Traditional Forecasting Applied to Retail Demand

Research Paper Business Analytics

Cor Woudt

Supervisor: Ger Koole

November 2018

Abstract

This paper elaborates on conventional traditional forecasting methods applied to retail demand. The methods are compared using several error metrics to evaluate the method which suits retail demand best. A new method is introduced in this paper, namely, the LEA forecasting method [10], which is currently still in development. The paper first introduces the forecasting methods, elaborates on their aspects and afterwards, a case study is carried out for comparison.

Contents

1 Structure

This paper is focused on the forecasting possibilities for retail demand. The structure of the paper will be as follows: First, an introduction into retail, long-tail demand and forecasting, which will be followed by an extensive overview of common forecasting methods. Afterwards, the main focus will be on the error metrics used for evaluation and the usability of the different methods in practise.

At the last section, a test case will be worked out to compare the described methods, to end with the results, the conclusion and a discussion and further research section.

2 Introduction

The problem addressed in this paper is related to retail demand and forecasting this specific type of demand. This paper is aimed to answer the following problem definition:

What is the best forecasting method to forecast retail demand, regarding a single forecasting method for all items?

To answer this question, there will be a case study regarding forecasting retail demand. In this part, an introduction will be given regarding retail, long tail assortment and forecasting.

2.1 Retail

Retail refers to the activity of reselling. A retailer can be referred to as any person or organization who re-sells goods or services in relatively small quantities directly to consumers or end-users.

In the past, most of retailing was performed through local grocery stores, traveling sales men and shopping malls, where nowadays, retail is not only performed through stores, but is also opening up at the internet, where most retailers have online shops to reach broader markets and a wider public.

Within the retail industry, there is a wide range of assortment. Most retail companies are selling a substantial amount of products, but not every product has the same turnover in terms of sales per week. Compared to products which are defining for a company or products which are sold quite often, certain products are just in the assortment to maintain a diverse assortment or a niche assortment, even though there is a low turnover.

Forecasting the whole assortment therefore can be difficult when there is a spread in behaviour of the sales pattern of the different products.

This brings us to the next point of interest, long tail assortment.

2.2 Long Tail

Long tail can be referred to as low turnover but significant assortment. The reason that retailers have this kind of assortment can vary from high margins to maintaining a big assortment or to must-have products within the branch. Multiple reasons can be discussed for having long tail assortment, but they all have one thing in common. Long-tail assortment can be logistical challenging, namely, forecasting this low turnover assortment. Forecasting can be challenging when there are a lot of zero sales realizations. These zero sales realizations are also referred to as intermittent demand. Forecasting is a critical part of retail. Within retail, the supply chain is the logistical connector of the operation. To maintain a smooth and efficient supply chain, many conditions must be met. One of these conditions is to estimate what the future demand will be and how to prepare for it.

2.3 Forecasting

To get estimates of the future and to prepare the supply chain, forecasting is necessary. Forecasting is the process of predicting or estimating future time series realizations, mostly based on extrapolation of past data and taking into account that events can occur which can influence the demand extremely, for example holidays.

The necessity of forecasting can be found in the next citation which is the mission statement of the Etos Supply Chain.

"The right item at the right place at the right time against the right costs" - Etos Supply chain

To get the right item available in the first place, there have to be an estimate of the future demand, the forecast.

3 Forecasting Methods

There are various methods of forecasting, ranging from averaging to smoothing and from linear to multidimensional extrapolation. Frequently, forecasting simply is to extrapolate the current state and momentum into the future. As Makridakis, which is known as a key figure of modern forecasting, describes [1], the most commonly used traditional forecasting methods are the following, described and explained below. The following notation will be used throughout the paper:

$$Y_t = \text{Realization at time } t \mid F_{t+h} = \text{Forecast for time } t+h \text{ made at time } t$$

3.1 Naive Forecasting

Naive forecasting is known in two ways, Naive 1 and Naive 2.

- Naive 1
  The most simple method, Naive 1, is to use $F_{t+1} = Y_t$, which means, the forecast for the next period is the realization of the present period.
- Naive 2
  A little more sophisticated is the Naive 2 Forecast. The second naïve method is based upon seasonality. The forecast is constructed the following way, $F_{t+1} = Y_{t+1-s}$, where s is the number of seasons in the data. For monthly data, $s = 12$, for weekly data, $s = 52$ and so on. The forecast in this case becomes the realization of the former period of that season.

The Naive forecasting methods are mostly used as reference or baseline method. The advantage is that both methods are extremely simple and intuitive. Therefore, these methods are explainable and computational cheap compared to more sophisticated methods.

3.2 Mean Forecasting

Another quite naive method, but more robust to variation and trend, is the mean forecasting method. The mean forecasting method simply averages all known realizations, and that will be the forecast for the next time period. Notation:

$$F_{t+1} = \frac{\sum_{i=1}^{t} Y_i}{t}$$

The mean forecast method is also rather simple and intuitive. In the case of the mean forecasting method, the computational task is slightly harder than the former two methods, but still computationally easy and cheap.

3.3 Moving Average of order k

The Moving Average of order k, also denoted by $MA(k)$, is also an averaging method. The past k realizations will be averaged and that will be the next forecast. Notation:

$$F_{t+1} = \frac{Y_{t-k+1} + \cdots + Y_t}{k}$$

$$or$$

$$F_{t+1} = F_t + \frac{1}{k}(Y_t - Y_{t-k+1})$$

What can be seen in the second equation is that the higher the number of periods taken into account, the more stable the forecast. The change in forecast will be affected by $\frac{1}{k}$ times the difference of the last used realization and most present realization. Moving Averages are quite stable, but are reactive forecasts. The Moving Average will always "follow" the actual realizations. However, compared to the Mean Forecasting method, the Moving Average is more responsive, by allowing at maximum k realizations, where the Mean Forecasting method will use all realizations to compute the forecast. The $MA(k)$ method is computationally a little easier than the Mean Forecasting method, due to the limited amount of realizations taken into account. This however, is only valid when the parameter k is determined. The optimization of the parameter can be difficult, but is mostly easier than optimizing smoothing parameters, which will be introduced later on. Advantages of $MA(k)$ are the weighting of several realizations. This allows the forecast to follow the trend and ensures the method does not overreact to a sudden outlier in the realizations. A disadvantage is that the method is always following the realizations just as the Mean Forecasting method and also will be slow in detecting a possible trend in the series. This is especially an issue when there is strong seasonality in the series.

3.4 Single Exponential Smoothing

One of the more popular methods is Single Exponential Smoothing, mainly referred to as SES. SES is based on the previous forecast and an adjustment for the forecast error of the previous forecast. Notation:

$$F_{t+1} = F_t + \alpha(Y_t - F_t) = \alpha Y_t + (1-\alpha)F_t$$

or

$$F_{t+1} = \alpha Y_t + (1-\alpha)\alpha Y_{t-1} + (1-\alpha)^2 F_{t-1}$$

SES is, just as the Moving Average method, based on the last realizations. However, the weights for each past realization decay exponentially, whereas the weights are equal at the Moving Average method. This allows the forecast to be more reactive regarding the actual realizations and thus be capable of reacting to a trend in the series.

SES is computationally a little harder than the former methods except for the $MA(k)$, due to the parameter tuning. Compared to $MA(k)$, the SES parameter is ranging between $[0, 1]$, where $MA(k)$ is ranging over the set of integer numbers. The parameter tuning must be done before the method is deployed. Although this process mainly occurs only before using the method in production, this has to be taken into account. Computing new forecasts is rather simple, as only the last forecast and last realization are taken into account. The foremost advantage of the SES is that it can react quite responsive, depending on the chosen α, whereas the former methods react much slower or don't react at all. The method also extrapolates its error of the former forecast onto the next forecast, which helps the method to be as reactive as possible.

3.5 Computational aspects

One of the reasons why these methods are commonly used, also in retail, is the fact that these forecasting methods require limited use of computational power. The computational power was back in the days not as equally wide available and powerful as nowadays. Imagine the need to forecast thousands of items on weekly basis, for thousands of stores, it would have been rather expensive to use computational challenging methods to derive the needed forecasts. Commonly known but more sophisticated and computational harder methods are described in the section below.

3.6 Holt's Linear Exponential Smoothing

Holt, as described in the following papers [5] [7], extended the single exponential smoothing to deal with trend data. This method is also known as Double Exponential Smoothing(DES), notation:

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + b_{t-1})$$

or

$$L_t = \alpha Y_t + (1-\alpha)(F_{t-1})$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

$$F_{t+h} = L_t + h \times b_t$$

The first component is an estimate of the level part. One can notice the same underlying logic as was used with the Single Exponential Smoothing method when $h = 1$.

The second component is an estimate of the trend. This addition provides better estimates for future realizations due to the fact that the momentum of the series is empowered by an extra factor. This method is computationally harder than SES, by updating one more equation each time, using four different components instead of two. Also, a bit more work compared to SES is concerned with the fact that Holt's method requires to estimate the parameter β next to α.

Holt's linear exponential smoothing has the advantage that the trend is extrapolated, whereas the detrended series is forecasted by single exponential smoothing. This is an application of decomposition, where both series are forecasted comparable with single exponential smoothing. The disadvantage of this method is that it is lacking a seasonal component. The seasonal fluctuations will be seen as a trend or as the regular series.

3.7 Holt Winters' Trend and Seasonal Smoothing

The smoothing and averaging methods described before can deal quite well with non-seasonal data. However, there are numerous items which show seasonal behaviour. To be able to forecast these type of items, a method with respect to the seasonal component is required. Holt Winters' Trend and Seasonal Smoothing [8] is a method which deals with this type of data. The method is also known as Triple Exponential Smoothing. Notation:

$$L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1})$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

$$S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-s}$$

$$F_{t+m} = L_t + b_t m + S_{t-s+m}$$

This method is not only capable of handling a trend next to the level component, but can now also handle a seasonal component.

This method is again computationally harder than SES, but also harder than Holt's linear method, by updating one more equation each time, using five different components instead of two. There is also more work concerned compared to both previous exponential methods due to the the fact that there are now three parameters to be estimated and optimized.

The advantage of Holt Winters' method is the fact that it can deal with seasonal data, as well as trend or regular data. However, Holt Winters' method requires a large data-set time-wise. The initialization of the method requires two years of data and an additional third year to verify the chosen parameters. When a retailer starts a company, most of the time there is no three years of data available. This makes it hard to use the Holt Winters' method in practise, especially for new items. During this case study, the parameters will be estimated using the available data.

3.8 Croston's

A method specified for intermittent or sporadic demand is called the Croston's Method [3]. Croston developed a method based on the Single Exponential Smoothing concept, however, the method differs on the aspect of forecasting not only demand but also inter-demand intervals. This method is mainly used for forecasting spare parts, but might be applicable in the field of retail demand forecasting regarding the long tail assortment. Slightly different notations will be used for Croston's method. These notations are considered by Rob Hyndman [11].

- Y_t denotes the demand during time t
- X_t denotes an indicator which is 1 if there was demand in period t
- j_t denotes the number of periods with non zero demand up until time t, i.e., $j_t = \sum_{i=1}^{t} X_t$, j_t will be referred to as j
- Y_j^* denotes the size of the jth non-zero demand and Q_j the inter-arrival time between Y_{j-1}^* and Y_j^*, so we can write $Y_t = X_t Y_j^*$, where j is referred to as the number of periods with non zero demand up until time t. This relationship is simply the indicator if there was demand times the associated demand for time t itself.

Using the SES concept, Croston's method constructs Z_j and P_j, which represent the forecast for the $(j+1)$th demand size and inter-arrival time. Notation for both:

$$Z_j = (1-\alpha)Z_{j-1} + \alpha Y_j^*$$

$$P_j = (1-\beta)P_{j-1} + \beta Q_j$$

With α and β the smoothing parameters. Let $l = j_t$ denote the last period of demand. Then the mean demand rate, which is used as the 1-step ahead forecast for the demand at time $t+1$, is estimated by the following ratio:

$$\hat{Y}_{t+1} = \frac{Z_l}{P_l}$$

$$F_{t+1} = \frac{Z_l}{P_l}$$

Croston was the first to decompose the forecast into two separate parts and forecast both parts independently. However, several drawbacks have arisen since. Firstly, the forecast is biased, as shown by Syntetos & Boylan

in their paper [4]. The bias arises from the fact that the inter demand time and the demand itself is not independent.

Secondly, the initial method made the weak assumption that both zero demand interval and the demand itself could be smoothed with the same parameter, which in this case study is addressed by using two parameters to optimize the method.

3.9 Decomposition Method

Classical decomposition assumes that the time series, a period of sales in this case, is following a pattern, which can be extracted, and an additional error and level part, denoted by E and L. This pattern is usually constructed by multiple components. The trend component T and the seasonal component S. One important notion here is that there can be two methods of decomposing data. The pattern can be multiplicative or additive. Forecasters are mostly using the multiplicative method. Notation:

Additive

$$Y_t = L_t + S_t + T_t + E_t$$

Multiplicative

$$Y_t = L_t \times S_t \times T_t \times E_t$$

Additive vs Multiplicative

$$S_t = Y_t - Y_{t-s} \text{ vs } S_t = \frac{Y_t}{Y_{t-s}}$$
$$T_t = Y_t - Y_{t-1} \text{ vs } T_t = \frac{Y_t}{Y_{t-1}}$$
$$L_t = Y_t - S_t - T_t \text{ vs } L_t = \frac{Y_t}{S_t \times T_t}$$

The trend can be obtained by subtracting the realization of $t-1$ of the realization of t. The same goes for the seasonal aspect, by subtracting realization $t-s$ of realization t. This is called differencing. The use of differencing shows the pattern for the trend and seasonal aspect. Removing the trend and season from the realizations leaves the level component. The level part is sometimes referred to as noise, depending on the nature of the series. In this case study, the level part is the steady sales pattern.

Forecasting via decomposition is a time consuming manner, due to the decomposition itself, where afterwards a component-wise forecast will be produced.

Decomposition is mainly focused on separating and forecasting the different components all on their own, and assemble them back to one forecast. Decomposition can be very useful if the underlying series are behaving significantly different. This separation makes it easier to understand the underlying series and its behavior. The disadvantage of decomposition is the time required to decompose properly and to detect whether the multiplicative or the additive method is required.

3.10 LEA forecasting

In this case study, a derivation of the decomposition is used. The LEA forecasting method is inspired on spline smoothing, where one tries to minimize the inflection points and the error, developed by [10]. There by, the method considers week and trend components, which are assigned via Linear Programming. The objective is shown below:

$$\min \alpha \sum D_{2_t} + (1-\alpha) \sum e_t$$

With D_{2_t} the second difference of the trend for realization t and e_t the error for realization t and α the smoothing parameter to be optimized.

The second difference is the used part to minimize the inflection points. The second difference is the trend of the trend so to say. Mathematically, it is a method to estimate the second derivative.

This method is still in development but will be used during the test case.

The unique part of this method is that it minimizes the inflection points, which is an uncommon way to prevent the method from over-fitting the training data. The h step ahead forecast is constructed the following way:

$$F_{t+h} = S_{t+h} + T_t + h \times D_t$$

Where S_{t+h} is the weekly component for time $t+h$, which is bounded by $[1, \cdots, 52]$ and thus if $t+h > 52, S_{t+h} = S_{t+h-52}$. Moreover, T_t is the trend component at time t and D_t is the first difference at time t. This first difference is used to estimate the future, by multiplying the current difference by the steps to be forecasted.

3.11 Further knowledge

Makridakis [1], has been organizing forecasting competitions to evaluate the practical use of forecasting methods. The M-competitions have shown that simple forecasting methods do as well or better than the more sophisticated methods [2]. Forecasting methods which work on paper and which are scientifically suited best, are not a guarantee for success. The M-competitions point out that indeed not always the theoretical best method is also the best choice in practise.

The M-competitions are mainly focused on yearly, quarterly and monthly data, whereas this paper is aimed at weekly data. Thereby, the M-competitions are dealing with non-zero sales only, where the periods are at least containing sales of 30 or more, where this paper is also concerning the intermittent demand in particular.

Several methods considered in the M-competitions are therefore not relevant for this case study.

All the methods mentioned in the above subsections are mainly suited best for low variable series, where often, the series are fluctuating around a strict positive number, like sales for example.

Nevertheless, for low rotating assortment, as long tail assortment for instance, these methods are not quite applicable. The traditional methods can be used but are expected to not cope well with zero sales. These irregular and infrequent sales, or intermittent or sporadic demand, are difficult to forecast due to the randomness in the inter demand period next to the usual randomness in the demand itself.

4 Evaluation

Regarding forecasts, there is one thing which must be kept in mind during the forecasting, namely, forecasts are almost always wrong. This is important to keep in mind and not just mop up the method due to a wrong forecast. Nevertheless, to be able to make meaningful conclusions about the methods described in the former sections, error metrics will be used to evaluate the performance of these methods.

The error metrics are generally divided into four categories, which are described below. [6]

- Scale dependent metrics
  As the name suggests, the scale dependent metrics accuracy are related to the scale of the series. Examples of scale dependent metrics are the mean absolute error and mean squared error. Scale dependent metrics are easy to understand and good for single series comparison, but due to its scale dependency, it is not useful for comparing against different series.

- Percentage error metrics
  Percentage error metrics gives an impression of the error compared to the series itself, which is a scaled metric. An example of a Percentage error metric is the Mean Absolute Percentage Error. The scale independent character is an advantage compared to the scale dependent metrics, but are not capable of handling zero sales series well.

- Relative error metrics
  Relative error metrics compares the error with the error of a baseline method, such as the naive 1 forecasting method. An example is the mean relative absolute error. For relative error metrics, the same advantages and disadvantages come as with the percentage error metric, with the division by zero problem.

- Scale free error metrics
  Scale free error metrics expresses each error as a ratio compared to the average error from a baseline method, such as the naive 1 method. The biggest difference is that this method is scale free and the relative error metric is not due to the single comparison against the average comparison. An example for the scale free error metric is the mean absolute scaled error. The scale free metrics are a solution to both scale dependencies and intermittent series, as the method is capable of handling zero sales series.

The error metrics used in this case study are the following:

- Mean Absolute Percentage Error (MAPE)
  The MAPE error metric is a widely used method for evaluating forecasts. However, the MAPE metric is not able to cope with zero sales series.

$$MAPE = \frac{100\%}{N} \sum_{t=1}^{N} \frac{\mid Y_t - F_t \mid}{Y_t}$$

  The MAPE is sometimes favoured over the Mean Squared Error, due to the high penalty for big errors. Sometimes, some items are more interesting for the business compared with others. For this instance, the

following adaption can be used:

$$WMAPE = \frac{100\%}{N} \sum_{t=1}^{N} \frac{w \cdot \mid Y_t - F_t \mid}{w \cdot Y_t}$$

The method takes the weighted average of all the errors relative to the weighted actual sales. This can be of particular value when one favours certain items above others, think of relevance for the business or gross profit. Setting $w = 1$ in the equation will lead to the above mentioned MAPE, which will be used in this case study. The usage of WMAPE is used when certain items for instance are more relevant to the business. The w will than increase the error weighting towards the important items to obtain insights into methods with a special regard to the important items. The range of w can be anything, as long as the $w > 0$ to prevent a division by zero.

- Mean Squared Error (MSE)
  The MSE is the average of all the squared errors, denoted by:

$$MSE = \frac{1}{N} \sum_{t=1}^{N} (Y_t - F_t)^2$$

The MSE is a well known measure and also widely used for forecasting, but the method is sensitive to outliers. Due to the square, outliers will result in a high MSE, while the method could perform quite well over the rest of forecasts. The MSE is a metric which, compared to the MASE, is capable at handling zero sales series.

- Mean Absolute Scaled Error (MASE)
  The MASE is an adjusted MAPE, developed by forecasting specialists [17].

$$MASE = \frac{1}{N} \sum_{t=1}^{N} \mid q_t \mid$$

$$q_t = \frac{Y_t - F_t}{\frac{1}{n-1} \sum_{i=2}^{n} \mid Y_t - Y_{t-1} \mid}$$

Rewritten, it can be seen as:

$$MASE = \frac{MAE}{MAE_{insample,naive_1}}$$

$$MAE = \frac{1}{N} \sum_{t=1}^{N} \mid Y_t - F_t \mid$$

With the MAE the mean absolute error. In the numerator, the MAE with respect to the given forecast is used and in the denominator, the MAE with respect to the Naive 1 method is used.

When the MASE metric is smaller than one, the used forecasting method is performing better than the benchmark method, i.e., Naive 1.

The MASE metric has an advantage compared to the MAPE, where the MASE is only dividing by zero if the complete series is zero. When a series only has zero sales, there is no point in making forecasts at all.

5 Case

This case study is considering retail demand data and Poisson generated data. The data is originating from 10 stores and one fictive store, composed via an aggregation of the 10 stores to model a more stable series. For these 11 different stores, 10 items are chosen, ranging from sporadic to high rotating demand. The test case will be worked out in python [13] and R [12], where the forecasting will be implemented in Python, and the evaluation is further worked out in R.

5.1 Data

The forecasting methods will be tested on sales data originated from Etos [9]. The sales time series will cover several practical cases like store-item sales, aggregated store-item sales, event driven sales and intermittent or high rotating sales. Due to this wide range of different behaving series, the methods will be tested on several

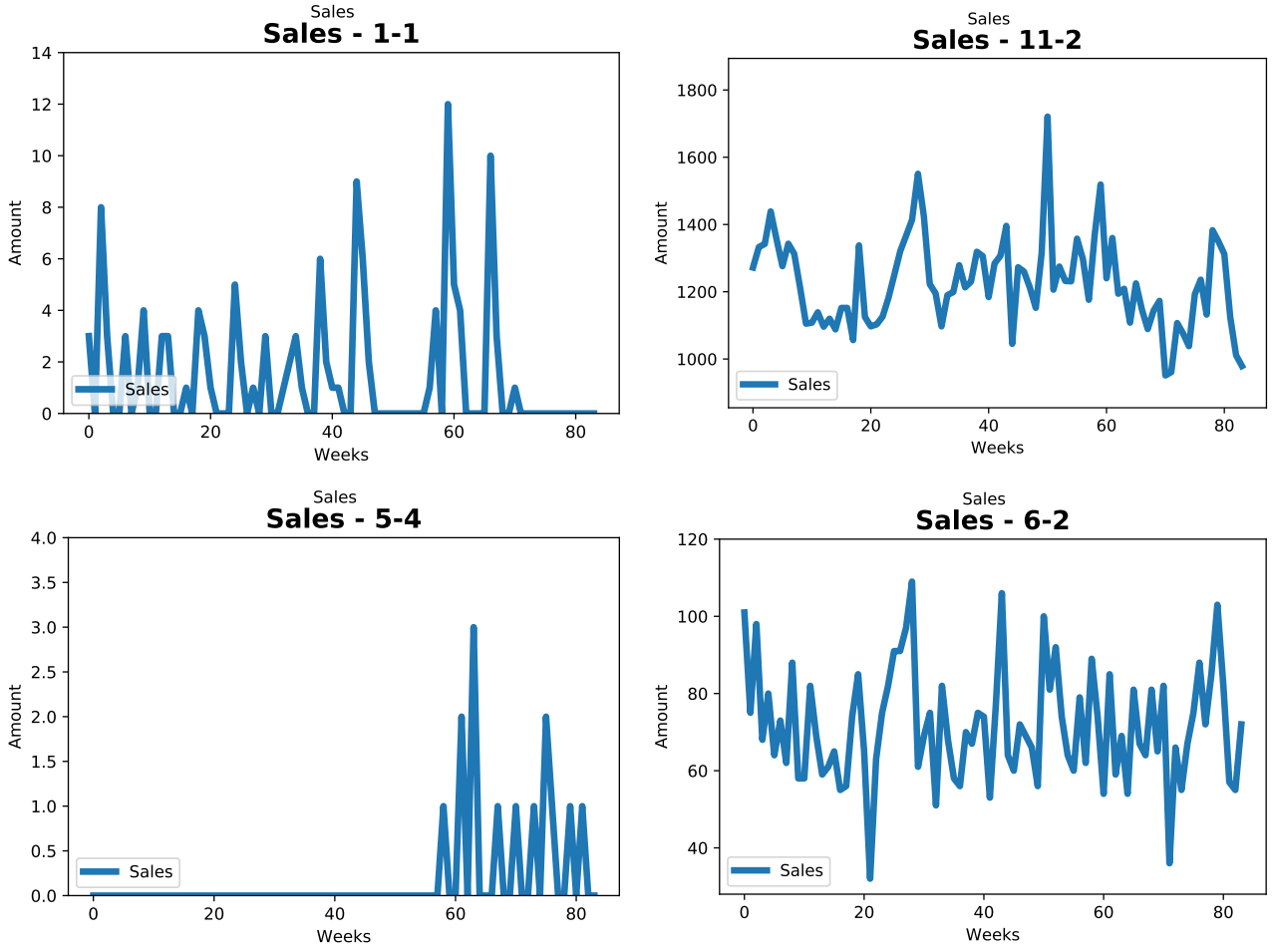
aspects, which will give more practical information about the methods and show their robustness. The data consists of 84 weeks of sales for several series.

Due to the fact that the case data consists of only about one and a half year of data, the expectation is that both Holt Winters' and the LEA-algorithm might have trouble initializing the settings regarding the seasonal aspects. The normal initialization period for seasonal data is at least two years and the data used for this case only contains an initialization period of one year.

Next to the real world data, Poisson distributed data will be generated with the same length as the original data, to make a more scientific comparison between the chosen methods and their parameter settings. The chosen Poisson series are modeled with the following means: $0.1, 0.25, 0.5, 1, 5, 10, 25, 50, 100, 250\&500$.

The Poisson distribution is chosen as it is used quite often to simulate the customer arrival process.

Four series are shown below, providing some feeling for the data to cope with.



The series are all behaving different, which will provide some interesting insights. Especially series $5-4$ is interesting, as the series is representing a new item, so this will be harder to forecast based on the past information as there is almost no data to rely on.

The other series are showing intermittent demand$(1-1)$ and frequent demand$(6-2,11-2)$. The latter series is looking to behave like a seasonal series, which is on of the parts this case study also addresses, although the estimates are quite fragile due to the small amount of available data.

5.2 Parameters

To ensure that all methods are tested to their fullest, a parameter grid is used for comparison of all the possible combinations. The defined parameters over all forecasting methods are $\alpha, \beta\&\gamma$ and are setup in a grid ranging from $\{0.1, 0.2, \cdots, 1\}$.

The methods will produce a forecast for the latter part of the series. This to let the methods learn from one year of data and start forecasting from there on.

The forecasting of the SES with $\alpha = 1$ and the mean forecasting of step size 1 are both neglected due to their complete similarity with the Naive 1 method.

6 Results

The forecasted series are compared using the three error metrics described earlier to evaluate the performance of the forecasting methods. As explained before, the MAPE metric is not capable of handling zero sales and is thus of less practical value, due to the high number of occurrences of zero sales. This sounds unpractical, but is the consequence of the choice to forecast retail demand series on store level. Nevertheless, the comparison can show the practical usability of the different methods.

The performance is measured per metric, where for each series the best performing methods are considered. The top three will be shown with their score. The score is the percentage of series it was the best forecasting method for. As mentioned, the MAPE metric has only measured a few series due to the zero sales realizations.

The results will be addressed per metric and per test set. The sets that have been tested are the following. The first set, set "All", is the set with all series, including the intermittent and even completely zero series.

The second set, set "74", contains all sets which have at least one realization bigger than zero.

The last set, set "18", is the set with only the series which have each realization bigger than zero.

The set names represent the number of series in the set.

Table 1: Results

Error Metric	**MASE**			**MSE**			**MAPE**		
Covering Set	**All**	**74**	**18**	**All**	**74**	**18**	**All**	**74**	**18**
Croston's Method	*58,97%*	*58,11%*	*38,89%*	*10,38%*	*14,86%*	*22,22%*	*27,78%*	*27,78%*	*27,78%*
Mean Method	*11,54%*	*12,16%*	*33,33%*	*37,74%*	*48,65%*	*44,44%*	*33,33%*	*33,33%*	*33,33%*
Single Exp Smh	*11,54%*	*12,16%*	*27,78%*	*21,70%*	*31,08%*	*22,22%*	*27,78%*	*27,78%*	*27,78%*
Naive 2	*10,26%*	*10,81%*	-	*0,94%*	-	-	-	-	-
Naive 1	*3,85%*	*4,05%*	-	*26,42%*	*1,35%*	-	-	-	-
Moving Average	*1,28%*	*1,35%*	-	*2,83%*	*4,05%*	*11,11%*	*5,56%*	*5,56%*	*5,56%*
Double Exp Smh	*1,28%*	*1,35%*	-	-	-	-	*5,56%*	*5,56%*	*5,56%*
LEA Algorithm	*1,28%*	-	-	-	-	-	-	-	-

Table 1 shows the performance of the tested methods per metric. The best performing methods are underlined for each set and each metric. The performance is measured as a percentage, where the number represents the amount of times the method is the best method. Dashed results means that the method has not been able to be one of the best methods for that set and error metric.

The results are interesting. Croston's method is performing equally compared to the Single Exponential Smoothing method. It is only better when the zero sales only series are considered, as can be seen by looking at the column "All".

Also remarkable is the fact that the Naive 1 and 2 methods are present in the best performing methods. However, one can see that when the set size decreases and the intermittent part is neglected, both methods are not present anymore. The Naive 2 method is the only method suited to handle the seasonality, which is present in a small matter in this case study. Originally, Holt Winter's(Triple Exponential Smoothing) was expected to perform better regarding seasonal series, but the method requires more data to make sure the initialization is done properly and the method is sufficient ready to produce high quality forecasts. The LEA-algorithm also relies on its initialization, which is not informative enough in this case study to obtain more relevant forecasts. The LEA-algorithm also has the ability to handle seasonality well, but is performing less then the Naive 2 method. This is the result of the fact that the LEA-algorithm also extrapolates the trend part, which can be to reactive for the series causing it to produce more outlier like forecasts, which are not an issue for the Naive 2 method. The Naive 2 method is performing well given the absence of big trend changes.

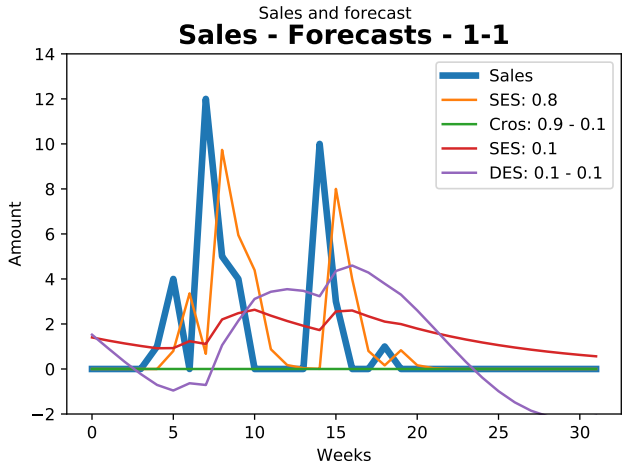
Croston's method and the Mean method are the best methods considering the MASE and MSE metric. The fact that the mean method is performing next to the more sophisticated methods like Croston's and Single Exponential Smoothing suggests that the series are quite stable.

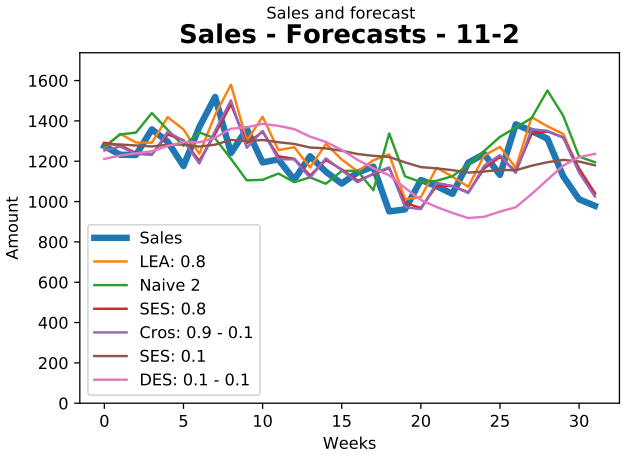
Looking at the forecasts on series level, there is, as expected, a high similarity between Croston's and Single Exponential Smoothing. The fact that Croston's method is making an adjustment for the inter-demand time, is the main reason for the difference in performance.

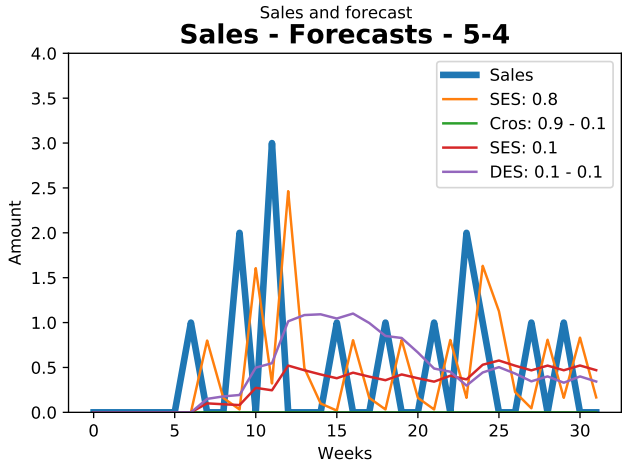
The Double Exponential Smoothing is performing less than the Single Exponential Smoothing, which can be explained due to the fact that there are a lot of zero sales periods, making it hard to maintain a stable trend forecast. As shown by the MAPE metric, the Double Exponential Smoothing is performing better when there is stable non zero demand.

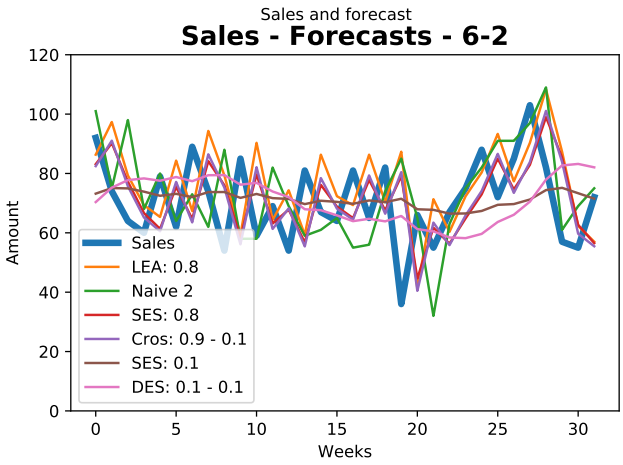
As mentioned earlier, there were aggregated series used to determine the behaviour on a higher level. These series tend to behave more stable.

A few examples of the forecasts are shown below, the same four series as shown at chapter five.









In the figures it can clearly be seen that the Naive 2 method is indeed performing well when there is an returning pattern in the data, but the trend is quite stable. Furthermore, the performance of the Naive 1, LEA algorithm and the SES with a high alpha are showing similar behaviour, following the realizations.

The performance of the Mean forecasting is mostly related to the steady series without to much trend movement. This makes it harder for the more sophisticated methods, which are aiming to address all possible components of the series and therefore interpret certain movements wrongly as seasonality or trend. The robustness of the simple methods is the key for the performance, next to the absence of more past data to suit the more advanced methods.

7 Conclusion

This case study has shown that sales data is difficult to forecast with one method. The fact that (intermittent)retail demand can be unstable makes it hard to come up with a good solution. There are always series which behave different than the operating window of the chosen method. Therefore, by making choices about which method to use in practise, I suggest to use a flexible method which suites the biggest part of the assortment to be forecasted.

If the chosen method is good in general, there are always workarounds to come up with in practise, for the ill behaving series. This way, the focus can be shifted from finding the perfect method to focusing on handling the ill behaving series.

Also shown by this case study, both Holt Winters'(TES) and the LEA-algorithm are performing not as good as they could. The problem lies outside the reach of these algorithms as the data to be forecasted lacks enough past realizations to obtain a qualitative initialization, making these methods perform worse than these methods actually can. The fact that there is less data than needed for a proper initialization makes it hard for both methods to come up with good seasonal estimates and make the clear distinction between seasonal and trend movements. The methods are now categorizing movements as seasonality while it might be a trend difference or just noise. The fact that the methods are not capable of handling these past data lacking series makes them under performing for these series, and might suggest that the methods are not applicable for items with a short life cycle.

As shown by the case study, coping with intermittent demand is difficult, as most methods are just converging to a zero forecast, which is useless in practice. Arguably the best method to cope with this intermittent demand is to just use an average and aggregate these averages to come up with a more stable series. The aggregation step is interesting for supply chain forecasting. When the store level is too intermittent, there can be more practical value in the aggregated series. The total demand is the demand to satisfy via the distribution centers, so this can be used one on one, as a replacement of the store level forecast.

The case study might suggest to use Mean forecasting as best forecasting method. This method is static and can therefore not adapt to new changes in the series. The ideal method would be one which is robust, working with some kind of lower- and upper bound, but still is able to extrapolate changes in the series pattern. The suggestion would be Single Exponential Smoothing, which can be enriched to make sure the method is not sensitive to the shown cases at this case study. The method on itself would be suitable, enriched with certain bounds and cleaning of past realizations to maintain a flexible but robust operating window.

8 Further Research

One of the important parts that have not been treated properly in this case study is more extensive research into seasonality and the time required for methods to adjust for this. The methods aimed at handling seasonal series, such as Holt Winters'(TES) and the LEA-Algorithm, are having difficulties due to the lack of sufficient initialization data. This troubles the potential of the models and makes the method perform worse than they eventually could. To make sure that these methods can be used for short life cycle items can be investigated, as how there might be adaptions for these algorithms to cope with less available data. This can be an interesting topic for further research as one might be able to obtain a more robust version of the LEA-algorithm.

An interesting part, which has been consciously not considered this case study, is the long term behaviour of the forecasting methods. The methods are now considered for a one step ahead forecast. Forecasting multiple steps ahead leads to more uncertainty, but is of more interest regarding the practical application of the methods and can be tested as an extension. Forecasts can then be used as estimates for future realizations, which do make the uncertainty bigger. However, long term forecasts can be useful and the uncertainty can be bounded by confidence intervals, maintaining a meaningful forecast.

Besides the described method, there are more forecasting methods nowadays. A linear combination of basis functions can be seen as an extension to regression, maintaining the property that there always will be optimal parameter settings due to a bounded solution space and a differentiable error metric. The use of linear basis functions and forecasting via the Bayesian framework are both not addressed in this paper. There might be a chance that these methods are coming up more as forecasting methods for ordinary retail demand, whereas these methods are now mainly used within the Machine Learning world. These methods are still in the phase where there is a lot of research but only few to none practical cases known towards forecasting demand and usage is computationally hard due to the lack of appropriate implementations and available frameworks.

The Bayesian framework is aimed at drawing conclusions about the data based upon the known observations and using that to make a prediction, based on probabilities [14] [15].

Linear basis functions are a lot more interpretive as they are simply a linear combination of several (linear)functions. These functions can be all from the same family of functions [16] but can also be from a wider range of functions to open up possibilities to better fit the data and the underlying distribution.

As shown in [18], where the main focus is on forecasting time series throughout machine learning methods, one can see that the Gaussian Process, which is relying on the Bayesian framework as well, is performing almost as good as the Neural Networks and other advanced methods in the Machine Learning field.

This is a nice illustration that advanced and complex models are not a guarantee for success. Keeping methods simple and explainable makes it easier to interpret the results and allow changes in the process, which are understandable and predictive in terms of the expected change and behaviour of the algorithm.

"It is our job to solve problems, not to brag about how shiny our tools are. Shiny tools are nice and all, but it often distracts away from the problem, this is bad. First try the simple thing before considering the complex thing. This is a reasonable approach to anything. A simple trick tends to work for another reason, it forces you to think about the problem. One can be tempted to assume a fancy algorithm to solve all problems, this assumption is a dangerous lie." [19]

As stated in the quote above, there is no necessary need for advanced and complex algorithms. Considering a simple method can provide insights and can always be used as a benchmark for a more advanced and complex algorithm.

References

[1] Makridakis, S., Wheelwright, S. & Hyndman, R. (1998). *Forecasting : Methods and applications* (3rd ed.). New York: John Wiley.

[2] Makridakis, S., & Hibon, M. (2000). *The m3-Competition: Results, conclusions and implications.* International Journal of Forecasting, 16(4), 451-476.

[3] Croston, J. (1972). *Forecasting and stock control for intermittent demands.* Operational Research Quarterly (1970-1977), 23(3), 289-303.

[4] Syntetos, A., & Boylan, J. (2001). *On the bias of intermittent demand estimates.* International Journal of Production Economics, 71(1-3), 457-466.

[5] Holt, C. C. (2004). *Forecasting seasonals and trends by exponentially weighted moving averages.* International Journal of Forecasting, 20, 5 – 10.

[6] Hyndman, R. J. *Another look at forecasting-accuracy metrics for intermittend demand* FORESIGHT Issue 4, June 2006.

[7] Holt, C. C. (2004). *Author's retrospective on Forecasting seasonals and trends by exponentially weighted moving averages.* International Journal of Forecasting, 20, 11 – 13.

[8] Winters, P. R. (1960). *Forecasting sales by exponentially weighted moving averages.* Management Science, 6, 324 – 342.

[9] Etos B.V. Retrieved September 1st, 2018 from https://www.etos.nl/

[10] Koole & Li "The LEA forecasting algorithm", working paper

[11] Shenstone, L., & Hyndman, R. (2005). *Stochastic models underlying croston's method for intermittent demand forecasting.* Journal of Forecasting, 24(6), 389-402.

[12] R Core Team (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[13] Python Software Foundation. Python Language Reference, version 3.6. Available at http://www.python.org

[14] Angers, J., Biswas, A., & Maiti, R. (2017). *Bayesian forecasting for time series of categorical data.* Journal of Forecasting, 36(3), 217-229.

[15] Corberan-Vallet, A., Bermudez, J., & Vercher, E. (2013). *Bayesian forecasting of demand time-series data with zero values.* European Journal of Industrial Engineering, 7(6), 777-796. doi:10.1504/EJIE.2013.058394

[16] Hachino, T., & Kadirkamanathan, V. (2011). *Multiple gaussian process models for direct time series forecasting.* Ieej Transactions on Electrical and Electronic Engineering, 6(3), 245-252. doi:10.1002/tee.20651

[17] Hyndman, R., & Koehler, A. (2006). *Another look at measures of forecast accuracy.* International Journal of Forecasting, 22(4), 679-679.

[18] Ahmed, N., Atiya, A., Gayar, N., & El-Shishiny, H. (2010). *An empirical comparison of machine learning models for time series forecasting.* Econometric Reviews, 29(5-6), 594-621.

[19] Warmerdam, V. (2018). Retrieved from https://www.linkedin.com/in/vincentwarmerdam/