System Analysis Laboratory
Helsinki University of Technology

Mat-2.108 Special assignment in applied mathematics

Comparison of Forecasting Algorithms in Spare Part Demand Planning

7th December 2006

Anssi Käki (58282L)
Department of Engineering Physics and Mathematics
1 Introduction

Forecasting is an essential part of decision-making field. There are several approaches to forecasting, depending on the subject of the forecast. In this text forecasting is considered as a part of demand planning. Demand planning is an important part of supply chain management (SCM), which is a growing section in operations research in both academic and business worlds.

A modern approach to SCM is presented in a book of Supply chain management and advanced planning (2004). In the book, Wagner (2004) divides demand planning into forecasting, simulation/what-if-analysis and safety-stock calculation. Furthermore, forecasting is divided into statistical forecasting, incorporation of judgemental factors and collaborative decision process. Using the same division, the focus of this text is in forecasting and particularly in statistical forecasting. The judgemental forecasting is not, however, ignored because it is an essential complement of statistical forecasting.

The statistical forecasts are inaccurate, especially when using "black box" techniques. The black box technique refers to approaches, where little or no knowledge of the subject of forecast is available and the data is handled as “anonymous”. That is often the case with demand forecasting. The unaccuracy (in general) stems from the fact that a precise forecast of a time series data demands a precise knowledge of the model behind the data. In the real world, where randomness is attended, such data sets do not exist and, therefore, the forecasting error is always present. Particularly demand patterns are impossible to recognise as there are numerous causal factors affecting the demand.

The approach to forecasting accuracy can be of theoretical nature but more approachable way is to use empirical methods, i.e. measure how different forecasts fit to data and how they manage to forecast the future. There is a large amount of empirical research on the measurement of statistical forecasting accuracy and also criticism of it (for more of this discussion see e.g. Makridakis 1984). In this study, the empirical methodology
plays a major role. It is because of complexity (or non-existence) of purely theoretical methodology and, on the other hand, because of the aim of the study: the examination of a Demand Manager (DM) module of certain statistical forecasting system. This module uses empirical means in order to find an optimal statistical forecasting algorithm and therefore the same kind of black box approach is suitable for the study. A detailed description of the research problem is considered next.

1.1 Problem definition

This study addresses questions posed by a project team of a Finnish-Swedish SCM-consulting company *ROCE Partners*\(^1\). The client of the company, a large Finnish high-tech firm, wanted to improve its spare-part supply chain operations. When this study was made, the project was moving from testing phase to implementation, and the project team sought to learn how to improve the statistical forecasting of spare-part demand. The general definition of the research problem can be put as: *How to create the most optimal component forecast?*

The project involves implementing a widely used advanced planning system (APS). The part to which this assignment contributes is the system’s DM, which produces demand forecasts from sales, install base and repair forecast data for numerous spare parts. The forecasts are mainly used in resource planning (short term forecasts), but also in strategic business units that operate with longer horizon (long term forecasts). The contribution of the study is to provide information about different forecasting algorithms, how to use them and how to improve the forecasting process in general.

For the assignment, real demand data was provided. In order to keep the scope of the study reasonable, only component demand and the demand type (“ABCD-classification”) data were used in the analysis. That is, different component characteristics were not used,

\(^1\)http://www.roce.com
although they could have been available. The data consisted of 100 time series divided to four demand types and the division was based on real distribution, i.e. the amount of series of type D was proportional to real proportion of D items. Also some documentation about the DM and its statistical forecasting functionality was provided.

The project team specified the research problem with following subproblems:

- What algorithm candidates should DM use when choosing the optimal forecast algorithm for various ABCD components (the so called Pick Best phase)?

- What is the best way to assess, which algorithm is the most accurate?

- Should the same algorithms be used for short and long term forecasts?

- How to choose the best parameters for e.g. the moving average algorithm?

- How much historical data should be used?

This report aims at giving answers or suggestions to previous questions and to forecast optimality in general. Both approaches, statistical and judgemental forecasting, are taken into account. The methodology of the both is presented in the next section along with the different error measures.

The evaluation of statistical algorithms is essential when producing forecasts based on time-series data. The Section 3 aims at comparing the algorithms by means of a literature review and tests for simulated and real demand data. The simulated data was used to get generic results for different algorithms using large sets of data. These results were then compared with real data test results in order to examine, whether the general results apply for the purposes of the project team.

The results are discussed in section 4. As the results show that statistical forecasting alone can be very inaccurate, some possibilities of using judgemental forecasting means are presented. The last section concludes the main findings and presents some suggestions for future study subjects.

3
2 Forecasting techniques and error measuring

In order to examine the forecasting system, the different forecasting concepts have to be defined. Here the forecasting framework is divided into three parts: judgemental forecasting, statistical forecasting algorithms and forecast error measuring. They provide the essential tools for forecasting algorithm evaluation and discussion of accuracy improvement possibilities.

2.1 Judgemental forecasting

"The key to effective demand management relies as much on human judgment and communication as it does on pure quantitatively based forecasting algorithms.”
- The Demand Manager Product Brief

Although this section does not answer all questions specifying the research problem in section 1.1, it still essentially relates to forecast improvement. It is intuitively somewhat clear that adding “expert information” to a forecast should improve (or at least not worsen) its accuracy. On the other hand, using purely judgemental forecasting is unpractical. This relates to the fact that automated forecasting systems like DM are planned to create and update forecasts for thousands of demands, a task which no human being is capable of doing. Also, as stated by Makridakis & Wheelwright (1989), the most of mathematical algorithms are unbiased. Judgemental forecasts can easily be biased by e.g. wrong assumptions or exaggerations.

An extensive study by Webby & O’Connor (1996) adds up the comparison of judgemental and statistical forecasting in empirical studies. In the literature, there is no clear evidence that either judgemental or statistical forecasts would outperform each other in general. Instead, there are some elements that can provide advantage to one or another. The study also discusses the combination of the two. Webby & O’Connor reach two major
conclusions:

- Especially the availability of contextual information (defined later) can make a judgemental forecast superior to a statistical one

- The major contribution of judgemental forecasting is the ability to integrate information not contained in the time series into a statistical forecast

2.1.1 Integration of judgemental and statistical forecasts

Contextual information is defined in Webby & O’Connor (1996) as “information, other than the time series and general experience, which helps in the explanation, interpretation and anticipation of time series behaviour”. Other descriptive words for contextual information are e.g. product knowledge or extra-model knowledge. Common examples of contextual information are upcoming promotion activities or inside knowledge of product replacements in the near future. Other sorts of subjective information are technical and causal knowledge, which can be expressed together as experience. Experienced forecaster can exploit e.g. the causality between company sales and spare part demand, the form of growth (exponential, logarithmic) of a new product or the reason for data discontinuities when adjusting the forecast. As stated before, the forecast automation (black box forecasting) is often needed as a base forecast. There are several ways to integrate judgemental information into statistical forecasts achieved by automated system.

Webby & O’Connor present four different integration alternatives: model building, forecast combination, judgemental adjustment and judgemental decomposition. An illustrative presentation of the four is presented in Figure 1. Basically, model building means that the user chooses the algorithm and its parameters, after which the chosen algorithm is used in forecasting. In combination approach, forecaster combines statistical and judgemental forecasts either in an objective manner (e.g. taking arithmetic mean of the two) or through some subjective reasoning. By judgemental adjustment is meant a statistical forecast which is tuned afterwards by the user in order to e.g. remove extreme values or
Figure 1: Integration of statistical and judgemental forecasts according to Webby & O’Connor (1996)

dampen the trend. The decomposition approach is somewhat similar to adjusting with a difference in timepoint when adjusting is made: in decomposition the contextual information is utilised (adjusting phase) before the statistical forecast is made. For more practical examples of the mentioned, see Section 4.

Wagner (2004) also suggests using similar techniques as above and in addition reminds of something very important, namely the feedback mechanisms. That is, the forecaster has to be aware of the effect of his judgemental contribution to the forecast. As simplest, the feedback mechanism can be implemented by comparing the statistical forecast with and without the judgemental input.

2.2 Statistical forecasting algorithms

In this study, the word algorithm is used when a way to calculate forecasts is meant. Many authors use the word method or sometimes model, which have a different meaning. Examples of these two would be the least squares forecasting method or random walk
with drift model. Here, the goal is to discuss about different algorithms that produce forecasts for unknown demand models.

Next is presented the principles of forecasting algorithms that are included in the DM. The symbol \( \hat{x}_t \) represents the forecasted value of \( x \) at time \( t \). The first three algorithms represent simple ones as the latter two are more complicated and might need exterior references (e.g. Makridakis & Wheelwright 1989) in order to be properly explained.

### 2.2.1 Moving average

The simple moving average (SMA) is often used as a reference algorithm in comparing the accuracy of more sophisticated methods. The only parameter the algorithm needs is the window length \( N \), being the amount of historical values taken into account in forecasting. The algorithm is defined as

\[
\hat{x}_{t+1} = \frac{x_t + x_{t-1} + \cdots + x_{t-N+1}}{N}.
\]

### 2.2.2 Simple exponential smoothing

Another simple and yet very common forecasting algorithm is the single exponential smoothing (SES). The idea behind SES is to use all observed values of the time series by taking a weighted average of the latest forecast (includes all the values till \( t-1 \)) and the latest observation (at time \( t \)). The weighting constant (or smoothing value) \( \alpha \in [0, 1] \) is chosen either subjectively or calculating the optimal value with the use of historical data (i.e. choosing \( \alpha \) that minimizes e.g. the mean square error of the forecast). The SES is defined as

\[
\hat{x}_{t+1} = \alpha x_t + (1-\alpha)\hat{x}_t
\]
The idea of SES lies also behind the next two algorithms. The first one reduces to SES if the series data is regular (i.e. non-intermittent) and the second one is an extension of the algorithm with trend and seasonal components included.

2.2.3 Modified Croston’s

Especially designed for the forecasting needs of intermittent demand, the method discovered by Croston (1972) takes advantage of both nonzero demand values and the interval between them when creating a forecast. A data set including consecutive zero demands and, when nonzero, random demand values, is considered as intermittent. In the implementation project, a data set for which 33% of the time periods consist of zero values and there are at least two consecutive periods with value zero, is classified as intermittent. The demand at time t is noted with \( x_t \) (and the forecast with \( \hat{x}_t \)) and the interval between two nonzero records \( p_t \) (\( \hat{p}_t \)). The most descriptive single quantity for demand forecasting in the case of intermittent data is the demand per period, namely \( \hat{Y}_t = \frac{\hat{x}_t}{\hat{p}_t} \).

If we denote the expected demand with \( \mu \) and expect the demand to occur with probability \( p \) (demand occurrence is Bernoulli distributed with parameter \( \frac{1}{p} \)), one would prefer using an unbiased forecasting algorithm for which \( E[\hat{Y}_t] = \frac{\mu}{p} \). Croston suggested using exponential smoothing to both demand and its inter-arrival time yielding the following algorithm, which is used only when nonzero demand appears:

\[
\hat{x}_{t+1} = \alpha x_t + (1-\alpha)\hat{x}_t \quad (3)
\]

\[
\hat{p}_{t+1} = \alpha p_t + (1-\alpha)\hat{p}_t \quad (4)
\]

\[
\hat{Y}_{t+1} = \frac{\hat{x}_{t+1}}{\hat{p}_{t+1}} \quad (5)
\]

In case of a zero demand, the forecast does not change, that is:

\[
\hat{Y}_t = \hat{Y}_{t-1}. \quad (6)
\]

Although the algorithm seems appealing, Syntetos & Boylan (2001) showed that, especially with large values of \( \alpha \), it yields biased results. Not going into details, the bias
is easily seen when one calculates $E[\hat{Y}] = E[\hat{x}]E[\frac{1}{\hat{p}}]$, which certainly is not $\frac{E[\hat{x}]}{E[\hat{p}]} = \frac{\mu}{\hat{p}}$.

Therefore, an unbiased version of the estimator (5) is more preferable. There are several ways to achieve (almost) unbiased version of Croston’s, yielding an algorithm called the modified Croston’s.

Based on $E[\hat{x}] \approx \frac{\mu}{\hat{p}} \frac{\hat{p} - \alpha - \alpha \hat{p}}{\hat{p} - \alpha}$, Teunter & Sani (2006) present the following, practically unbiased estimator:

$$\hat{Y}_t = \frac{\hat{x}_t (2\hat{p} - \alpha \hat{p})}{\hat{p}_t (2\hat{p} - \alpha)} = (1 - \frac{\alpha}{2}) \frac{\hat{x}_t}{\hat{p}_t - \frac{\alpha}{2}}. \quad (7)$$

The initialization of the algorithm is made by taking the average of all nonzero demands as $\hat{x}_0$ and the average time between the historical nonzero demands as $\hat{p}_0$. With this information, the one-step forecasts can now be calculated whenever new nonzero demand is observed with (3), (4) and (7). If a zero demand is observed, the forecast is achieved by (6). Forecasting more steps ahead is fairly simple: when forecasting demand, one simply uses the latest demand estimator $\hat{x}$ that should occur every $\hat{p}$th period. Demand per period $\hat{Y}$ is the constant ratio of the latest estimators $\frac{\hat{x}}{\hat{p}}$.

### 2.2.4 Triple Plus / Holt-Winter’s

The Holt-Winter’s linear and seasonal exponential smoothing is also known as triple exponential smoothing, which in order to keep the terminology analogous with the DM is called here the triple plus (3PLUS). The principle of 3PLUS is the same despite of the source, but the initialization and parameter update appearing in this text are taken from Meyr (2004).

The algorithm is based on three component model of a time-series, which comprises of level $(a)$, trend $(b)$ and seasonal factors $(c)$. Therefore, the one-step forecast is achieved by

$$x_{t+1} = (a_t + b_t)c_t. \quad (8)$$
Next we consider the estimation of the coefficients in (8) and, as new observations are made, the update and forecast computations.

To initialise the algorithm, the estimates of all components \((\hat{a}_0, \hat{b}_0\) and \(\hat{c}_0\)) have to be defined. The seasonal coefficients are determined with ratio-to-moving average decomposition, for which the moving average of the time series is calculated first. If the season is e.g. a week and data begins on Monday, the first moving average is calculated for Thursday, while it is the first day, when all days of the first week are included in the average. For Friday, the first Monday is replaced by the second Monday and so on (see table 1, column \(\bar{x}_7\) as an example). The minimum demand for data is two full seasons, so that the moving average can be computed for each day at least once. When the actual values are then divided with the moving averages, the proportional effect of a day (or a month etc.) can be obtained (see column \(\hat{c}\)). However, all data should be exploited and the seasonal coefficients are determined by taking the average of all timepoint coefficients (e.g. Thursday, column \(\hat{c}_i\)). Due to randomness, the coefficients may not sum up to full season length (e.g. seven for weeks), which can be normalised by multiplying the factors by the ratio of season length and sum of factors (column \(\hat{c}_i\)). After the calculations, one should have seasonal factor estimates \(\hat{c}_i\) to explain the seasonal behaviour of data.

The level and trend coefficients are achieved by means of linear regression. The data has to be deseasonalised first, which is simply achieved by \(\hat{d}_t = \frac{\hat{d}_t}{\hat{c}_i}\). Thereafter, the regression line is fitted and the intercept is taken as \(\hat{a}_0\) and the slope as \(\hat{b}_0\).

Putting all together, let us assume that the season length is \(T\) and the current time \(t = 0\). By using the initialization above, we should retain the numerical estimators for the current data level \(\hat{a}_0\), the trend factor \(\hat{b}_0\) and the seasonal components \(\hat{c}_{T+1}, \hat{c}_{T+2}, \ldots, \hat{c}_0\).

The 3PLUS applies the idea of SES to trend and seasonal components of a time series. As in SES, the weighting constant \(\alpha\) describes how much weight is put on the last observation vs. the previous forecast. Similarly, the constants \(\beta\) and \(\gamma\) weigh the trend and seasonal components. The choice of constants can be made by trying a set of different
Table 1: An example of calculating seasonal factors.

<table>
<thead>
<tr>
<th>Time</th>
<th>$x$</th>
<th>$\bar{x}_7$</th>
<th>$\hat{c}$</th>
<th>$\hat{c}_i$</th>
<th>$\hat{c}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>12.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue</td>
<td>14.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wed</td>
<td>18.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu</td>
<td>22.0</td>
<td>17.71</td>
<td>1.242</td>
<td>1.196</td>
<td>1.195</td>
</tr>
<tr>
<td>Fri</td>
<td>25.0</td>
<td>17.86</td>
<td>1.400</td>
<td>1.400</td>
<td>1.400</td>
</tr>
<tr>
<td>Sat</td>
<td>28.0</td>
<td>18.00</td>
<td>1.556</td>
<td>1.556</td>
<td>1.555</td>
</tr>
<tr>
<td>Sun</td>
<td>5.0</td>
<td>18.29</td>
<td>0.273</td>
<td>0.273</td>
<td>0.273</td>
</tr>
<tr>
<td>Mon</td>
<td>13.0</td>
<td>18.29</td>
<td>0.711</td>
<td>0.711</td>
<td>0.711</td>
</tr>
<tr>
<td>Tue</td>
<td>15.0</td>
<td>18.43</td>
<td>0.814</td>
<td>0.814</td>
<td>0.814</td>
</tr>
<tr>
<td>Wed</td>
<td>20.0</td>
<td>19.00</td>
<td>1.053</td>
<td>1.053</td>
<td>1.052</td>
</tr>
<tr>
<td>Thu</td>
<td>22.0</td>
<td>19.14</td>
<td>1.149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fri</td>
<td>26.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sat</td>
<td>32.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sun</td>
<td>6.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
values and choosing which fits the historical data best or by a subjective-basis procedure. However, after the initialization has been made and weighting constants established, the one-step forecasts can be obtained with (8) by replacing \( a_t, b_t \) and \( c_i \) with \( \hat{a}_t, \hat{b}_t \) and \( \hat{c}_i \). The seasonal coefficient is chosen to match the forecasted time point, i.e., when forecasting Monday, one uses the seasonal coefficient of Monday. Formally it is obtained with \( i + 1 = (t \mod T) + 1 \). The update of coefficients is made as follows:

\[
\hat{a}_t = \frac{a_t x_t}{\hat{c}_i} + (1 - \alpha)(\hat{a}_{t-1} + \hat{b}_{t-1}) \\
\hat{b}_t = \beta(\hat{a}_t - \hat{a}_{t-1}) + (1 - \beta)\hat{b}_{t-1} \\
\hat{c}_i = \gamma \frac{x_t}{\hat{a}_t} + (1 - \gamma)\hat{c}_{i-1}.
\]

(9) (10) (11)

After the updates are made, the next one-step forecast is calculated again with equation (8). Forecasting \( k \) periods ahead is done with the newest estimates of \( \hat{a}, \hat{b} \) and \( \hat{c} \):

\[
\hat{x}_{t+k} = (\hat{a}_t + \hat{b}_t k)\hat{c}_i.
\]

(12)

### 2.2.5 Bayesian forecasting

Bayesian forecasting offers a dynamic approach to time series forecasting. Here we present the idea and progress of the algorithm. The adjusting of the Bayesian algorithm in the DM is restricted, so a detailed consideration of Bayesian models is not made in this study, while the results would not be usable in the DM.

Bayesian forecasting consists of the assumed distribution of observed time-series and dynamical update of the distribution parameters when the amount of observations increase. The distribution of the system before an observation is called a *prior* (or prior). After a new observation is made, the distribution is updated to a *posterior* (or posterior). At the initialisation phase, the forecaster chooses an initial distribution of the data. This could be e.g. normally distributed demand with a linear growth of 5% and a variance of 1%. After the first observation, the initial distribution is updated to a posterior one. In the next phase, the posterior becomes the new prior, which is then updated again.
Figure 2: The system state forecasting and updating according to Welch & Bishop (2004) as a new observation appears. The idea of Bayesian forecasting relates to a state space model of a time series. The system state at time $k$ is marked with $Y_k$ ($Y \in \mathbb{R}^{n \times k}$) and the observation with $z_k$ ($z \in \mathbb{R}^k$).

The key between the stochastic model of a time series and the optimal forecasting (in a least squared error of a state-space model sense) is the Kalman filter. The comprehensive introduction of the Kalman filter by Welch & Bishop (2004) offers a straightforward algorithm to the usage of the discrete version, which is presented next.

A discrete time linear process can be expressed as a stochastic difference equation system

\begin{align}
Y_k &= AY_{k-1} + BU_{k-1} + w_{k-1} \\
z_k &= HY_{k-1} + v_{k-1},
\end{align}

where $A$ represents the state transition matrix, $B$ relates the control input $U$ to the state $Y$ and matrix $H$ links the observation to the state. The randomness is included in both equation by the white noise of $w \sim N(0,Q)$ and $v \sim N(0,R)$. $Q$ is called the process noise covariance matrix and $R$ the measurement noise covariance matrix.
After the equations (13) and (14) are set to represent the wanted system, the system state can be updated after each measurement with the Kalman filter. Without going into details, the updating computations are presented in figure 2.

The DM can apply Bayesian forecasting to linear level and trend models. The level model refers to a model where the observation fluctuates around its true level, that is: $z_t = x_t + \varepsilon_t$, where $\varepsilon_t$ is the white noise at time $t$, $\varepsilon_t \sim N(0, \sigma)$. A more usable model includes an additive trend component, meaning that the current level is updated from the previous level adding a trend factor ($\beta$) and white noise, i.e. $x_t = x_{t-1} + \beta_t + \gamma_t$. Also the trend is assumed to be a random walk, so $\beta_t = \beta_{t-1} + \delta_t$ ($\gamma_t$ and $\delta_t$ are defined as $\varepsilon_t$ above). Comparing this presentation to equations (13) and (14), the following equation system is achieved:

$$
Y_{t+1} = AY_t + W
$$
$$
z_{t+1} = Hx_{t+1} + V,
$$

where

$$
Y_t = \begin{pmatrix} x_t \\ \beta_t \end{pmatrix}; A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}; H = \begin{pmatrix} 1 \\ 0 \end{pmatrix}^T
$$

As an assumption, the observation noise is $V \sim N(0, R)$ ($R = \gamma$) and the system noise is $W \sim N(0, Q)$, where covariance matrix $Q$ is

$$
Q = \begin{pmatrix} \gamma + \delta & 0 \\ 0 & \delta \end{pmatrix}
$$

while the level and trend are assumed to be uncorrelated. Both the transition matrices and variance matrices are assumed to be constant, i.e. not time-dependent.

After the forecaster has decided initial values for level and trend ($= \hat{x}_{t-1}$) and the error margins (covariance matrix) for intialisation ($=P_{t-1}$), the forecasting algorithm presented in Figure 2 is usable. Using the Bayesian forecasting in more than one period long forecasts is done with

$$
\hat{x}_{t+k} = \hat{x}_t + \hat{\beta}_t k,
$$

14
where \( \hat{x}_t \) and \( \hat{\beta}_t \) are the latest level and trend estimates, which by definition are the most accurate ones.

### 2.2.6 Combinations

One alternative for a statistical forecasting algorithm is a combination of single algorithms. In general, there are numerous possibilities to combine single algorithms (e.g., linear combinations, regression techniques). The DM offers only weighted average, which is basically a linear combination of single forecast values:

\[
\hat{x}_t = \sum_{i=1}^{N} \hat{x}_{i,t} w_i,
\]

(18)

where \( \hat{x}_{i,t} \) is forecast of the \( i \)th method at time \( t \) and \( \sum_i w_i = 1 \). Clearly, when \( w_i = \frac{1}{N} \), (18) becomes the arithmetic mean.

A considerable amount of work has been done in the field of combinations. De Menezes et al. (2000) present practical guidelines for choosing the appropriate forecast combining method. As the DM offers only averaging, the results of De Menezes et al. are not presented here.

### 2.3 The forecast accuracy measurements

The necessity of forecast accuracy measurement stems from the fact that fitting a certain model to an existing data does not imply that an algorithm based on this model would produce the best (most accurate) future forecasts, while one can rarely argue that the demand really follows a specific model. This makes measuring the forecast accuracy sensible, as we quite seldom can reason before measuring that a certain algorithm produces the absolute optimal forecasts. An interesting problem is choosing the right (i.e. best suitable) measure.
Intuitively, measuring the average error between the observed value and its forecasted value might sound tempting, but is often misleading as the positive and negative values cancel each other out. Therefore, a better measure is the mean absolute deviation (MAD):

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \hat{x}_i|. \quad (19)$$

There are some disadvantages when using MAD, mainly considering the comparability between differentially scaled time series and the unstability when the time series is non-stationary. The latter means that the error value increases as the time series level increases although the magnitude of error would remain the same. The answer to these issues is relative error measuring, of which the mean absolute percentage error (MAPE) is the most common one:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|x_i - \hat{x}_i|}{x_i}. \quad (20)$$

MAPE is not ideal either. First of all, it demands ratio-scaled data. This is not a problem when speaking about demand (there exists a natural zero in demand). MAPE also puts a larger weight to forecasts that exceed the actual value vs. the ones that are under it. Therefore, it should not be used if large errors are expected. Another choice would naturally be mean absolute percentage error forecast (MAPE-F):

$$\text{MAPE - F} = \frac{1}{n} \sum_{i=1}^{n} \frac{|x_i - \hat{x}_i|}{\hat{x}_i}, \quad (21)$$

which on the other hand puts more weight on the below values. Quite commonly used is the adjusted mean absolute percentage error (MAPE-A)

$$\text{MAPE - A} = \frac{1}{n} \sum_{i=1}^{n} \frac{|x_i - \hat{x}_i|}{\frac{x_i + \hat{x}_i}{2}}, \quad (22)$$

which is unfortunately (being a solution to error sign favoring) not included in the DM. There exists however another solution, namely mean absolute percentage error max (MAPE-M) that can be computed by replacing the average in the nominator of (22) with $\max(x_i, \hat{x}_i)$.

Outside the scope of the DM lies some interesting methods, of which one useful (and a fairly simple one) is the relative absolute error (RAE). It is based on comparing 2 different
forecast errors produced by different methods. Defined as in Armstrong & Collopy (1992):

$$\text{RAE} = \frac{|x_i - \hat{x}_i|}{|x_i - \hat{y}_i|},$$

(23)

where $\hat{y}_i$ is the forecast from a random walk model $\hat{y}_{i+1} = x_i$. RAE is more adaptive to large fluctuations and should be preferred to MAPE if possible. When averaging, one must notice that the mean of RAE cannot be arithmetic (as in MAPE is the arithmetic mean of absolute percentage errors), but can be achieved by taking the geometric mean of RAE (GMRAE):

$$\text{GMRAE} = \left( \prod_{i}^{N} \text{RAE} \right)^{1/N}.$$  

(24)

In order to avoid zero and infinite products, it is suggested to use a cut-off filter to RAE values, so that $RAE \in [0.01, 10]$.

More modern and interesting discussion of accuracy measurements can be found in e.g. Hyndman & Koehler (2005).

3 Algorithm evaluation

The evaluation of algorithms can be done theoretically or by empirical means. By using the latter, a literature review and tests of algorithm accuracy are presented next.

Perhaps the most extensive empirical forecasting study is the M3-competition by Makridakis & Hibon (2000). The idea of the competition is that the competitors (representing different statistical algorithms) have to provide forecasts to a wide range of time-series (3003 different series to be exact) and the results of accuracy are gathered in order to find the best forecasting algorithm. Based on the results and similar competitions before, Makridakis & Hibon make the following propositions:

- Statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones.
- The accuracy of combining various methods outperforms, on the average, the individual methods being combined and does very well in comparison to other methods.

- The accuracy of the various methods depends upon the length of the forecasting horizon involved.

However, Koning et al. (2005) found that after running statistical tests of competition results, two postulates of the three should be modified. First of all, the combination algorithm used in the competition was not more accurate than single algorithms. Second, some methods are significantly better than others independent of the time horizon.

In addition to general findings of Makridakis & Hibon, there were also some suggestive results for single algorithms. The algorithms (“methods”) used in the competition included all presented in chapter 2.2 except the one based on Bayesian forecasting. CROSTON was only represented in a commercial package of ForecastPro\(^2\), which selects one from the following algorithms: SES, Box Jenkins ARMA, Poisson and negative binomial models, CROSTON and SMA. It was also methodically the closest one to the DM examined here.

ForecastPro does, on average, very well in the competition. Another remark made is that a fairly new algorithm \textit{theta} (presented in Assimakopoulos & Nikopoulos 2000) does remarkably well. However, Hyndman & Billah (2003) showed that theta is actually a special case of SES with drift and is therefore not covered in this study in more detail. One more conclusion is that trend dampening should be used, as the \textit{dampen trend exponential smoothing} performs quite well. The good performance of the \textit{combination} of three simple methods encouraged to a research by Hibon & Evgeniou (2002), which supported following results.

First, in the choice of the best forecasting algorithm, the combinations should be treated as a single method as there is no evidence that they always outperform the single methods

\(^2\)http://www.forecastpro.com/
or vice versa. Second, there is some evidence that the longer the forecast horizon, the better the combinations perform in proportion to single algorithms. Therefore it might be sensible to prefer combinations in the long run forecasts.

One specific and important area of forecasting is the intermittent demand, while many demand data in fact are intermittent. Syntetos & Boylan (2005) compare SMA, SES, original Croston and its modification, and conclude that if the demand interval is close to one, their modification performs better than the others but in general, SES and SMA are even better than both Croston algorithms. However, the Croston modification (by Teunter & Sani) presented in section 2.2.3 differs from the one Syntetos & Boylan used and therefore the results are not necessarily applicable. Willemain et al. (2004) present a modern technique using a bootstrap method to intermittent time series. The results show that the algorithm performs better than SES and the original Croston. Once again, one cannot say for sure that the Croston as in (7) would be less accurate than the algorithm by Willemain et al., while it was not used in the study.

There are also some suggestions considering the automated algorithm choosing system, the so called “Pick Best” phase. Wagner (2004) suggests using only 1-3 forecasting algorithms with equal parameter settings for a certain product group. This is based on (in addition to practical experience) the fact that pick best options can often be only a little (if all) modifiable, which may lead to unwanted results. The second problem relates to the first and to error measures, that is: the error measure used in the pick best phase does not tell you anything about, for example, the algorithm robustness. Therefore, forming the candidate group from a small amount of well behaving algorithms is the best way to use black box algorithm selection.
3.1 Forecasting test with simulated data

To compare the literature findings with praxis, a simulation test was implemented in MATLAB. The structure of the test was as follows. First, three sets of 100 x 100 data were generated: one representing the fast movers (dataset "A"), one the slow movers (dataset "B") and one the intermittent demand. The generation of data \((X_t)\) sets consisted of creating noise with multiplicative trend \((\beta_t)\), which was modelled as a random walk:

\[
\beta_t = \beta_{t-1} + \varepsilon_{\beta} \\
X_t = Y_t \beta_t,
\]

where \(Y_t \sim N(\mu_X, \sigma_X)\) and \(\varepsilon_{\beta} \sim N(0, \sigma_{\beta})\). The fast movers are items with high and frequent demand (high \(\mu_X\), low \(\sigma_X\)) as the slow movers have a smaller demand with
Table 2: The numerical characteristics of the simulated data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>( \bar{x} )</th>
<th>min ( x )</th>
<th>max ( x )</th>
<th>( \bar{\sigma} )</th>
<th>( P_{\text{demand}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>543.98</td>
<td>213.51</td>
<td>1456.5</td>
<td>47.03</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>53.03</td>
<td>2.09</td>
<td>188.44</td>
<td>27.35</td>
<td>0</td>
</tr>
<tr>
<td>Intermittent</td>
<td>1096.00</td>
<td>0.74</td>
<td>3484.20</td>
<td>541.42</td>
<td>0.31</td>
</tr>
</tbody>
</table>

more fluctuation (low \( \mu_X \), high \( \sigma_X \)).

The numerical characteristics of the data can be seen in Table 2. For the intermittent demand, only nonzero demands were used in calculations. Illustrative examples of each data set are presented in Figure 3.

In the case of intermittent data, the demand values occurred randomly, i.e. \( X_t = IX_t \) with \( I \sim Bernoulli(P_{\text{demand}}) \). Seasonality was only treated as a special case by a separate data set (n=10), which was generated like the non-seasonal series multiplied by a sine function.

3.1.1 Regular data forecasts

The length of each data set was 100 points of which as a historical part was considered the first half, i.e. the beginning point \( T_0 = 50 \). With each of the algorithms, a forecast was generated with time-horizons of 1, 10 and 20 which causes the corresponding amount of reference points to be 50, 41 and 31. The forecasting algorithms were then compared with the help of forecasted and “real” values using different error measures.

For the regular data sets A and B the algorithms of SES, SMA, BAYES and 3PLUS were used. In addition to choosing algorithms, one has to define the parameters used and e.g. in the case of 3PLUS, some other things such as whether to use the coefficient update as new datapoints are observed or initialise the whole algorithm from the beginning. The
Table 3: The forecast of regular data: GMRAE results.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>GMRAE (Set A)</th>
<th>GMRAE (Set B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t = 1$ $t = 10$ $t = 20$</td>
<td>$t = 1$ $t = 10$ $t = 20$</td>
</tr>
<tr>
<td>SES ($\alpha = 0.5$)</td>
<td>0.84 0.89 0.95</td>
<td>0.85 0.87 0.88</td>
</tr>
<tr>
<td>SES (optimal $\alpha$)</td>
<td>0.79 0.84 0.87</td>
<td>0.81 0.81 0.84</td>
</tr>
<tr>
<td>SMA (window = 6)</td>
<td>0.83 0.91 0.95</td>
<td>0.84 0.83 0.86</td>
</tr>
<tr>
<td>SMA (optimal window)</td>
<td>0.77 0.82 0.85</td>
<td>0.83 0.81 0.83</td>
</tr>
<tr>
<td>BAYES</td>
<td>0.82 0.82 0.78</td>
<td>0.83 0.86 0.94</td>
</tr>
<tr>
<td>3PLUS (optimal $\alpha$)</td>
<td>0.78 0.75 0.68</td>
<td>0.82 0.81 0.84</td>
</tr>
<tr>
<td>3PLUS (no update)</td>
<td>0.85 0.81 0.76</td>
<td>0.81 0.80 0.83</td>
</tr>
</tbody>
</table>

SES was implemented with $\alpha = 0.5$ and with optimal $\alpha \in [0.1, 0.9]$, defined separately for each time series using minimal MAPE as criterium. The SMA implementation was made with window length of 6 and optimal length $\in [1, 10]$. For the Bayesian algorithm, only the measurement variance of 1000 (same magnitude as the data variance) was used. The 3PLUS was run with and without the smoothing updates presented in (9)-(11) and, when updating, using optimal overall smoothing constant $\alpha \in [0.1, 0.9]$ and $\beta, \gamma$ being equal to 0. The latest choice of zero smoothing constants is made based on the assumption that there is no need for trend or seasonal coefficient update. In practice, when time horizon is more than 1, this choice is critical for algorithm stability. The “without update” alternative means that the algorithm is initialised from the scratch after every timestep.

In Table 3, the GMRAE results for regular data are presented. Basically the results show in what proportion is the algorithm error in comparison to a random walk model (i.e. moving average with window 1). From the table, some observations can be drawn. First, algorithms more sophisticated than the random walk will provide more accurate forecasts. Second, the more complex methods BAYES and 3PLUS perform better than the simple ones, at least with the less scattered data (Dataset A). Some more results can
be achieved by analysing the MAPE-M values of the forecasts (see Table 4).

The MAPE-M values can be compared with a statistical paired $t$-test for dependent samples. For all the test results presented here, the significance level of 95% was used. For the dataset A with time horizon $t = 1$, the 3PLUS with optimal $\alpha$ was clearly the best in every pairwise comparison. The 3PLUS without updates was the worst algorithm together with SES with $\alpha = 0.5$. With $t = 10$, optimal 3PLUS was still the best and 3PLUS without update and BAYES performed equally good and also significantly better than the simple ones. Also the optimal versions of SES and SMA performed equally - and significantly better than the ones with constant parameters. The same results apply for $t = 20$ with the addition that 3PLUS without update performs significantly better than BAYES.

For the dataset B, 3PLUS without updates performed the best and SES with $\alpha = 0.5$ the worst for $t = 1$, other algorithms being equally good. With a longer time horizon, 3PLUS without updates remained the best, the worst ones being SES with $\alpha = 0.5$ and BAYES. With the longest time horizon the same division seems to remain (see Table 4), but actually there is no (statistically significant) difference between the four best methods.
3PLUS(no updates), SES and SMA’s. The BAYES algorithm performs extremely bad.

The algorithms were also tested with a small sample (10 series) of seasonal data. The presence of seasonality highlights some facts. First, the mean forecast error of all algorithms with time horizon 10 is only 65% compared with the random walk - considerably better than with the non-seasonal data (see Table 3). Second, the statistically significant comparisons of MAPE-M show that the 3PLUS without update performs best (MAPE-M of 15.18% compared to second one: 3PLUS with optimal $\alpha$ and MAPE-M of 17.15%) and that especially SES with constant smoothing and the BAYES algorithm perform poorly compared to the others. It is quite clear that when seasonality is present, the 3PLUS is the way to go.

Before going into intermittent data, the error measures are compared. One way to approach the measure comparison is to evaluate how the best algorithm of each series differs by the error measure. For dataset A, the MAPE and MAPE-M concur almost completely. There are only a couple of series, in which the most accurate algorithm differs by the measure and in these exceptions, the marginals are very narrow (i.e. the best according to MAPE was second best according to MAPE-M etc.). RAE seems to concur with MAPE and MAPE-M for around 80% of the series and, when not concurring, it still describes the algorithm ranking quite similarly, i.e. the ones having a small MAPE/MAPE-M tend to have a small RAE too. When comparing the errors of the dataset B some changes arise: RAE and MAPE-M agree on the best algorithm in approximately 80% of the cases, which also corresponds with the results of the dataset A. MAPE, however, concurs with RAE only in 50% of the series, meaning that MAPE disagrees on the most accurate algorithm with MAPE-M and RAE very often.
Table 5: The forecast of intermittent data: MAPE and MAPE-M results.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MAPE</th>
<th></th>
<th></th>
<th>MAPE-M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t = 1$</td>
<td>$t = 10$</td>
<td>$t = 20$</td>
<td>$t = 1$</td>
</tr>
<tr>
<td>SES (optimal $\alpha$)</td>
<td>0.896</td>
<td>1.289</td>
<td>1.281</td>
<td>0.455</td>
</tr>
<tr>
<td>SMA (optimal window)</td>
<td>0.767</td>
<td>1.150</td>
<td>1.267</td>
<td>0.486</td>
</tr>
<tr>
<td>CROSTON (optimal $\alpha$)</td>
<td>0.405</td>
<td>1.203</td>
<td>1.202</td>
<td>0.302</td>
</tr>
</tbody>
</table>

3.1.2 Intermittent data forecasts

The intermittency in data offers more challenges to forecasting. The aim of the simulation test described below is to justify, which one of the (for intermittent data regularly used) forecasting algorithms, SES or CROSTON, suits best for intermittent data. As a reference algorithm, the SMA was also included.

The data used in the test is described in Table 2. All algorithms (SMA, SES and CROSTON) were used with the optimal parameters (in minimum MAPE sense) and the time horizons were the same as before (1, 10 and 20). Because of large amount of zero-demand (i.e. large amount of constant demand per period dates), RAE was not useful as an error measure and therefore only MAPE and MAPE-M were considered. The results of the test are in Table 5. The quantity that was measured is not the demand but demand per period, i.e. the single demand quantities were divided to preceding inter-demand time.

Statistical testing shows that (in MAPE-M-sense), CROSTON outperforms other algorithms by far, independent of the time horizon. With $t = 1$, SES performs better than SMA, but otherwise, there is no significant difference between the two worse algorithms.

As the algorithm errors in general are quite high, the information about the robustness of the algorithms is also interesting. For each time horizon, the maximum value of MAPE is gained by SES, the second largest by SMA and the smallest maximum by CROSTON. The minimum values are achieved by SES or CROSTON, which indicates that the SMA
Table 6: The forecast of intermittent data: range and variance of MAPE.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>VAR(MAPE)</th>
<th>MAX(MAPE)-MIN(MAPE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t = 1$</td>
<td>$t = 10$</td>
</tr>
<tr>
<td>SES (optimal $\alpha$)</td>
<td>0.96</td>
<td>2.15</td>
</tr>
<tr>
<td>SMA (optimal window)</td>
<td>0.52</td>
<td>1.28</td>
</tr>
<tr>
<td>CROSTON (optimal $\alpha$)</td>
<td>0.16</td>
<td>0.92</td>
</tr>
</tbody>
</table>

never succeeds particularly well. Altogether, as Table 6 indicates, the smallest MAPE range is achieved by CROSTON, the second smallest by SMA and the largest by SES. The variances of MAPE follow the exactly same order, so all in all one can state that CROSTON is the most robust of the tested algorithms.

3.2 Forecasting test with real data

The real demand data provided consisted of 100 data sets, which were classified by item type to 4 different classes, A (5 series), B (15 series), C (35 series) and D (45 series). The proportional size of each class corresponds with the real item distribution.

The class A means fast movers and B and C respectively items with lower demand, the slow movers. The class D stands for intermittent demand. The data length varied from around 30 to 150 points per set and the values are weekly demands. In the data, there were some missing values, which were ignored with A,B and C items and considered as zero demand with class D items. The seasonality of data was not tested, while the plots of data indicated no signs of seasonality. The numerical characteristics of the data sets can be found in Table 7 and an illustrative example of each series in Figure 4.

The time series, even the ones belonging to the same demand class, differed from each other significantly. For example in the class C, the smallest maximum of a single series was only 25 compared to the largest of 6027 and the standard deviation differed enormously
from 5.28 to 675.49. The series also clearly represent items in different stages of their life-time: in some cases the demand is strongly emphasised in the beginning and is very close to zero in the end, whereas some items have a growing demand throughout the series. Many series also resemble intermittent demand pattern, as there occurs single demand spikes among otherwise stable (nonzero) demand.

Running some test forecasts shows that the time series, which contain a lot of values close to zero, behave problematic in the forecast error sense. The problems caused are twofold: first, 3PLUS and BAYES (which model both level and trend) can forecast negative demands if the trend is heavily diminishing. Second, the forecast accuracy measures tend to reach impractical (i.e. very high) values if the real demand is small and/or the forecast value is negative. So, in order to gain somewhat sensible results, the following
Table 7: The numerical characteristics of the real demand data.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>$\bar{x}$</th>
<th>min $x$</th>
<th>max $x$</th>
<th>$\sigma$</th>
<th>$P_{demand}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1503.6</td>
<td>16</td>
<td>26025</td>
<td>1430.8</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>364.6</td>
<td>1</td>
<td>5715</td>
<td>382.1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>93.2</td>
<td>1</td>
<td>6027</td>
<td>146.3</td>
<td>0</td>
</tr>
<tr>
<td>Intermittent</td>
<td>225.1</td>
<td>1</td>
<td>39000</td>
<td>360.0</td>
<td>0.76</td>
</tr>
</tbody>
</table>

actions were made:
- The forecast values are forced to be non-negative, i.e. $\hat{x}_t = \max(\hat{x}_t, 0)$
- Time-series, for which MAPE was over 5 for at least 2 algorithms, were excluded

One should notice that the actions presented do not remove the problemacy of incompetent forecasts. They only help us to remove the influence of these forecasts from the error comparison, so that the best algorithm that suits for most of the cases can be discovered. In the practise, an automated alerting system should be utilised so that the extreme values can be adjusted subjectively.

The results of the test are divided in two dimensions. First, the overall results of every class are discussed. Second, some results on individual algorithms are investigated.

For the fast-movers, the amount of data was remote (5 series) and therefore the results not very reliable. It is somewhat surprising to notice that BAYES seems to work best with the data and that, regardless of the time horizon, the 3PLUS (with optimal $\alpha$ and smoothing updates) does not perform significantly better than the simpler algorithms. With the exception of 0.76 (BAYES with horizon of 20), the GMRAE values vary between 0.89 and 1.15, which indicates that the forecasting is remarkably more inaccurate than with the simulated data (see Table 3).

The dataset B differs only a little from the results for dataset A. The BAYES algorithm is no longer superior and both versions of 3PLUS and BAYES perform significantly
worse ($t$-test for MAPE-M values) than the simpler algorithms (both with constant and optimal parameters) with all time horizons. The average GMRAE remains at around 0.9 and the 3PLUS without updates, being the worst of algorithms in GMARAE-sense, has over 1 in over 50% of the series.

![Graph showing forecasting problems](image)

Figure 5: An example of forecasting problems.

The class C consisted of 35 series, of which some (6%, 9% or 23% of the series corresponding to $t=1$, 10 or 20) were abandoned because of too high MAPE values. One example of such a series can be found in Figure 5, which points out two major problems that arise from the provided data. First, single extreme values (such as in the figure at time 91) distort the forecasts and make the forecast measures practically unusable. Second, although the demand mean was high (in the figure 258.1), the trend is strongly diminishing, which causes some of the algorithms to be zero in the end. At the same time, the MAPE values are very high because the nominator can be small compared to absolute error value. The MAPE-M and GMRAE results of accepted forecasts are to be
Table 8: The MAPE-M and GMRAE results of data C forecasting.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MAPE-M</th>
<th>GMRAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t = 1$</td>
<td>$t = 10$</td>
</tr>
<tr>
<td>SES ($\alpha = 0.5$)</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>SES (optimal $\alpha$)</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>SMA (window = 6)</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>SMA (optimal window)</td>
<td>0.49</td>
<td>0.52</td>
</tr>
<tr>
<td>BAYES</td>
<td>0.52</td>
<td>0.61</td>
</tr>
<tr>
<td>3PLUS (optimal $\alpha$)</td>
<td>0.56</td>
<td>0.65</td>
</tr>
<tr>
<td>3PLUS (no update)</td>
<td>0.65</td>
<td>0.69</td>
</tr>
</tbody>
</table>

It is challenging to draw any kind of conclusions from the results. The difference between the ranks by measure is explained by the fact that for roughly every fourth of the series, algorithms using trend in calculations have a zero forecast for many (or all) data points, which implies MAPE-M = 1. Therefore, the MAPE-M values seem useless. The GMRAE results indicate that 3PLUS with optimal $\alpha$ and SMA might perform better than the others. However, at least one finding of Makridakis & Hibon (2000) can be confirmed from the results: there is no remarkable difference between the simpler and more complicated methods.

For the intermittent data (D class) all forecasts were taken into account when $t=1$ and the rejection percents for $t=10$ and $20$ were 9% and 29% respectively. Many of the rejected forecasts consisted of heavily diminishing demand, where the historical part of the data included high demand season and the forecasted part almost nonzero demand period. Without any external information, this kind of behaviour is impossible to forecast. Anyway, the MAPE and MAPE-M results (Table 9) show that (in MAPE-M sense) CROSTON outperforms other algorithms for every time horizon and that there is no
Table 9: The MAPE and MAPE-M results of intermittent data forecasting.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MAPE</th>
<th>MAPE-M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t = 1$</td>
<td>$t = 10$</td>
</tr>
<tr>
<td>SES (optimal $\alpha$)</td>
<td>1.34</td>
<td>2.71</td>
</tr>
<tr>
<td>SMA (optimal window)</td>
<td>1.58</td>
<td>2.63</td>
</tr>
<tr>
<td>CROSTON (optimal $\alpha$)</td>
<td>0.33</td>
<td>3.01</td>
</tr>
</tbody>
</table>

While the overall results provided only little knowledge of the differences between the algorithms, there is still something to learn about single algorithm performance. For example, for SES and CROSTON, the choice of the smoothing factor(s) can make a difference in algorithm accuracy. With the dataset C, the optimal $\alpha$’s were considerably lower with $t=1$ when compared to $t=10$ or 20. The values however differed a lot and with all time horizons: every $\alpha \in [0.1, 0.2, \ldots, 0.9]$ was chosen to be the optimal at least once. Therefore, if no external reason points otherwise, the choice of optimal smoothing factor(s) should be utilised whenever possible.

The optimal window length distributions for SMA are plotted in Figure 6. The mode for longer time-horizons is 1, which corresponds to the random walk model. Still, every other value is represented at least once, so the same conclusion as with smoothing factor apply: if possible, the optimal window length should be chosen individually for each forecasted series. If this is not possible, the best suggestion based on the distributions presented is 4, which is the median of every time horizon.
4 Discussion of results

The testing of simulated data highlights some characteristics of the algorithms and error measures. The results were not always self-explanatory and therefore, the conclusions drawn are only indicative. With the regular data, the 3PLUS seems to work quite well. A clear division between the simple and more complicated algorithms favoring BAYES and 3PLUS was observed with dataset A but, with dataset B, the BAYES algorithm turned out to be the worst with longer time-horizons and altogether less robust than 3PLUS. The largest difference between accuracies is observed with dataset A in long time-horizons, which indicates that the trend factor missing from SES and SMA is an important factor in forecasting accuracy. The simple algorithms tend to underestimate the demand if data has a growing trend and vice versa. With more fluctuation (dataset B), the importance of trend is diminished and actually, the 3PLUS with updates and
optimal $\alpha$ seems to perform worse than its simpler version, SES. In the case of seasonal data, the 3PLUS should always be preferred.

The results with the real data were poor compared to the ones with simulated data or in the literature. The main reason for this was the randomness of data, which was greater than in the simulated data. The results support two conclusions. First, combined with the results from the simulation test, some algorithm usage suggestions can be given. As long as the data randomness is high and no evident patterns (trend, seasonality) occur, the choice of the forecasting algorithm can be made freely among SMA, BAYES and 3PLUS, with and without the trend (SES), using the logic of minimising some forecast error measure. When patterns do arise, 3PLUS should be used. A trivial fact considering all forecasts is that, whenever possible, algorithm parameter optimisation should be exploited if there is no need for judgemental adjustment of parameters.

With intermittent data, the CROST ON algorithm seems to outperform SMA and SES in every case, with both simulated and real demand data. It also appears to have a smaller variance and range of variation in accuracy and therefore, it is more reliable choice than the other algorithms.

The choice of the best algorithm is based on minimising some error measure and the measure choice plays an important role. The simulation test, especially with dataset B, showed that MAPE disagrees with RAE (which is a robust and versatile measure) clearly more often than MAPE-M. Therefore, MAPE-M can provide more robust results and should be used.

The second major conclusion revealed from the results is the demand for explaining the data behaviour. The results indicate that major problems arise, when extreme values or unexpected patterns occur in the data. If this behaviour can be explained and taken into account, otherwise useless forecasts can be used and/or the forecast errors diminish substantially. Some ways to solve these kinds of problems are offered by integration of judgemental forecasting.
The DM used in the project provides a great deal of integration possibilities. Starting from the simplest algorithm possible, namely SMA, the choice of window length can be adjusted subjectively. If drastic changes in demand pattern are expected, one must decide whether to “follow” the changes with a small window or dampen the forecast noise with a longer window. As with all smoothing techniques, the choice of weight constants can be critical. Instead of determining the constant(s) by minimising the forecast error, the weight constant could be forced close to one in order to prefer the present knowledge to history. The DM also allows to specify a “decay parameter” to some algorithms, which basically means adding a trend dampening factor. So, if the forecaster sees that demand growth is diminishing, the dampening factor should be set under its assumed value 1. In the data provided, diminishing trend was present in many series and taking it into account in the algorithms would probably have been beneficial.

Many demand planning software also provide a vast amount of data manipulating tools. Data manipulation can be thought as judgemental forecasting in two different sense. The manipulation of input data (removing/replacing missing values, using cut-off filters etc.) is a part of model building whereas the modifying of output data (forecast algorithm results) is considered as judgemental adjustment. In both ways, the forecast accuracy or usability is most likely to improve and in fact, the forecasting test of real data showed that taking the extreme values into account can be crucial to forecast accuracy (see e.g. Figure 5).

5 Conclusions

The aim of the study was, as presented in section 1.1, to examine how the accuracy of forecasting spare part demand could be improved. Based on some findings on the literature, a forecasting test of three simulated time series sets and a test with real demand data, the statistical black-box forecasting was found to be fairly inaccurate and, in some cases, not applicable to forecasting at all. However, if forecasts for thousands of
items have to be generated, using automated systems like the DM is a necessity.

While even a minor improvement in the forecast accuracy can provide significant savings in form of e.g. diminished waste caused by overstocking, some suggestions considering the algorithm use based on the forecasting tests are literature findings were given. One major conclusion was that in the search for the best algorithm, narrowing the alternatives to few, well performing algorithms is the best way to use such black-box systems. Also some more detailed suggestions were provided.

Still the role of judgemental forecasting must be emphasized: the integration of judgemental and statistical forecasts can improve the forecast usability significantly. Some ideas based on the forecasting tests were presented in section 4. Other judgemental and statistical forecasting techniques, which were not in scope of this study but could turn out to be useful and worth further studying, are presented next.

The combination techniques could be easily implemented in practise. This demands putting judgemental information into numbers and combining the results with a statistical forecast. An example of this would be collecting sales and product demand estimates based on expert opinions (using e.g. the delphi method), computing purely statistical forecast and averaging these two. A detailed discussion of statistical algorithm combinations was also excluded from the study. They could, however, bring more accuracy and above all, more robustness to algorithm variety.

As the choice of algorithm certainly makes a difference if data has patterns (seasonality, differentially shaped trends, etc.), some kind of a pattern recognition system could improve the functionality of an automatic algorithm selection system. For example, if a statistical test shows that there exists seasonality in the data, always choose 3PLUS. Or, if there exists a statistically significant linear trend, choose SES with drift, 3PLUS or BAYES and so on. Especially with long time horizons, the algorithms including trend modelling do perform better if the data de facto has a trend, and therefore the evaluation of existance of the trend is essential.
Also the possibilities of using regression analysis were completely ignored in this text - a subject that can improve the forecast accuracy significantly if used correctly (Meyr 2004). Likewise something as simple as taking product group cross correlations into account could provide a substantial improvement to forecast accuracy.

Finally, however tempting it would be to use the forecasting results without critical evaluation, one should bear in mind the statement made by Spyros Makridakis, one of the most experienced researcher in the area of empirical forecasting:

"It is unfortunate but true that users of quantitative forecasting techniques have no simple, reliable way to predict what will happen when established patterns or relationships change."
References


*The Demand Manager User Guide*


