Forecasting with multivariate temporal aggregation: The case of promotional modelling

Nikolaos Kourentzes a,*, Fotios Petropoulos b

a Lancaster University Management School, Department of Management Science, Lancaster, Lancashire, LA1 4YX, UK
b Logistics and Operations Management Section, Cardiff Business School, Cardiff University, UK

1. Introduction

Demand forecasting is central to decision making and operations in organisations. As the volume of forecasts increases, for example due to an increased product customisation that leads to more SKUs being traded, or a reduction in the length of the forecasting cycle, there is a pressing need for reliable automated forecasting. Conventionally, companies rely on a statistical baseline forecast that captures only past demand patterns, which is subsequently adjusted by human experts to incorporate additional information such as promotions. Although there is evidence that such process adds value to forecasting, it is questionable how much it can scale up, due to the human element. Instead, in the literature it has been proposed to enhance the baseline forecasts with external well-structured information, such as the promotional plan of the company, and let experts focus on the less structured information, thus reducing their workload and allowing them to focus where they can add most value. This change in forecasting support systems requires reliable multivariate forecasting models that can be automated, accurate and robust. This paper proposes an extension of the recently proposed Multiple Aggregation Prediction Algorithm (MAPA), which uses temporal aggregation to improve upon the established exponential smoothing family of methods. MAPA is attractive as it has been found to increase both the accuracy and robustness of exponential smoothing. The extended multivariate MAPA is evaluated against established benchmarks in modelling a number of heavily promoted products and is found to perform well in terms of forecast bias and accuracy. Furthermore, we demonstrate that modelling time series using multiple temporal aggregation levels makes the final forecast robust to model mis-specification.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
reliability and relatively good accuracy (Makridakis and Hibon, 2000; Gardner, 2006). The exponential smoothing family of methods is capable of modelling a wide variety of time series with or without trend and seasonality. With the incorporation of exponential smoothing in a state-space framework its statistical underpinnings were researched, resulting in an elegant and effective automatic model selection procedure (Hyndman et al., 2002, 2008). The basis of this model selection is to fit the various forms of exponential smoothing and choose the most appropriate based on a pre-selected information criteria, typically Akaike’s (Hyndman et al., 2002; Billah et al., 2006). This approach has been implemented in various statistical software (Hyndman and Khandakar, 2008) and is widely regarded as a benchmark for automatic univariate forecasting that is at the core of FSSs.

More recently, further refinements in the automatic specification of exponential smoothing have appeared in the literature. From one hand, Kolassa (2011) argued that identifying a single model by using information criteria may not always perform well and investigated the performance of combining models via Akaike weights, instead of choosing a single one. He found this approach to be superior, resulting in more reliable and accurate forecasts. On the other hand, Kourentzes et al. (2014) looked at the combination of exponential smoothing models that are fitted across multiple temporally aggregated versions of the initial time series. They argued that their approach, named MAPA (Multiple Aggregation Prediction Algorithm) has advantages over conventional exponential smoothing modelling because different time series components are attenuated or strengthened at different temporal aggregation levels, resulting in a more holistic estimation of the time series structure and more accurate forecasts. Their approach builds on the extensive literature on the effects of temporal aggregation on forecasting (for recent examples see: Zottori et al., 2005; Silvestrini and Veredas, 2008; Andrawis et al., 2011; Spithourakis et al., 2014; Rostami-Tabar et al., 2013). However, these approaches are not able to make use of additional information such as promotions. Nonetheless, promotional modelling is crucial for many areas such as manufacturers of fast moving consumer goods and retailing. As argued above, automatic promotional forecasting is desirable. Regression type statistical models are often used to build promotional models (Fildes et al., 2008), which incorporate multiple exogenous marketing inputs. Such models are hard to automate and require substantial expertise to maintain. Significant advances have taken place in promotional modelling at a brand level, involving sophisticated forms of temporal aggregation. Temporally aggregating a time series can cause various of its components to become more or less prominent with direct effects on model identification and estimation. MAPA uses multiple temporal aggregation levels, allowing multiple views of the data to be considered during model building and subsequently combined in a final forecast.

MAPA can be seen as a three step procedure, where in the first step the original time series is aggregated in multiple aggregation levels using non-overlapping means of length k. The mean is used instead of the sum, as it retains the scale of the series across the various aggregation levels. Given a time series Y, with observations y_t, and t = 1,...,n, temporal aggregation can be performed as

\[ y_{(k)}^{[l]} = k^{-1} \sum_{t = (l-1)k}^{lk} y_t. \]  

(1)

The temporally aggregated time series is noted with a superscript [k] and has less observations than the original time series. For example for k=2 the resulting series Y^{[2]} will have half as many observations as the original time series. Note that the latter can be written under this notation as Y^{[1]}. Depending on the aggregation level k it may be that the division n/k has a non-zero remainder, in which case the n – \lfloor n/k \rfloor k first observations of the time series are ignored in the construction of the aggregated one. The aggregation operator in Eq. (1) acts as a moving average and the resulting time series is smoother than the original one. High frequency components are progressively filtered as the aggregation level increases, essentially attenuating the seasonal and random component of time series, while allowing the low frequency trend and level components to dominate, capturing these better. Petropoulos and Kourentzes (2014) suggested that aggregating up to time series of yearly time buckets it is sufficient, since all high frequency components will be filtered by then, allowing to clearly see all low and estimation and robustness advantages of MAPA. We investigate the performance of the proposed method using a real case study of heavily promoted demand series of cider SKUs (Stock Keeping Units) of a popular brand in the UK. We use as benchmark the extended exponential smoothing that includes external promotional information, to demonstrate the advantages of using multiple temporal aggregation levels, and a regression based promotional model from the literature. We find that multivariate MAPA outperforms all benchmarks substantially, providing a useful candidate for a fully automatic promotional model. Furthermore, we find that exponential smoothing performs very well against regression based promotional models. We argue that one of the major advantages of the proposed method is its robustness to model misspecification and therefore its reliability for practical implementations.

The rest of the paper is organised as follows: Section 2 describes MAPA and introduces our extension to model external variables; Section 3 describes the case study that will be used to empirically evaluate the proposed method, while Section 4 describes the experimental setup and the benchmarks used in this research; Section 5 presents the results, followed by a discussion on the benefits of temporal aggregation for promotional modelling and conclusions.

2. Methods

2.1. Multiple Aggregation Prediction Algorithm

The Multiple Aggregation Prediction Algorithm (MAPA) was proposed by Kourentzes et al. (2014) to take advantage of the time series transformations that can be achieved by non-overlapping temporal aggregation. Temporally aggregating a time series can cause various of its components to become more or less prominent with direct effects on model identification and estimation. MAPA uses multiple temporal aggregation levels, allowing multiple views of the data to be considered during model building and subsequently combined in a final forecast.

This paper investigates the use of multiple temporal aggregation to construct enhanced and automated exponential smoothing based promotional models. We extend the MAPA approach to include external variables, using a similar formulation to multivariate exponential smoothing. The motivation is to combine the simplicity and reliability of exponential smoothing with the
high frequency elements of the series, although it is possible to consider even higher levels.

Subsequently, in the second step of MAPA a forecasting model is fitted at each aggregation level. Due to the aggregation operator it is expected that the original time series components will change. For fast moving consumer goods this means that seasonality may be present or trend easy to observe only some levels (Kourentzes et al., 2014), while for slow moving items the intermittency characteristics will change across different aggregation levels, until the time series becomes non-intermittent (Petropoulos and Kourentzes, 2015). Obviously, the underlying structure of the time series is constant, however due to the different sampling frequencies at the various aggregation levels, different elements of it become easier, more difficult or impossible to observe and estimate. Kourentzes et al. (2014) argued that this is a strength of the MAPA, as instead of selecting a single model, which may be wrongly identified, by repeating the process at each temporal aggregation level and combining the resulting models, potential problems due to errors in model selection and parameterisation are mitigated. However a new problem is introduced that results in the dampening of the estimated time series components. For example, let us assume that for a time series a seasonal model is estimated at one level, while a non-seasonal model is estimated at another. By combining the forecasts of these two levels the seasonal part is halved, assuming unweighted averaging is used. This is an undesirable property of forecast combination in the context of temporal aggregation, as it is expected that the time series components will not be present at all levels. To overcome this problem MAPA performs combination by time series components. The reader is referred to the discussion by Kourentzes et al. (2014) for more details.

Although in theory MAPA could use any forecasting method at each aggregation level, exponential smoothing is very suitable, as it separates a time series into level, trend and seasonal components during modelling. Exponential smoothing (ETS) models the level ($l_t$), trend ($b_t$) and seasonality ($s_t$) of a time series explicitly. These components are smoothed, and the level of smoothing is controlled by the smoothing parameters of ETS: $\alpha$ for the level, $\beta$ for the trend and $\gamma$ for the seasonal component. The smoothed components are then combined to give a forecast. Depending on the nature of the time series under consideration, these may interact in an additive or multiplicative way. Furthermore, the trend can be linear or damped, which is controlled by parameter $\phi$. Table 1 provides the error correction forms of exponential smoothing with additive errors. The following notations are used: $N$ for none, $A$ for additive, $Ad$ for additive damped, $M$ for multiplicative and $Md$ for multiplicative damped. The forecast is denoted by $\hat{y}_t$ and $e_t$ is the white noise error. Similar models exist for multiplicative error terms. To identify the correct form of ETS for each time series and temporal aggregation level the Akaike Information Criterion (AIC) is used, as it is suggested by Hyndman et al. (2002) for ETS modelling.

For MAPA we are interested in the last state vector $x_k^{[t]}$ of ETS, which contains the updated values of each $l_t$, $b_t$ and $s_t$: $x_k^{[t]} = (l_k^{[t]}, b_k^{[t]}, s_k^{[t]}, s_k^{[t-1]}, \ldots, s_k^{[t-M+1]})$. Using this information we can produce forecasts for any desirable horizon. Note that additive and multiplicative components will have different scale, as the later is expressed as a ratio of the level. This makes the combination by components difficult. To overcome this Kourentzes et al. (2014) proposed to first transform multiplicative components into additive using the formulae in Table 2.

The additive translation of the components is only used for constructing the out-of-sample component predictions that will be combined. Note that as these components are coming from different temporal aggregation levels, their length will be different. For example predicting at the monthly level a year ahead will result in twelve values, while in annual level will result in a single value. The translated component forecasts are returned to the original time domain using

$$z_t = \sum_{j=1}^{k} a_j z_t^{[j]}$$

where $z_t^{[j]}$ is the vector to be returned to the original time domain and $t = 1, 2, \ldots, n$ and $i = [t/k]$. Eq. (2) acts as a piecewise constant interpolation. The weights $a_j$ are equal to $k^{-1}$, resulting in an unweighted disaggregation scheme, which has been found to perform well (Nikolopoulos et al., 2011).

The last step of MAPA involves the combination of the components estimated across the different aggregation levels. Two

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>State space exponential smoothing equations for additive error.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trend</th>
<th>Seasonal</th>
<th>N</th>
<th>A</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_t = l_{t-1}$</td>
<td>$\mu_t = l_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + s_{t-m}$</td>
</tr>
<tr>
<td>$l_t = l_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + \alpha e_t$</td>
</tr>
<tr>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
</tr>
<tr>
<td>$\mu_t = l_{t-1} + b_{t-1}$</td>
<td>$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$</td>
</tr>
<tr>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
</tr>
<tr>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
</tr>
<tr>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
</tr>
<tr>
<td>$\mu_t = l_{t-1} + \phi b_{t-1}$</td>
<td>$\mu_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$</td>
</tr>
<tr>
<td>$l_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$</td>
<td>$l_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$</td>
<td>$l_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$</td>
<td>$l_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$</td>
<td>$l_t = l_{t-1} + \phi b_{t-1} + s_{t-m}$</td>
</tr>
<tr>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
</tr>
<tr>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
</tr>
<tr>
<td>$\mu_t = l_{t-1} + b_{t-1}$</td>
<td>$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$</td>
<td>$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$</td>
</tr>
<tr>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
<td>$l_t = l_{t-1} + b_{t-1} + \alpha e_t$</td>
</tr>
<tr>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
<td>$b_t = b_{t-1} + \beta e_t$</td>
</tr>
<tr>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
<td>$s_t = s_{t-1} + \gamma e_t$</td>
</tr>
</tbody>
</table>

Please cite this article as: Kourentzes, N., Petropoulos, F., Forecasting with multivariate temporal aggregation: The case of promotional modelling, International Journal of Production Economics (2015), http://dx.doi.org/10.1016/j.ijpe.2015.09.011
combination methods were originally proposed: using unweighted mean and median, which were found to perform very similarly. In the case of the unweighted mean, each component is combined using

$$T_{t+h} = K^{-1} \sum_{k=1}^{K} d_{k}^{(i)}.$$  

(3)

$$\bar{T}_{t+h} = K^{-1} \sum_{k=1}^{K} b_{k}^{(i)}.$$  

(4)

$$\bar{S}_{t+h} = K^{-1} \sum_{k=1}^{K} g_{k}^{(i)}, \text{if } (m/k) \in \mathbb{Z} \text{ and } k < m,$$

(5)

where $K$ is the maximum aggregation level considered and $K'$ is the number of aggregation levels where seasonality may be identified, i.e., when $m/k$ results in an integer and $k < m$, as ETS is not capable of capturing fractional seasonality. The following example illustrates this: Suppose a monthly sampled time series then $K' = 1, 2, 3, 4, 6, \ldots$ seasonality estimated and combined only at monthly, bi-monthly, quarterly, four-month, and semi-annual data. For trend, if at some temporal aggregation level no trend is fitted, then it is assumed that for that level the value of trend is zero.

To produce the final forecast for $h$ steps ahead, the forecast horizon of the original time series, the components can be simply added together, as they have been already translated into additively

$$\tilde{y}_{t+h}^{(1)} = \tilde{y}_{t+h} + \bar{B}_{t+h}.$$  

(6)

2.2. MAPA with exogenous variables

Here we will extend MAPA to include exogenous variables. Let $X_j$ with observations $x_{jt}$ be the $j$th explanatory variable to be included in our model and $j = 1, \ldots, J$. The formulations in Table 1 can be adjusted to include $X_j$ as follows:

$$\mu_t = \mu_1 + \sum_{j=1}^{J} d_{j,t},$$  

$$d_{j,t} = c_j x_{jt},$$  

(7)

where $d_{j,t}$ contain the effect of each $X_j$ variable at time $t$ and $c_j$ is its coefficient. Coefficients $c_j$ function in the same way as in a regression model, coding additive effects, while multiplicative

effects can be captured through logarithmic transformation of the data. This formulation is similar to the standard ETS with regressor variables (Hyndman et al., 2008), with the only difference being that the effect of each variable is measured separately in $d_{j,t}$ allowing to directly incorporate it in the MAPA framework. Estimation of $c_j$ is done simultaneously with the rest of the ETS states, $\mu$. This can be done either by least squares, maximum likelihood estimation or other desirable cost functions.

At each temporal aggregation level $k$ a separate $d_{j,k}^{(i)}$ is calculated, based on the estimated $d_{j,k}^{(i)}$ and temporally aggregated $X_j^{(i)}$. The resulting vectors are treated in the same way as the estimated time series components in the univariate case. First, they are translated into the original time domain using Eq. (2). Then these are combined into a single effect across all aggregation levels for each variable $X_j$

$$\bar{d}_{j,t+h} = K^{-1} \sum_{k=1}^{K} d_{j,k}^{(i)}.$$  

(8)

Finally, Eq. (6) that was used for the univariate forecast is adjusted to include the new multivariate effect estimations:

$$\tilde{y}_{t+h}^{(1)} = \tilde{y}_{t+h} + \bar{B}_{t+h} + \bar{S}_{t-m+h} + \sum_{j=1}^{J} \bar{d}_{j,t+h}.$$  

(9)

The parameters of the multivariate ETS at each temporal aggregation level will be optimised in the same way as the univariate ETS and the appropriate model form will be selected using AIC, as before. However, the temporal aggregation introduces one additional complexity for the multivariate models. As $X_j$ are aggregated, they become smoother as implied by the aggregation Eq. (1). This changes the correlation between explanatory variables and may introduce multicollinearity at higher aggregation levels, if more than one variable is included in the model. As an illustrative example consider the case of two different promotions or special events that occur only once per month at a different day of the month and are coded using binary dummies. At a daily level these variables are not collinear, but at a monthly temporal aggregation level both variables become the same, equal to a vector of ones. Clearly, if both variables were included in Eq. (7) estimating coefficients $c_j$ would not be possible. To avoid this it is desirable to transform the variables so that they become orthogonal. We can use principal components analysis to achieve this.
Principal component analysis generates a new set of variables, \( \hat{X}_j \) with \( j = 1, \ldots, J \), called principal components, which are linear combinations of the original variables. The weights of the linear combination are such that the resulting principal components are orthogonal to each other. Therefore, the new variables \( \hat{X}_j \) are no longer multicollinear and contain no redundant information (Jolliffe, 2002). These can now be used as inputs instead of the original \( X_i \) variables, overcoming the problems caused by temporally aggregating the explanatory variables. The principal components are constructed so that they are ordered in terms of variance, with the last components typically having very small variance. In practice we can omit these, thus reducing the number of inputs to less than the original \( J \). There are two commonly considered alternatives in choosing which components to retain. One can retain all components that are over a cut-off level in terms of variance. Alternatively, one can select to include only components that are significant in a regression context (Jolliffe, 1982). Here, for simplicity we use the first option, as conventional ETS parameter estimation does not typically provide standard errors of the estimated parameters that would allow the calculation of t-statistics. Note that it is still possible to obtain these by bootstrapping.

Therefore, by using principal components analysis we avoid the problem of multicollinearity of the inputs as the aggregation level increases and reduce the dimensionality of multivariate MAPA, making it less cumbersome to estimate.

Summarising, the extended MAPA works as follows. First the provided time series and promotions are temporally aggregated. At each aggregation level the data is processed as illustrated in the flowchart in Fig. 1. The promotional variables are first processed using principal components analysis and then incorporated in the exponential smoothing described by Eq. (7). From that the level, trend and seasonal components, as well as the promotional effect are extracted. The components are transformed to additive ones using the expressions in Table 2. Then, together with the promotional part these are returned to their original frequency using Eq. (2). Estimates from all temporal aggregation levels are combined using Eqs. (3)-(5) and (8) for each level, trend, season and promotion components respectively. Finally these are combined in the final forecast using Eq. (9).

3. Case study

We empirically evaluate the performance of the multivariate MAPA by exploring its performance over benchmarks in predicting the sales of products under multiple promotions. Data from one of the leading cider brands have been collected from a UK manufacturer. These forecasts are useful for the manufacturer to support production and inventory planning decisions. Demand for 12 variants of the brand, including SKUs with different package sizes and flavours, has been collected for 104 weeks. The manufacturer sells the SKUs to multiple retailers who are offered different promotions. The timing of each promotion has been provided and was coded as binary dummy variables. Each SKU may be under up to 6 promotions at any time, accounting for the different offers to each retailer, with a varying degree of success. The promotions in this case study are known in advance, as the company has control of the promotional plan.

Table 3 provides the average descriptive statistics across SKUs. Looking at the difference between the measures of central tendency and the maximum we can observe the impact of promotions on sales, which is also reflected in the skewness of the sales. It can also be seen that these SKUs are heavily promoted, having on average 3.25 different promotions that are active for 61.78% of the sample. Note that all SKUs in the case study are fast moving.

As an example Fig. 2 provides the sales and the timing of the promotions to various retailers for a single SKU of the case study, which is representative of other SKUs in the dataset. Periods when at least one promotion takes place are highlighted. As Fig. 2 illustrates the SKU is under some promotion in almost every period. Note that for modelling the time series each retailer-level promotion is input separately as not to assume that all have a similar effect.

4. Empirical evaluation

To evaluate the performance of the multivariate MAPA all SKUs available to us from the cider brand of our case study are used. For each time series the last 18 weeks are withheld. This test set will be used to assess the out-of-sample forecasting performance of the method against established benchmarks for horizons \( t+4, t+8 \) and \( t+12 \), which are relevant for decision making for the manufacturer in our case study. We employ a rolling origin evaluation scheme to collect as many error measurements as possible for the three different forecast horizons in the test set. Forecasts are produced from each origin (week) of the out-of-sample period for the target forecast horizons and the performance is evaluated for each period (for more details on rolling origin evaluation see Tashman (2000)). The rest of the data is used for fitting the

Table 3

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>33.00</td>
</tr>
<tr>
<td>Mean</td>
<td>4038.81</td>
</tr>
<tr>
<td>Median</td>
<td>1999.67</td>
</tr>
<tr>
<td>Maximum</td>
<td>28,151.75</td>
</tr>
<tr>
<td>Coefficient of variation</td>
<td>1.27</td>
</tr>
<tr>
<td>Skewness</td>
<td>2.56</td>
</tr>
<tr>
<td>Number of promotions</td>
<td>3.25</td>
</tr>
<tr>
<td>Periods under promotion</td>
<td>61.78</td>
</tr>
</tbody>
</table>

Fig. 1. Flowchart of calculation steps for each temporal aggregation level of MAPA with exogenous variables.

they will permit us to evaluate the gains in performance achieved
which has been shown to improve over the performance of
Although both
introduced by Kourentzes et al. (2014) is also used as a benchmark,
practice and research (Gardner, 2006). The univariate
calculation of percentage metrics impossible. Furthermore, with
models should outperform it in order to justify their additional
level promotional models. Here we implement as a benchmark the
principal components of the promotional dummies, following the
multivariate benchmarks are used. In the literature there
retaining only the first principal component at each
gains in performance achieved by their multivariate counterparts, if any. Due to limited estima-
tion sample we consider temporal aggregation up to approxi-
mately the monthly level, \( K=4 \).

Two multivariate benchmarks are used. In the literature there
most number of promotional models at SKU level (for
promotional models at brand level due to the different data
set and used to produce all the forecasts in the test set.
We track the forecasting bias and accuracy for the weekly
manufacturer sales of each SKU and horizon using the scaled Error
\( sE_t \) and the scaled Absolute Error \( sAE_t \), which are defined as
\[
sE_t = \frac{y_t - f_t}{n^{-1} \sum_{i=1}^{n} y_i}, \quad (10)
\]
\[
sAE_t = \frac{|y_t - f_t|}{n^{-1} \sum_{i=1}^{n} y_i}, \quad (11)
\]
where \( y_t \) and \( f_t \) are the actual and forecasted values at period \( t \) and
the denominator is the mean of the time series. Both error metrics
are scale independent and allow summarising the forecasting
performance across the different time series. These errors are used
instead of more common percentage metrics, such as the Mean
Absolute Percentage Error, because the time series used in this
study contain several periods of zero sales which makes the
Calculation of percentage metrics impossible. Furthermore, with
traditional percentage errors periods with very low demand will
have disproportionate impact. Both scaled metrics used here can
be approximately interpreted as percentage forecast bias and error
(Kolassa and Schütz, 2007). The error metrics are summarised across
origins and time series by calculating the mean, resulting in the
scaled Mean Error (sME) and scaled Mean Absolute Error (sMAE). For sME positive values imply under-forecasting and negative values imply over-forecasting.

The performance of multivariate MAPA is assessed using a
number of benchmarks. First, a random walk forecast is used that
will be referred to as \( \text{Naïve} \). As the random walk is a very simple
model that requires no parameter estimation, any more complex
complex models should outperform it in order to justify their additional
complexity. Next, univariate \( ETS \) is used as a benchmark. Exponen-
tial smoothing is commonly used in business forecasting and
has been found to be relatively accurate and reliable, both in
practice and research (Gardner, 2006). The univariate \( MAPA \)
introduced by Kourentzes et al. (2014) is also used as a benchmark,
which has been shown to improve over the performance of \( ETS \).
Although both \( ETS \) and \( MAPA \) are not capable of modelling the
available promotional information, they are useful benchmarks as
they will permit us to evaluate the gains in performance achieved

\[\text{Fig. 2. Sales and promotions of one SKU from the case study.}\]

5. Results

Table 4 presents the results of the empirical evaluation across
all available SKUs for the cider brand of the case study, in terms of
sME and sMAE. The best performance for each error metric and
horizon is highlighted in boldface. Values in parentheses represent
medians across all SKUs, while the rest represent mean errors
across SKUs. The last column in the table provides the mean rank
of each method across SKUs and target forecast horizons for sME
and sMAE. A method with rank of 1 is interpreted as being the best
for every single case, while with rank of 6 it is always the worst.

Overall, MAPAx is the best performer both in terms of average
bias and error. It is interesting to evaluate the improvements
achieved by extending the models to use promotional information.
To support the comparisons, Fig. 3 visualises the mean results
presented in Table 4. Focusing on the univariate \( ETS \) and \( MAPA \)
the latter performs better for horizons \( t+4 \) and \( t+8 \) and the former
for \( t+12 \). On average \( MAPA \) improves over \( ETS \), in accordance to
the findings by Kourentzes et al. (2014). This holds both in terms of
forecast bias and error. Interestingly for long term forecasts, \( t+12 \), the \( \text{Naïve} \) has similar errors to both \( ETS \) and \( MAPA \), attesting to the
difficulty of producing accurate forecasts for the time series of our
case study. When considering median errors the \( \text{Naïve} \) is more
accurate for long term forecasts that both \( ETS \) and \( MAPA \). In terms of
bias the \( \text{Naïve} \) is always less biased.

When promotional information is included in the models their
performance increases substantially. Starting from the benchmark
Regression the forecast errors drop over the univariate models for horizons t+4 and t+8. For horizon t+12 the performance is again relatively poor, being similar to the Naïve. In terms of bias Regression is consistently the most biased. ETSx performs overall better than Regression, with the latter having lower errors only for the t+4 forecast horizons. In terms of median errors ETSx is always better than Regression and Naïve. It should be noted that in many ways ETSx incorporates several aspects of Regression, such as using principal components for the promotional information and capturing the time series dynamics. The primary difference between them is the way that the time series structure is identified and modelled, with ETSx being arguably simpler. Furthermore ETSx has substantial performance improvements over its univariate counterpart, demonstrating the benefits of including the promotional information.

MAPAx exhibits the biggest improvement over its univariate counterpart. The observed improvements demonstrate again the benefit of including promotional information in the models. Considering mean sME across SKUs MAPAx gives the least biased predictions, with substantial differences over ETSx for all forecast horizons. However, when medians are considered MAPAx is second after the Naïve for longer horizons (t+8 and t+12). Nonetheless, it still exhibits substantial improvements over all other methods and in particular Regression and ETSx that are capturing the promotional information. In terms of accuracy MAPAx has lower errors than both multivariate benchmarks, considering either mean or median errors across SKUs. Overall, considering the mean errors of the best performing benchmark for each horizon, MAPAx is about 51.6% less biased and has about 12.0% lower forecast errors.

Focusing on the mean ranks provided in Table 4, MAPAx achieves the best ranking for both sME and sMAE, demonstrating its consistent performance. The value of the promotional inputs is highlighted in the mean ranks of sMAE, where Regression and ETSx rank better than the univariate benchmarks. This demonstrates that the promotional inputs are useful for improving forecasting accuracy. Note that Naïve performs better than the univariate ETS and MAPA providing evidence of the difficulty of producing accurate baseline forecasts for the time series of the case study.

In many ways the relative performance of MAPAx in comparison to ETSx replicates the pattern between the univariate MAPA and ETS. Using multiple temporal aggregation consistently results in better performance over conventionally modelled exponential smoothing.

Therefore the superior performance of MAPAx is a result of the combination of the quality of the forecasting method and the quality of information available to it. These results were found to be consistent using other error metrics, such as scaled Mean Squared Error.

### 6. Discussion

Considering the conventional ETS if there are strong promotional effects, as it is true in our case study, the parameter estimates and even the selected model, as it is conditional on the
estimated parameters, may be biased. By introducing the promotional information in the ETSx model this effect is mitigated. However, there is still uncertainty in the parameter identification and model selection, due to available sample and sampling frequency issues (Kourentzes et al., 2014) or the inherent limitations of information criteria for model selection (Kolassa, 2011). The original MAPA was developed with the motivation of addressing the later issues. The time series is modelled at multiple temporal aggregation levels, thus at each level filtering the higher frequency components of the time series, allowing to estimate lower ones appropriately. Combining the estimates across the different aggregation levels results in robust information.

The effect of including promotional information at low levels of aggregations is apparent, as at this level their effect will be stronger. However, at higher aggregation levels the size of the effect of promotions at each period becomes smaller and one could expect that it is no longer as important. Eq. (1) shows that aggregation acts as a moving average, therefore although the effect per period will be smaller, the promotion now is expanded to neighbouring periods. This results again in an important overall effect, which unless modelled explicitly it is bound to bias parameter estimates and potentially even the selection of the model for each aggregation level.

Let us consider the example of a simulated sales series with promotions. Fig. 4 plots the sales series at various temporal aggregation levels. The promoted periods are noted with black bars at the lower part of the plots. Furthermore, for comparison, the simulated sales as if there were no promotions are plotted with a dotted line. Observe that the sales do not contain any trend or seasonality and therefore the only non-promotional structure is the level. Single exponential smoothing would be appropriate to model this effect is mitigated.

<table>
<thead>
<tr>
<th>Aggregation level</th>
<th>ETS – sales without promotions</th>
<th>ETS – sales with promotions</th>
<th>ETSx</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.046</td>
<td>0.093</td>
<td>0.054</td>
</tr>
<tr>
<td>2</td>
<td>0.137</td>
<td>0.215</td>
<td>0.128</td>
</tr>
<tr>
<td>4</td>
<td>0.252</td>
<td>0.146</td>
<td>0.198</td>
</tr>
<tr>
<td>6</td>
<td>0.230</td>
<td>1.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Fig. 4. Sales with promotions at temporal aggregation levels 1, 2, 4 and 6. Periods under promotion are noted with black bars at the lower part of the plots.

The last column of Table 5, ETSx, lists the parameters for the simulated sales series without promotional effects, corresponding to the dotted line in Fig. 4. The second column, ETS – sales with promotions, lists the parameter estimates when the sales series includes the peaks due to promotions. Note that in all cases the parameters are substantially different, demonstrating the impact of the promotions not captured on model fit. Now the level of the time series model is wrongly estimated and the accuracy of the forecasts is expected to be poor. Interestingly, this is true even for high aggregation levels that the promotional uplift is seemingly small.

The last column of Table 5, ETSx, lists the parameters of ETSx that model promotions as an additional input. Although the parameters are not identical to those of the first column, they are much closer demonstrating the advantage of including such information when available, even when its effect is relatively smoothed due to the temporal aggregation. Now the level dynamics are captured more accurately and the resulting forecasts are expected to perform better.

It is also interesting to note that there may be cases, as is for aggregation level 6 in this example, where the smoothing parameter is apparently misestimated. In these cases the forecasts of ETSx will be of poor quality, while the ones of MAPAx that combine the estimates from multiple aggregation levels will be better. A similar observation can be made with regards to the fitted model at each aggregation level as argued by Kourentzes et al. (2014). Potentially models fitted at some aggregation levels may be
misspecified in their form. By using multiple temporal aggregation, as the outputs of the various models at the different aggregation levels are combined, we do not rely on a single one, which might have been misspecified. This is a useful property for practical implementations of MAPAx, as it makes it robust against mis specification at some aggregation levels and crucially at the original time series, which is the only view of the data conventional time series modelling focuses on. Therefore MAPAx provides a reliable automatic forecasting procedure that includes external variables.

MAPAx is useful for practice as it incorporates explanatory variables in an automated way, such as promotions, and provides reliable and accurate predictions. This makes it useful for supply chain forecasting, where typically a large number of SKUs need to be forecasted for inventory and planning purposes. Therefore it is interesting to consider the implications of using MAPAx for such cases. Stock calculations are typically based on the following formula: expected demand over lead time plus demand uncertainty over lead time. The first quantity is essentially the expected value of the forecast, which ideally should have a forecast bias of zero, otherwise the expected value of the forecast does not match the expected value of the demand. The second quantity, which is essentially the safety stock, is a pre-set percentile of the distribution of forecast error size, which is often approximated as the mean squared error of the forecast multiplied by some factor to account for the target service level and the lead time. Therefore, a good forecast for such purposes should have small bias and magnitude of forecast errors. Table 4 provided evidence of the superior performance of MAPAx in our case study, both in terms of forecast bias and error, demonstrating that it has desirable behaviour and out-performs the various benchmarks in both dimensions.

7. Conclusions

This paper extended the univariate Multiple Aggregation Prediction Algorithm that was recently proposed in the literature, and has been shown to have good performance both for fast and slow moving items, to the multivariate case. To demonstrate the performance and the efficacy of the proposed formulation we investigated the usage of the MAPAx to model the demand of SKUs of a popular cider brand in the UK, including promotional information.

MAPAx was found to outperform all benchmarks, which included a recently proposed in the literature SKU-level promotional model and exponential smoothing with regressor inputs, appropriately preprocessed. In particular, the main differences between ETSx and MAPAx is the use of multiple temporal aggregation levels, which provides the latter approach its superior performance and also makes it robust against model misspecification. The overall better performance of MAPAx over its exponential smoothing counterpart follows similar findings for the univariate case in the literature, providing evidence of the merits of this alternative approach to forecasting time series, based on modelling time series at multiple temporal aggregation levels.

In the discussion we attempted to highlight the implications of using MAPAx for baseline forecasting in a supply chain context. Future research should explore in detail the inventory implications of using MAPAx when external variables are available and important for capturing the demand behaviour. Another aspect of using MAPAx in a supply chain context that warrants further research is the interaction of human experts with the statistical forecast. As forecasting methods become more complex, here to introduce promotional information at SKU level, their transparency to experts is reduced, complicating the adjustment process.

Acknowledgements

We would like to thank the anonymous reviewers for their comments that helped us improve the paper. Any remaining errors are the responsibility of the authors.

References


