
2017-09-19	14534	2
2017-09-19	14648	100
2017-09-19	14546	10
2017-09-19	14887	2
2017-09-19	14525	2
2017-09-19	14534	2
2017-09-19	14648	1
2017-09-19	14546	10
2017-09-19	14545	1
2017-09-19	14525	2
2017-09-19	14534	2
2017-09-19	14648	1
2017-09-19	14546	10
2017-09-19	14545	1
2017-09-19	14528	2
2017-09-19	14605	1
2017-09-19	14728	1
2017-09-19	14748	2
2017-09-19	14545	1
2017-09-19	14545	1
2017-09-18	14868	1
2017-09-18	14866	1
2017-09-18	14645	1
2017-09-18	14748	4
2017-09-18	14568	1
2017-09-18	14646	1
2017-09-18	14887	1
2017-09-18	14525	3
2017-09-18	14527	2

Appendix B

Dummy Variables

B.1 Date Variables

The following date variables are created.

1. month
2. quarter
3. weekday
4. week of year
5. week of month
6. introduction year

B.2 Holidays

According to the Dutch system.

Holiday	**Calculation**
New year's Eve	<year>-12-31
New Year	<year>-12-01
First Christmas day	<year>-12-25
Second Christmas day	<year>-12-26
Kingsday	**if** year >2014: <year>-4-27 **else**: <year>-4-30 (-1 day if occurs on Sunday)
Liberation day	<year>-5-5
Valentine's day	<year>-2-14
'Sinterklaas'	<year>-2-14
Mothers day	second Sunday of may
Fathers day	third Sunday of June
Commemoration day	<year>-5-4
Three Kings' Day	first Sunday of January
Easter	Butcher's Algorithm
Easter Monday	Easter +1 day
Good Friday	Easter - 2 days
Ascension day	Easter + 29 days
Whit Sunday	Ascension day + 10 days
Whit Monday	White Sunday + 1 day

Appendix C

Dendogram Example

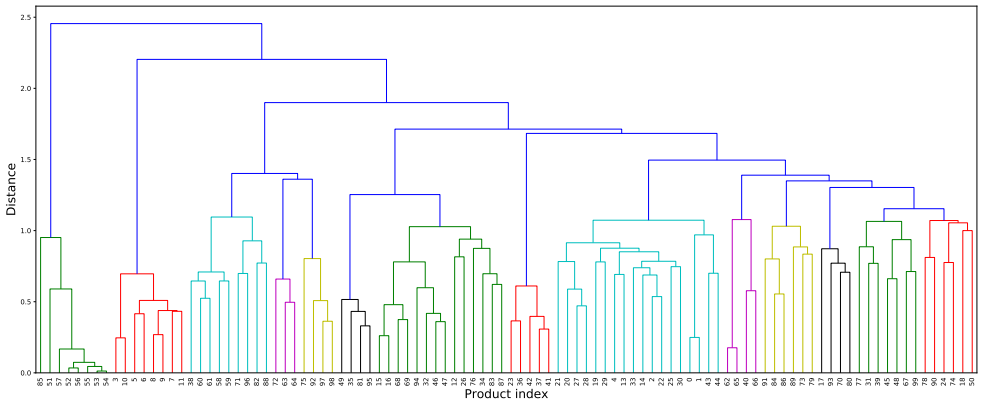

FIGURE C.1: Dendogram with sample size n=100.

Appendix D

Clustering sample

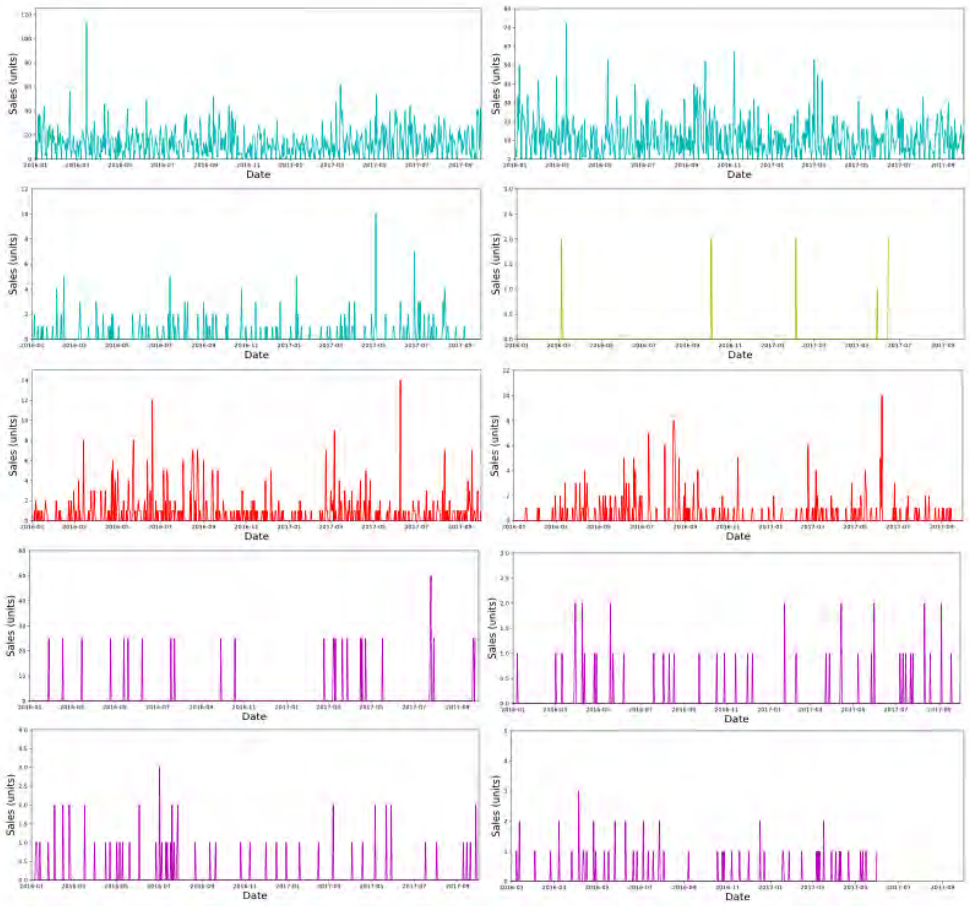

FIGURE D.1: Cluster sample for n=18 and k=5

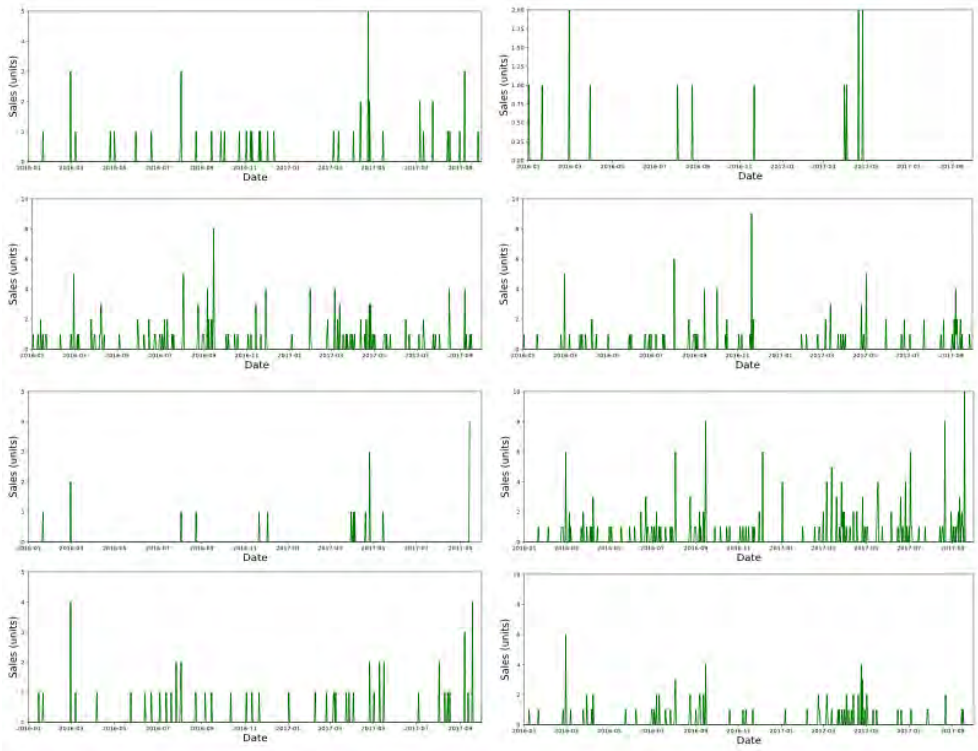

FIGURE D.2: Cluster sample for n=18 and k=5

Bibliography

Azzali, S. and M. Menenti (2000). "Mapping vegetation-soil-climate complexes in southern Africa using temporal Fourier analysis of NOAA-AVHRR NDVI data". In: *International Journal of Remote Sensing* 21.5, pp. 973–996. DOI: 10.1080/014311600210380.

Babai, M.Z, M.M. Ali, and N. Kourentzes (2012). "Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis". In: *Omega* 40.6, pp. 713–721. DOI: 10.1016/j.omega.2011.09.004.

Barnea, A. and J. Lakonishok (1980). "AN ANALYSIS OF THE USEFULNESS OF DISAGGREGATED ACCOUNTING DATA FOR FORECASTS OF CORPORATE PERFORMANCE". In: *Decision Sciences* 11.1, pp. 17–26. DOI: 10.1111/j.1540-5915.1980.tb01122.x.

Bergen, K.M. et al. (2005). "Change detection with heterogeneous data using ecoregional stratification, statistical summaries and a land allocation algorithm". In: *Remote Sensing of Environment* 97.4, pp. 434–446. DOI: 10.1016/j.rse.2005.03.016.

Bijma, F., M. Jonker, and A. Van der Vaart (2013). *Epsilon uitgaven 76: Inleiding in de statistiek (Dutch Edition)*. Utrecht: Epsilon Uitgaven. ISBN: 978-90-5041-135-6. URL: https://www.amazon.com/Epsilon-uitgaven-76-Inleiding-statistiek/dp/9050411355?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=9050411355.

Bühlmann, P. and S. Van de Geert (2011). *Statistics for High-Dimensional Data*. Berlin: Springer-Verlag GmbH. ISBN: 3642201911. URL: https://www.ebook.de/de/product/14667862/peter_buehlmann_sara_van_de_geer_statistics_for_high_dimensional_data.html.

Chase, C. (1995). "Measuring forecast accuracy". In: *Journal of business forecasting methods and systems* 14.3, pp. 134–143.

Chen, H. and J.E. Boylan (2008). "Empirical evidence on individual, group and shrinkage seasonal indices". In: *International Journal of Forecasting* 24.3, pp. 525–534. DOI: 10.1016/j.ijforecast.2008.02.005.

Davydenko, A. and R. Fildes (2013). "Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts". In: *International Journal of Forecasting* 29.3, pp. 510–522. DOI: 10.1016/j.ijforecast.2012.09.002.

Evans, J. and R. Geerken (2006). "Classifying rangeland vegetation type and coverage using a Fourier component based similarity measure". In: *Remote Sensing of Environment* 105.1, pp. 1–8. DOI: 10.1016/j.rse.2006.05.017.

Flach, P. (2018). *Machine Learning*. Glasgow: CAMBRIDGE UNIV PR. 410 pp. ISBN: 1107096391. URL: https://www.ebook.de/de/product/20035222/peter_flach_machine_learning.html.

Fliedner, G. (1999). "An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation". In:

Computers & Operations Research 26.10-11, pp. 1133–1149. DOI: 10.1016/s0305-0548(99)00017-9.

Gross, C.W. and J.E. Sohl (1990). "Disaggregation methods to expedite product line forecasting". In: *Journal of Forecasting* 9.3, pp. 233–254. DOI: 10.1002/for.3980090304.

Gunst, M.C.M de (2013). "Statistical Models". Vrije Universiteit Dictaat.

Harvey, A.C. (1993). *Time Series Models: 2nd Edition*. London: The MIT Press. ISBN: 978-0262082242. URL: https://www.amazon.com/Time-Models-Andrew-C-Harvey/dp/0262082241?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0262082241.

Hyndman, R.J. and A.B. Koehler (2006). "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22.4, pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

Jain, A.K., M.N. Murty, and P.J. Flynn (1999). "Data clustering: a review". In: *ACM Computing Surveys* 31.3, pp. 264–323. DOI: 10.1145/331499.331504.

Kedem, B. and K. Fokianos (2002). *Regression Models for Time Series Analysis*. 2nd ed. New Jersey: Wiley. ISBN: 0-471-36355-3.

Kourentzes, N., F. Petropoulos, and J.R. Trapero (2014). "Improving forecasting by estimating time series structural components across multiple frequencies". In: *International Journal of Forecasting* 30.2, pp. 291–302. DOI: 10.1016/j.ijforecast.2013.09.006.

Lhermitte, S. et al. (2011). "A comparison of time series similarity measures for classification and change detection of ecosystem dynamics". In: *Remote Sensing of Environment* 115.12, pp. 3129–3152. DOI: 10.1016/j.rse.2011.06.020.

Ma, S., R. Fildes, and T. Huang (2016). "Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information". In: *European Journal of Operational Research* 249.1, pp. 245–257. DOI: 10.1016/j.ejor.2015.08.029.

Makridakis, S., S.C. Wheelwright, and R.J. Hyndman (1997). *Forecasting: Methods and Applications*. New York: PAPERBACKSHOP UK IMPORT. 656 pp. ISBN: 0471532339. URL: https://www.ebook.de/de/product/3598915/spyros_g_makridakis_steven_c_wheelwright_rob_j_hyndman_forecasting_methods_and_applications.html.

Manning, C.D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. New York: Cambridge University Pr. 496 pp. ISBN: 0521865719. URL: https://www.ebook.de/de/product/7455223/christopher_d_manning_prabhakar_raghavan_hinrich_schuetze_introduction_to_information_retrieval.html.

Orcutt, G., H.W. Watts, and J.B. Edwards (1968). "Data aggregation and information loss". In: *American Economic Review* 58, pp. 773–787. URL: https://www.jstor.org/stable/1815532.

Ragnerstam, E. (2015). "How to calculate forecast accuracy for stocked items with a lumpy demand". MA thesis. School of Innovation, Design and Engineering.

Rostami-Tabar, B. et al. (2015). "Non-stationary demand forecasting by cross-sectional aggregation". In: *International Journal of Production Economics* 170, pp. 297–309. DOI: 10.1016/j.ijpe.2015.10.001.

Shou-Hsiung, C., C. Shyi-Ming, and J. Wen-Shan (2016). "Fuzzy time series forecasting based on fuzzy logical relationships and similarity measures". In: *Information Sciences* 327, pp. 272–287. DOI: 10.1016/j.ins.2015.08.024.

T. Tabar, Bahman (2013). "ARIMA demand forecasting by aggregation". Theses. Université Sciences et Technologies - Bordeaux I. URL: https://tel.archives-ouvertes.fr/tel-00980614.

Tijms, H. (2013). *Operationele Analyse. Een inleiding in modellen en methoden*. 4th ed. Amsterdam: Epsilon. ISBN: 978-50-5041-075-5.

Troncoso, A., M. Arias, and J.C. Riquelme (2015). "A multi-scale smoothing kernel for measuring time-series similarity". In: *Neurocomputing* 167, pp. 8–17. DOI: 10.1016/j.neucom.2014.08.099.

Zotteri, G. and M. Kalchschmidt (2007). "A model for selecting the appropriate level of aggregation in forecasting processes". In: *International Journal of Production Economics* 108.1-2, pp. 74–83. DOI: 10.1016/j.ijpe.2006.12.030.

Zotteri, G., M. Kalchschmidt, and F. Caniato (2005). "The impact of aggregation level on forecasting performance". In: *International Journal of Production Economics* 93-94, pp. 479–491. DOI: 10.1016/j.ijpe.2004.06.044.