Demand forecasting by temporal aggregation: using optimal or multiple aggregation levels?

Nikolaos Kourentzes[a,*], Bahman Rostami-Tabar[b], Devon K. Barrow[c]

[a]Lancaster University Management School
Department of Management Science, Lancaster, LA1 4YX, UK
[b]Cardiff Business School, Cardiff University
Cardiff, CF10 3EU, UK
[c]School of Strategy and Leadership, Faculty of Business and Law
Coventry University, Coventry, West Midlands, CV1 5FB, UK

Abstract

Recent advances have demonstrated the benefits of temporal aggregation for demand forecasting, including increased accuracy, improved stock control and reduced modelling uncertainty. With temporal aggregation a series is transformed, strengthening or attenuating different elements and thereby enabling better identification of the time series structure. Two different schools of thought have emerged. The first focuses on identifying a single optimal temporal aggregation level at which a forecasting model maximises its accuracy. In contrast, the second approach fits multiple models at multiple levels, each capable of capturing different features of the data. Both approaches have their merits, but so far they have been investigated in isolation. We compare and contrast them from a theoretical and an empirical perspective, discussing the merits of each, comparing the realised accuracy

[*]Correspondance: N Kourentzes, Department of Management Science, Lancaster University Management School, Lancaster, Lancashire, LA1 4YX, UK. Tel.: +44-1524-592911
Email address: n.kourentzes@lancaster.ac.uk (Nikolaos Kourentzes)

gains under different experimental setups, as well as the implications for business practice. We provide suggestions when to use each for maximising demand forecasting gains.

Keywords: Forecasting, demand planning, temporal aggregation, model selection, exponential smoothing, MAPA

1. Introduction

Demand forecasting plays a crucial role in the operations of modern organisations (Fildes et al., 2008; Syntetos et al., 2016). It supports a variety of business decisions, from operational, to tactical, to strategic level, such as capacity planning (Miyaoka and Hausman, 2008), resource planning (Barrow, 2016; Jalal et al., 2016), advertising and promotional planning (Trapero et al., 2014; Ma et al., 2016), demand planning (Trapero et al., 2012; Syntetos et al., 2015), analysing competition effects (Merino and Ramirez-Nafarrate, 2016), tactical production planning (Sagaert et al., 2017), among others. Accordingly, practitioners need to define the forecast objective in terms of forecast horizon and time bucket (e.g. daily, weekly, monthly, etc.), so as to support the appropriate decisions.

An important assumption is that the level of required forecasting matches the level of available collected data. However, often this not true. For example, in many organizations, managers from several departments are involved in forecast generation and adjustment, that supports decisions for production, inventory management, logistics, procurement, and others (Lapide, 2004); with each function having different decision horizons. For example, budget forecasts are not required at the, typically, weekly resolution of inventory management, and refer to much longer horizons than the latter.

As a remedy the original data series can be aggregated over time (temporal aggregation, TA) to align the decision parameters with the forecast modelling, or alternatively disaggregated. Recently there has been a resurgence in researching TA for forecasting. In the past the research had mostly focused in modelling macroeconomic time series, but current work has demonstrated its usefulness for forecasting business time series, and in particular for the purpose of demand forecasting to support decision-making in operations management (Babai et al., 2012; Kourentzes and Petropoulos, 2015; Boylan and Babai, 2016). Using TA a time series is modelled at a pre-specified aggregation level, instead of its original sampling frequency. Forecasts are then created at the aggregate level, which may be disaggregated to the original frequency, if so needed. The motivation for using TA is that it smooths the original series, removing noise and even some of its component, simplifying the generation of forecasts, which is desirable in itself (Green and Armstrong, 2015). The exact effects depend on the selected aggregation level, a critical consideration for the effectiveness of TA.

To this end, the econometric literature has explored the effect of TA, mainly on AutoRegressive Integrated Moving Average (ARIMA) processes (Silvestrini and Veredas, 2008), providing some evidence of the benefits and caveats of the practice, while more recent forecasting research has helped identify analytically the optimal aggregation level for a small number of processes, under specific modelling conditions (Rostami-Tabar et al., 2013, 2014). Nonetheless, general guidelines for how to best select the aggregation level do not exist, and this introduces substantial uncertainty in the modelling process. This has lead Kourentzes et al. (2014) to propose using multiple TA levels instead of a single one. In this case, modelling happens at multiple levels and the output is a combined forecast.

Therefore, although there is a strong theoretical and empirical evidence that TA can be beneficial to forecasting, there is no consensus as to how best perform it. The two alternative schools of thought recommend from the one hand to use a single optimal TA level, and from the other hand to use multiple levels, since the identification of a single level is problematic. The aim of this paper is threefold: (i) we contrast the two approaches both from a theoretical and empirical perspective; (ii) we benchmark these against heuristic based alternatives; and (iii) provide additional evidence of the usefulness of TA for demand forecasting over traditional time series modelling, at the original sampling frequency.

We find that overall TA is beneficial for demand forecasting over conventional time series modelling. Each school of thought offers different advantages and has different limitations. The main limitation of identifying an optimal single aggregation level is that it assumes knowledge of the demand process at both the original and the aggregate level, with the obvious implications for practice. On the other hand, using multiple levels is particularly robust to model uncertainty and is found to provide accuracy improvements for wide number of cases. However, the forecast is suboptimal by design in the strict sense of mean squared error fit. Finally, we translate these findings to implications for business forecasting practice.

The rest of the paper is organised as follows: section 2 provides an overview of the use and developments of TA in demand forecasting and section 3 describes the two alternative approaches in using TA. Section 4 describes the datasets used and the setup of our evaluation, while section 5 presents the results, followed by concluding remarks in section 6.

2. Temporal aggregation in business forecasting

Non-overalapping TA can be seen as a filter of high-frequency components of the time series. As we aggregate, low frequency components will dominate and depending on the level of aggregation higher frequency components will become weaker or vanish altogether. For example, consider a monthly seasonal time series that is aggregated to an annual series. The high frequency seasonal component is filtered, while the observed variance of the time series will be mostly due to the trend/cycle component.

In the econometric literature TA has been researched for several decades and the focus has mainly been on its effects on ARIMA processes. The key theoretical results can be summarised as follows: (i) TA reduces the number of available observations; hence causing loss of estimation efficiency; (ii) the dynamics of the underlying ARIMA process become more complicated, mainly due to the moving average component; and (iii) the identifiable ARIMA converge to relatively simple IMA processes, often IMA(1,1) (Wei, 1978; Rossana and Seater, 1995). The literature provides evidence of accuracy gains of forecasting directly using temporally aggregated data, rather than aggregating forecast from disaggregate series (Silvestrini and Veredas, 2008).

2.1. Temporal aggregation at a single level

More recently there has been substantial research on TA for business forecasting and supply chain management. Nikolopoulos et al. (2011) recommend using TA for modelling and forecasting intermittent time series in a supply chain context. Their main motivation is to avoid modelling the intermittency at the sampling frequency directly and instead model the series with conventional forecasting methods, once the intermittency has been reduced substantially. They demonstrate that on average TA provides accuracy improvements. This finding has been validated several times in the context of intermittent demand forecasting (Babai et al., 2012; Petropoulos and Kourentzes, 2014a). It is important to note that Nikolopoulos et al. (2011) do not provide a conclusive solution with regards to the identification of the appropriate TA level. Instead, they recommend a heuristic that is meaningful for inventory management: aggregate to the level that corresponds to the lead time plus review period. Petropoulos et al. (2016) demonstrate that some intermittent demand forecasting methods, such as Croston's method, can be interpreted as special cases of TA and propose various alternative setups of TA, which in turn can reduce the variability of the the non-zero demand or the inter-demand intervals and demonstrate benefits for forecast accuracy.

Spithourakis et al. (2011) extended the work by Nikolopoulos et al. (2011) to fast moving demand data, validating that TA leads to forecast accuracy improvements. Jin et al. (2015) utilise a large set of paired order and point-of-sale data in a retail supply chain to examine the impact of TA on forecast accuracy. They show that it increases forecast accuracy and reduces computational intensity of forecast generation. Luna and Ballini (2011) use TA to predict daily time series of cash withdrawals and find similar or better forecast accuracy to modelling the daily series directly.

Exploring further the impact of TA for demand forecasting Rostami-Tabar et al. (2013) and Rostami-Tabar et al. (2014) derive analytically the optimal aggregation level when the underlying demand process follows AutoRegressive AR(1), Moving Average MA(1), AutoRegressive Moving Average ARMA(1,1) and exponential smoothing is used to produce the forecasts. The choice of forecasting model is motivated by the problem context, where single exponential smoothing is the norm for producing demand forecasts for non-trended and non-seasonal time series. They determine analytically the conditions under which non-overlapping TA outperforms the traditional modelling approach. Using the optimal TA levels, they demonstrate accuracy improvements and show that TA's superiority is a function of the demand process parameters, forecasting method parameters, and aggregation levels. However there are no expressions for more complex ARIMA forms or different forecasting models. This is an important limitation given the prevalence of seasonal and trended demand series in practice. Moreover, it should be noted that the ARIMA type processes can only represent fast moving items. For slow moving items, the consideration of other process such as Integer ARMA (INARMA) processes is relevant (Mohammadipour and Boylan, 2012).

2.2. Multiple temporal aggregation levels

The majority of the aforementioned literature had taken the approach to explore how to best model the time series at a single aggregate level instead of the original that the time series was sampled. Kourentzes et al.

(2014) argue that there are two concerns with this approach: (i) for the majority of time series we do not have a way to identify the optimal TA level; and (ii) even if there was one, due to sampling, there is a substantial uncertainty about the underlying process and the appropriate model to apply to a time series. Based on these, they recommend using multiple levels of TA and combining the separate forecasts. This approach not only benefits from managing the modelling risk, but also utilises the established gains of forecast combination (Barrow and Kourentzes, 2016; Blanc and Setzer, 2016). Kourentzes et al. (2014) provide empirical evidence to demonstrate gains over conventional forecasting. Since, modelling with multiple TA levels has been used successfully to intermittent demand, promotional modelling and inventory management (Petropoulos and Kourentzes, 2014a; Kourentzes and Petropoulos, 2015; Barrow and Kourentzes, 2016). An advantage of this approach is that it is not restricted to specific demand processes and allows for a wider variety of forecasting models.

Although both approaches for using TA (single optimal and multiple levels) have demonstrated forecasting gains, so far there is no comparative study between the two. Arguably using multiple levels is suboptimal in the strict sense (see section 3.2), yet more flexible and widely applicable. On the other hand using a single optimal level is expected to provide better performance, assuming that the identification of the underlying demand process is reliable.

3. Temporal aggregation approaches

In this section we briefly outline the two alternative approaches for using non-overlapping TA for demand forecasting. In general, given a time series with observations y_t and $t = 1, \ldots, n$, non-overlapping TA can be performed as:

$$y_i^{[k]} = \sum\nolimits_{t=1+(i-1)k}^{ik} y_t, \tag{1}$$

where k is the aggregation level. We denote the temporally aggregated time series with a superscript $[k]$. Eq. (1) implies that the first $n - \lfloor n/k \rfloor k$ observations of the time series may be ignored in the construction of the aggregated series, depending on k. Naturally, the resulting time series has less observations than the original one.

Note that Eq. (1) acts as a moving average on the original series, filtering high frequency components. To exemplify the effect of this we illustrate it in Figure 1. A series is sampled at a monthly frequency and exhibits a clear repeating seasonal pattern and an outlier at the beginning of 2015. Both outlier and seasonality are high-frequency components. As the series is aggregated at a quarterly level ($k = 3$) the effect of the outlier is mitigated, but the seasonality remains, although it is smoother. At the annual aggregation level ($k = 12$) the seasonality is fully removed and a slight trend that is present in the series becomes apparent. This trend was not observable at the lower aggregation levels. The forecasting literature has taken advantage of this effect of TA to improve the quality of forecasts (for example, Pedregal and Trapero, 2010; Kourentzes et al., 2014).

3.1. Identifying the optimal temporal aggregation level

Rostami-Tabar et al. (2014) evaluate analytically the impact of non-overlapping temporal on demand forecast accuracy. They assume that the underlying series follow an ARMA(1,1) and its special cases AR(1) and

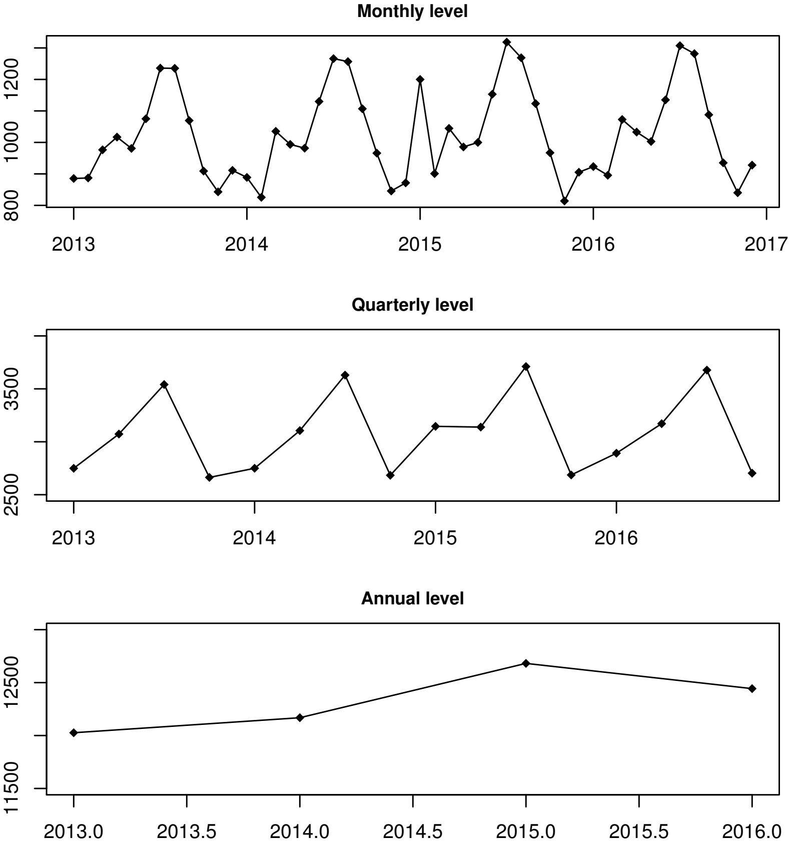

Figure 1: Original time series sampled at a monthly frequency and aggregated at quarterly and annual levels.

MA(1) that can be mathematically written in Eq. (2), Eq. (3) and Eq. (4) respectively:

$$y_t = C + \phi y_{t-1} - \theta \epsilon_{t-1} + \varepsilon_t, \tag{2}$$

$$y_t = C + \phi y_{t-1} + \varepsilon_t, \tag{3}$$

$$y_t = C - \theta \epsilon_{t-1} + \varepsilon_t, \tag{4}$$

where C is constant, $|\phi| < 1$ is the autoregressive parameter, $|\theta| < 1$ is the moving average parameter and ε_t is an independent random variable for underlying demand series in period t, normally distributed with zero mean and variance σ^2.

The forecasting method considered is the Single Exponential Smoothing (SES). Using SES, the aggregate forecast of demand is calculated as follow

$$\hat{y}_i^{[k]} = \alpha y_{i-1}^{[k]} + (1-\alpha)\hat{y}_{i-1}^{[k]}, \tag{5}$$

where $0 < \alpha < 1$ is the smoothing parameter used at the aggregated demand series. Forecast accuracy of the aggregated demand series is calculated using Mean Square Error (MSE):

$$\text{MSE} = \text{Variance}\left(y_i^{[k]} - \hat{y}_i^{[k]}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{[k]} - \hat{y}_i^{[k]}\right)^2 \tag{6}$$

The optimal aggregation level minimises the forecast MSE of the aggregated demand for each demand process under consideration, when SES is used as the forecasting method. These are analytically identified. The aggregation level that leads to the most error reduction is determined by comparing variances at the aggregate level. This approach works as follows: first, buckets of aggregated demand are created at the aggregation level k, using Eq. (1); then SES, Eq. (5), is applied to these aggregate series and finally the variance of aggregated forecast error is calculated using Eq. (6).

According to Rostami-Tabar et al. (2014), the MSE of aggregate forecasts can be derived as follows for ARMA(1,1), AR(1) and MA(1) demand processes:

$$\begin{aligned}
\mathrm{MSE}_{ARMA} &= \frac{2\sigma^2 \left(k\left(1-2\phi\theta+\theta^2\right)+\left(\phi-\theta\right)\left(1-\phi\theta\right)\left(\sum_{i=1}^{k-1} 2\left(k-i\right)\phi^{i-1}\right)\right)}{\left(2-\alpha\right)\left(1-\phi^2\right)} \\
&+ \frac{2\sigma^2\alpha\left(\sum_{i=1}^{k}\left(i\phi^{(i-1)}\right)+\sum_{i=2}^{k}\left(i-1\right)\phi^{(2k-i)}\right)\left(\phi-\theta\right)\left(1-\phi\theta\right)}{\left(2-\alpha\right)\left(1-\phi^k+\alpha\phi^k\right)\left(1-\phi^2\right)},
\end{aligned} \tag{7}$$

$$\begin{aligned}
\mathrm{MSE}_{AR} &= 2\sigma^2\left(\frac{k+\sum_{i=1}^{k-1} 2\left(k-i\right)\phi^{i}}{\left(1-\phi^2\right)\left(2-\alpha\right)}\right) \\
&- \frac{2\sigma^2\alpha\left(\sum_{i=1}^{k}\left(i\phi^{(i-1)}\right)+\sum_{i=2}^{k}\left(i-1\right)\phi^{(2k-i)}\right)}{\left(2-\alpha\right)\left(1-\phi^k+\alpha\phi^k\right)\left(1-\phi^2\right)},
\end{aligned} \tag{8}$$

$$\mathrm{MSE}_{MA} = \frac{\sigma^2\left(2k\left(1+\theta^2\right)-2\left(k-1\right)\theta+2\alpha\theta\right)}{2-\alpha}. \tag{9}$$

In order to obtain the optimal aggregation level, k^* for each process, the first derivative of Eqs. (7), (8) and (9) need to be calculated. The calculation of the first derivative of Eq. (9) with respect to k shows that MSE_{MA} is a decreasing function of k. A higher aggregation level results in a lower MSE_{MA}. Therefore, for MA(1) demand process the optimal aggregation level is the highest value in the considered range.

However, the calculation of the first derivative for Eqs. (7) and (8) is infeasible. Consequently, to determine the optimal aggregation level a numerical investigation is conducted across the range of $2 \leq k \leq K$, where K is the maximum value of the considered aggregation levels, and the aggregation level with the minimum value of MSE is selected.

3.2. Multiple aggregation prediction algorithm

The Multiple Aggregation Prediction Algorithm (MAPA) proposed by Kourentzes et al. (2014) models a time series at multiple TA levels to achieve better estimation of the various time series components. As TA strengthens and attenuates different components of the series, the multiple views permit capturing these better.

Modelling a time series with MAPA can be seen as a three step procedure. In the first step, multiple aggregated series are constructed from the original time series, creating K different series. It is recommended to aggregate up to the annual level, so as to filter any seasonal components fully, thus enabling to better capture the cycle/trend (Petropoulos and Kourentzes, 2014b). To avoid any complications introduced by changes in the scale of the aggregate time series Eq. (1) is divided by k, changing the summation to an average.

In the second step, each series is modelled independently. Although in theory one could use any forecasting method to this purpose, the authors demonstrated MAPA with state space exponential smoothing. The idea is that at each aggregation level, different components of the time series will be easier to capture. The underlying structure of the time series is constant, yet what is observable changes depending on the TA level we focus on. Furthermore, since modelling the time series does not depend on identifying a single model, which may be appropriately done or not, MAPA mitigates modelling uncertainty.

In the third step, the outputs of the various models are combined in a single forecast. In contrast to conventional forecast combination, MAPA prescribes to combine each state (or component) of the model separately. First, the various fitted states are extrapolated into the future. Then these are all oversampled appropriately to bring hem to the original sampling frequency. For example if at aggregation level $k = 2$ the predicted values of a state are

$(2, 4)$, these will be returned to the original sampling frequency as: $(2, 2, 4, 4)$, where each value is repeated k times. For the case of exponential smoothing that states may interact in an additive or multiplicative way, Kourentzes et al. (2014) provide formulas to transform them all to additive. Subsequently all estimated versions of a state, across all TA levels, are linearly combined and the combination may be weighted or not. For example, in the context of exponential smoothing a trend state may be estimated at each TA level. These are then linearly combined to a single trend state. When a trend state is missing it is assumed to be zero. The reasoning behind this choice is that if at an aggregation level no trend is estimated then this is evidence that the estimate trends at other levels may be wrong and should be damped, which is done through the combination. This is repeated for the remaining states. An exception is done for any seasonal states, where these would be impossible to estimate, for example when modelling annual data. Finally, the combined states are added to provide the final forecast.

This counter-intuitive approach to combination is necessary due to the different information that each state may encode. Consider for example combining the forecasts of a seasonal model constructed on a monthly time series and its non-seasonal counterpart constructed at an annual level series. Simply combining the two forecasts would result in a seasonal part half that is the size of what it should be. On the other hand, combining per state overcomes this problem.

The final MAPA forecast reconciles information from all TA levels, allowing the model to provide a more holistic representation of the high and low frequency components present in the time series, whereas when modelling at a single level modelling focuses mostly at the components that capture most of the variance of a time series. TA is therefore used as a device to transform information available on the original time series in different ways.

MAPA will be by definition suboptimal for any single TA level and obviously for the original time series in a mean squared fit error sense, as it reconciles information from different levels. In conventional model building one has to select the appropriate model and estimate its parameters. The latter typically happens either with maximum likelihood estimation or by minimising some relevant squared in-sample error (Gardner, 2006), so that the resulting parameters are optimal for the given sample. Similarly the model is selected so as to minimise some similar, typically penalised, criterion, such as the Akaike Information Criterion that has been showed to perform well for this task (Burnham and Anderson, 2002; Hyndman et al., 2008). With TA the same process is applied at the aggregated time series, and the resulting model is optimised for that view of the data. In the case of MAPA, this is repeated at every aggregation level, and once the forecasts are combined, the resulting final forecast is not optimal, in the sense described above, for any individual aggregation level and under-fits to all of them. Furthermore, as the selected model at each aggregation level may vary, there is no guarantee that the combined final forecast will reflect the 'best' model for any aggregation level. Hence, MAPA is suboptimal for any single TA level, and instead attempts to provide a holistic forecast from all aggregation levels (Petropoulos and Kourentzes, 2014b).

This has interesting implications for the choice of the appropriate forecasting model. While conventional forecast model building implies that the ‘best’ model is chosen, under MAPA it is explicitly understood that none of the fitted models is ‘best’ and the notion of model selection is only appropriate locally for each TA level. A generalisation of MAPA has been proposed by Athanasopoulos et al. (2017) that can operate even when the available predictions are not the product of statistical models.

4. Empirical evaluation

In this section we outline the setup of the empirical evaluation which is used to assess the performance of TA and of the two alternatives schools of though for demand forecasting purposes.

4.1. Data

We conduct the empirical evaluation using both real and simulated time series. The real time series have several advantages, capturing the complexity of real applications and having realistic sample size, therefore permitting us to draw direct conclusions for the usefulness of the TA approaches for business forecasting. However, the underlying data generating process is unknown, which can limit some aspects of the empirical evaluation. More specifically, for the selection of the optimal TA level, as outlined in section 3.1, there are derivations only for some types of data generating processes. By controlling that with simulated data, we can assess the performance of the method when the data generating process is correct, approximate or inappropriate for the existing derivations. Therefore, using simulated time series, although any insights are limited by the simplicity of the series, allows us to investigate the performance of the competing TA approaches in a controlled setup and explore the conditions under which each approach performs best, as well as the sensitivity of selecting a single optimal aggregation level when there are deviations from the expected data generating process.

We use two real datasets.The first is from a major UK fast moving consumer goods manufacturer and has 229 time series, of 173 weekly observations each. From these the last 43 weeks are used as a test set. All time series are non-seasonal. Forecasts for these time series are essential for inventory management purposes. The second dataset contains 133 weekly time series tracking call volumes of different types in a call centre of a major UK media company. The series range from 108 to 169 weeks. The last 43 time series are kept as a test set. Forecasts for this case are necessary for workforce planning in the call centre.

For the manufacturer dataset the augmented Dickey-Fuller test finds that 90% of the time series are stationary, suggesting that ARIMA(p,0,q) is appropriate, while for the remaining 10% of the series first differences are adequate, therefore suggesting ARIMA(p,1,q), where p and q are the orders of the autoregressive and moving average parts respectively. For the call centre dataset the respective values are 26% and 74%. Note that for this particular dataset any series that exhibited seasonality have been de-seasonalised prior to the experiment. We mirror these ARIMA orders to the simulated time series.

We simulate ARIMA(p,d,q) processes with $p = (0, 1, 2)$, $d = (0, 1)$ and $q = (0, 1, 2)$. For each process we generate 500 series. Each time series is 100 observations long, from which 60 are used as fitting sample and the rest are retained as a test set. The values of the autoregressive and moving average coefficients are randomly sampled for each series from a uniform distribution, while ensuring that any resulting process is stable and invertible.

Note that we restrict the empirical evaluation to non-seasonal time series. This is done to facilitate the comparison between the two TA approaches, as there are currently no analytical formulas to derive the single optimal aggregation level for the methodology presented in section 3.1 for seasonal time series. This is not a restriction when using multiple TA levels, as in this case knowledge of the underlying process is not required.

4.2. Evaluation scheme and metrics

The forecast horizon is set to be $h = 13$ periods, reflecting a quarter in weekly data. We use a rolling origin evaluation scheme. Using the complete training set the first forecast is created for the first 13 periods of the test set. Then we roll the forecast origin forward, including one additional observation in the training set. Models are re-optimised and new forecasts are produced. The process is repeated until forecasts are generated for the last 13 periods of the test set. The rolling origin evaluation scheme has the advantage that it permits sampling forecast errors multiple times and it mitigates the effect of outliers in either the forecast origin or the test period (Tashman, 2000).

We assess the performance of each forecast using a bias and an accuracy metric. Following the recommendations by Davydenko and Fildes (2013) we use relative error metrics, due to their ease of interpretation, good statistical properties and being scale independent, allowing us to summarise across different time series.

To measure accuracy we use the Average Relative Mean Absolute Error

(ARMAE, referred to as AvRelMAE by Davydenko and Fildes, 2013):

$$\text{ARMAE} = \sqrt[n]{\prod \left(\frac{\text{MAE}_i}{\text{MAE}_b}\right)},$$

where n is the number of time series over which the summary metric is calculated and MAE_i is the Mean Absolute Error of forecast i, calculated as:

$$\text{MAE} = m^{-1} \sum_{t=1}^{m} |y_t - \hat{y}_t|.$$

The MAE is calculated over m origins, for any given forecast horizon, and y_t and $\hat{y}_t$ are the actuals and forecasts respectively. MAE_b is the benchmark forecast, which in this case is the forecast produced on the original time series, without using any TA.

Davydenko and Fildes (2013) argue conclusively why ARMAE should be preferred over other accuracy metrics, such as the Mean Absolute Percentage Error or the Mean Absolute Scaled Error that are biased. ARMAE has both desirable statistical properties and is easy to interpret.

We define the Average Relative Absolute Mean Error (ARAME) to measure bias, as follows:

$$\begin{aligned} \text{ARAME} &= \sqrt[n]{\prod \left|\frac{\text{ME}_i}{\text{ME}_b}\right|}, \\ \text{ME} &= m^{-1} \sum_{t=1}^{m} (y_t - \hat{y}_t). \end{aligned}$$

ARAME is constructed following the arguments for ARMAE. Although it removes the direction information of bias, as measured by ME, it retains the magnitude of bias. Therefore it still allows us to assess whether a forecast is less or more biased. Removing the direction of bias is necessary so as to be able to summarise across time series using the geometric mean, which is appropriate for ratios. We avoid using bias metrics such as the Mean Percentage Error, because of the misleading interpretation it has due to calculation induced bias.

Both metrics are easy to interpret. If their value is under one then forecast i is more accurate, or has less bias, than the benchmark and vice-versa if their value exceeds one. Alternatively, calculating the difference of ARMAE or ARAME from 1 provides the percentage improvement over the benchmark used in the denominator.

4.3. Methods

To satisfy the aims of this analysis we consider three alternatives in terms of TA: (i) no aggregation, where modelling is done on the original time series; (ii) single aggregation level, where modelling is done on a temporally aggregated view of the series; and (iii) multiple aggregation levels, as prescribed by MAPA.

When using a single TA level we consider two options to identify the appropriate level. First, we use a simple heuristic proposed by Nikolopoulos et al. (2011) that prescribes the aggregation level to match the forecast horizon (more specifically that would be the lead time plus review period). Second we use the formulas for identifying the optimal aggregation level for AR(1), MA(1) and ARMA(1,1) processes (Rostami-Tabar et al., 2013, 2014). Obviously in a realistic situation the true model is always unknown. Therefore, we can only approximately find the most likely model, subject to sampling uncertainty. In practice we do this in the following way: initially for a time series all AR(1), MA(1) and ARMA(1,1) models are fitted and the best is chosen using the Akaike Information Criterion corrected for sample size (AICc, Burnham and Anderson, 2002). Subsequently we identify the optimal TA level as described in section 3.1.

Finally, the forecasts are generated using the Single Exponential Smoothing (SES, Hyndman et al., 2002):

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t,$$

where, α is a smoothing parameter between 0 and 1. Note that SES is equivalent to ARIMA(0,1,1), but it is widely used to model any 'level' time series, as defined in the exponential smoothing framework. This makes it appropriate to model series without seasonality or persistent trends (trend exponential smoothing is equivalent to ARIMA(0,2,2) model, Hyndman et al., 2008, with higher order of differencing than any of the series in our datasets). Furthermore, SES is the most widely used statistical method for demand forecasting (Gardner, 2006; Rostami-Tabar et al., 2013).

Combining the above, we obtain the following: (i) *Orig-SES*: using SES on the original time series; (ii) *Heur-SES*: using the heuristic to select the single aggregation level on which SES is fitted; (iii) *Opt-SES*: similar to Heur-SES, but at the optimal TA level; (iv) *MAPA-SES*: using MAPA, but restricted to use SES at all aggregation levels.

A main argument of MAPA is that the appropriate model for a time series is unknown and difficult to identify by investigating a single view of the time series. Furthermore, differences in the identified model are expected across the TA levels. Therefore we also use an unrestricted MAPA, where at each TA level any exponential smoothing model maybe selected, as it was originally proposed (Kourentzes et al., 2014). In order to have a comparable benchmark we also use exponential smoothing with model selection at the original time series. The selection of the models is based on AICc (Hyndman et al., 2008). This results in two additional forecasts over the four mentioned above: (v) *Orig-ETS*; and (vi) *MAPA*. Note that currently there are no derived analytical formulas for identifying the optimal aggregation level for other exponential smoothing forms than SES.

All forecasts are implemented in *R* (R Core Team, 2016) using the *forecast* package version 7.1 (Hyndman, 2016) and the *MAPA* package version 1.9.1 (Kourentzes and Petropoulos, 2016). Code for finding the optimal aggregation level, as per section 3.1, is implemented in function *get.opt.k* of the *TStools* package (Kourentzes and Svetunkov, 2016).

5. Results

Tables 1 and 2 provide the summary ARMAE and ARAME figures respectively. The summary values are geometric means across all origins, horizons and series. Each row corresponds to a particular subset of time series and the best performing forecast is highlighted in boldface. We also provide the geometric mean values across all ARIMA with zero, first and any differencing order: ARIMA($*$,0,$*$), ARIMA($*$,1,$*$) and ARIMA($*$,$*$,$*$). The last three rows refer to the results for the real datasets, with the last row being the average performance across both manufacturing and call centre sets.

We first focus on table 1 that provides the forecast accuracy. Orig-SES is used as the benchmark in the calculations and therefore has always an error equal to 1. Considering the simulated time series with no differencing, there is strong evidence that TA improves performance over forecasts produced on

Table 1: ARMAE

Demand	No aggregation		Single level		Multiple levels	
	Orig-SES	Orig-ETS	Heur-SES	Opt-SES	MAPA-SES	MAPA
Simulated series						
ARIMA(1,0,0)	1.000	0.979	0.974	0.975	0.972	**0.961**
ARIMA(0,0,1)	1.000	1.002	**0.960**	0.965	0.972	0.973
ARIMA(2,0,0)	1.000	0.971	0.986	0.983	0.973	**0.949**
ARIMA(0,0,2)	1.000	1.002	**0.969**	**0.969**	0.978	0.979
ARIMA(1,0,1)	1.000	1.001	0.966	0.971	0.964	**0.963**
ARIMA(2,0,2)	1.000	0.983	0.990	0.982	0.974	**0.953**
ARIMA(1,1,0)	**1.000**	**1.000**	1.439	1.223	1.062	1.004
ARIMA(0,1,1)	**1.000**	1.051	1.290	1.173	1.030	1.037
ARIMA(2,1,0)	1.000	**0.891**	1.444	1.207	1.062	0.916
ARIMA(0,1,2)	**1.000**	1.048	1.278	1.091	1.011	1.012
ARIMA(1,1,1)	1.000	**0.975**	1.349	1.191	1.056	0.990
ARIMA(2,1,2)	1.000	0.927	1.327	1.139	1.044	**0.922**
ARIMA(∗,0,∗)	1.000	0.989	0.974	0.974	0.972	**0.963**
ARIMA(∗,1,∗)	1.000	0.980	1.353	1.170	1.044	**0.979**
ARIMA(∗,∗,∗)	1.000	0.985	1.148	1.068	1.007	**0.971**
Real datasets						
Manufacturing	1.000	1.011	0.999	0.999	**0.992**	0.994
Call centre	1.000	1.005	1.121	1.080	0.980	**0.979**
Overall real sets	1.000	1.009	1.042	1.028	**0.987**	**0.987**

the original time series. Initially let us restrict the discussion to SES based forecasts. For ARIMA(1,0,0) the optimal aggregation level can be calculated. However, we observe that both Heur-SES and MAPA-SES, perform better than Opt-SES, if only marginally. We attribute this to the sampling uncertainty that affects both the identification of the underlying process, but also for the parameters estimation required to calculate the optimal aggregation level. This is echoed in the results for ARIMA(0,0,1), where Opt-SES performs worse than Heur-SES. Rostami-Tabar et al. (2013) showed that when the time series follows MA(1) process it is beneficial to aggregate as much as possible. Heur-SES does that, always aggregating to $k = 13$. On the other hand, Opt-SES because of the sampling uncertainty at times incorrectly identifies lower aggregation levels as appropriate. Similar comments can be made about ARIMA(1,0,1). In general for the remaining processes that the optimal aggregation level is only approximately identified the performance is similar to Heur-SES. This is reflected in the reported errors for ARIMA($*$,0,$*$).

When the unrestricted MAPA and Orig-ETS are considered, we see that the former offers substantial gains in accuracy and overall performs best for ARIMA($*$,0,$*$).

When looking at the simulated time series with first order differencing, it is evident that TA does not perform that well. However, this is to be expected as Orig-SES is optimal or approximately optimal for these time series, while forecasts on the temporally aggregated series are produced using substantially fewer observations. This is particularly evident in the performance of Heur-SES and Opt-SES. On the other hand, MAPA performs comparatively to the forecasts produced on the original time series. This is reflected in the overall performance for ARIMA($*$,1,$*$), where MAPA offers gains over Orig-SES and is marginally better than Orig-ETS.

Across all simulated series, looking at the results for ARIMA($*$,$*$,$*$) we observe that MAPA offers performance gains over both Orig-SES and Orig-ETS, demonstrating that TA can be beneficial for demand forecasting accuracy. However, MAPA-SES that does not take full advantage of the multiple views of the time series, being restricted to SES, is marginally worse that Orig-SES, while both Heur-SES and Opt-SES underperform. Like before, we attribute this to sampling uncertainties that make the estimation of the true time series models very challenging.

Looking at the results for the real datasets we observe that the most accurate forecasts are produced again by MAPA and MAPA-SES, both outperforming the conventional Orig-SES and Orig-ETS. It is interesting to note that for the real time series that the underlying process is both unknown, but also more complex, Orig-ETS performs worse than Orig-SES. In this case the model selection problem is more acute than for the simulated time series. On the other hand the multiple time series views afforded by MAPA result in minimal differences between MAPA and MAPA-SES.

In Figs. 2 and 3 we provide the ARMAE per forecast horizon for the simulated series, summarised as ARIMA($*$,0,$*$) and ARIMA($*$,1,$*$), and the real time series respectively. The horizontal black line is the benchmark Orig-SES, on which all other accuracies are calculated from.

Observe that for ARIMA($*$,0,$*$) TA offers substantial gains. As the forecast horizon increases the relative forecasting accuracy as measured by ARMAE increases. Notably as forecasts based on TA are build on aggregate data their good performance for long horizons is to be expected.

A different result emerges when we consider the performance of TA fore-

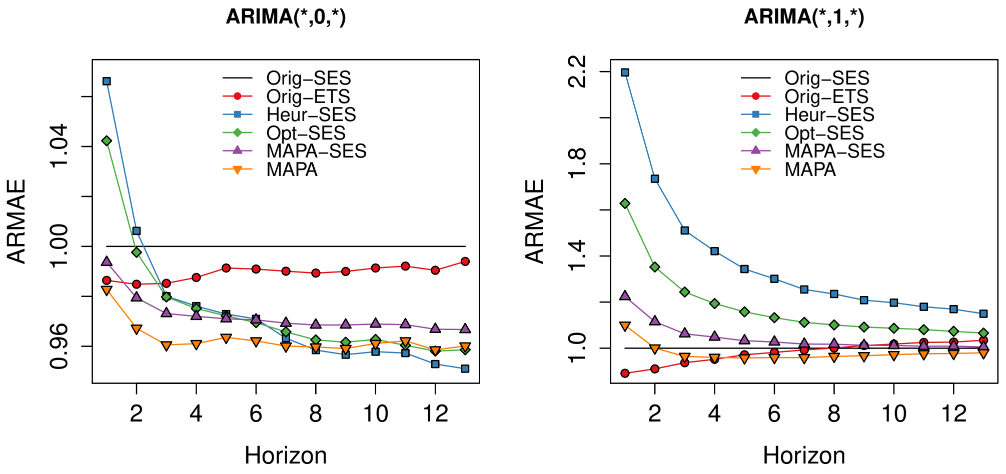

Figure 2: ARMAE per horizon for simulated ARIMA(∗,0,∗) and ARIMA(∗,1,∗).

casts for ARIMA(∗,1,∗) over the different forecast horizons. At shorter horizons Orig-ETS dominates, but for longer forecast horizons, as the uncertainty increases, its performance becomes indistinguishable to Orig-SES. The relative performance of both Heur-SES and Opt-SES to Orig-SES improves as the forecast horizon increases. This is attributed to the forecasts being produced at aggregated time series. Opt-SES performs better to Heur-SES over all horizons, demonstrating that even though the calculation of the optimal aggregation level is inappropriate, as no exact formulas have been derived for ARIMA(∗,1,∗) processes, the existing analytical derivations still offer benefits over Heur-SES, even when applied approximately. Finally the performance of the two MAPA based forecasts improves over horizons, but as discussed above, they do not offer benefits of the same magnitude as in the case of ARIMA(∗,0,∗).

The results for the real time series (Fig. 3), where the underlying models

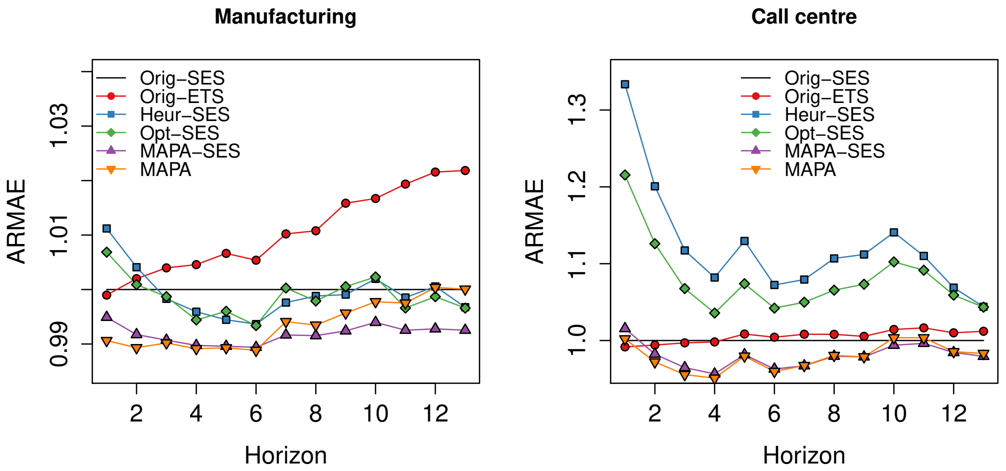

Figure 3: ARMAE per horizon for real datasets.

are both unknown and more complex, making SES suboptimal, demonstrate the benefits of TA. For the manufacturing dataset all Heur-SES, Opt-SES, MAPA-SES and MAPA perform better than Orig-SES at most forecast horizons. In contrast, the relative accuracy of Orig-ETS degrades as the the horizon increases. For the call centre dataset the Heur-SES and Opt-SES perform worse than the benchmark Orig-SES and Orig-ETS, across all horizons. We attribute this behaviour to the complexity of the original time series. On the other hand, both MAPA-SES and MAPA, which do not assume a specific underlying model for the series, perform better than the benchmarks. Finally, we focus on the relative accuracy in the pairs Orig-SES–MAPA-SES and Orig-ETS–MAPA, where similar shape of behaviour of errors is observed across horizons, but with the TA based forecasts being in both cases consistently and substantially better.

We turn our attention to the ARAME results in table 2. Overall we

Table 2: ARAME

Demand	No aggregation		Single level		Multiple levels	
	Orig-SES	Orig-ETS	Heur-SES	Opt-SES	MAPA-SES	MAPA
Simulated series						
ARIMA(1,0,0)	1.000	**0.984**	1.037	1.042	1.000	0.988
ARIMA(0,0,1)	1.000	1.014	1.004	0.998	**0.993**	1.000
ARIMA(2,0,0)	1.000	**0.990**	1.100	1.053	1.010	0.999
ARIMA(0,0,2)	1.000	1.005	0.987	**0.981**	0.986	0.994
ARIMA(1,0,1)	1.000	**0.990**	1.104	1.080	1.013	0.999
ARIMA(2,0,2)	1.000	**0.986**	1.067	1.029	1.003	0.998
ARIMA(1,1,0)	1.000	**0.821**	1.981	1.432	1.200	1.061
ARIMA(0,1,1)	1.000	**0.941**	1.850	1.454	1.213	1.134
ARIMA(2,1,0)	1.000	**0.700**	2.013	1.389	1.188	0.929
ARIMA(0,1,2)	1.000	**0.911**	1.864	1.331	1.170	1.094
ARIMA(1,1,1)	1.000	**0.773**	1.655	1.343	1.142	0.996
ARIMA(2,1,2)	1.000	**0.756**	1.751	1.330	1.158	0.942
ARIMA(∗,0,∗)	1.000	**0.995**	1.049	1.030	1.001	0.996
ARIMA(∗,1,∗)	1.000	**0.812**	1.848	1.379	1.178	1.023
ARIMA(∗,∗,∗)	1.000	**0.899**	1.392	1.192	1.086	1.010
Real datasets						
Manufacturing	1.000	0.992	1.051	1.075	1.014	**0.970**
Call centre	**1.000**	1.007	1.929	1.624	1.235	1.252
Overall real sets	1.000	**0.998**	1.314	1.251	1.090	1.065

observe that forecast built on the original time series are less biased than those built on aggregate versions of them, with only some exceptions. On the one hand, this is an intuitive result, given that both Orig-SES and Orig-ETS are modelled at the original time series, for which the fit will be unbiased. On the other hand, the TA based forecasts will be unbiased for their respective levels. This is in agreement with observations by Kourentzes et al. (2014).

For the real time series that the modelling uncertainty is much higher, MAPA performs well, as it takes advantage of TA to mitigate this. In contrast, the processes in the simulated data are relatively simple and therefore there is limited scope for MAPA to reduce the modelling uncertainty.

6. Conclusions

In this paper we investigated the use of TA for demand forecasting. More specifically we contrasted two different school of thoughts for how forecasting can be done using TA: (i) identifying and using a single optimal aggregation level; and (ii) using multiple levels. Each approach is shown to have its advantages.

Although theoretically using the optimal level would be advantageous, its performance is inhibited by two factors. First, analytical derivations exist only for a limited number of processes. Arguably this is not a substantial limitation as Rostami-Tabar et al. (2013) has shown a methodology how to develop the derivations for other processes. Nonetheless, with the available derivations, currently this TA modelling approach excludes direct modelling of seasonal time series, which can be important for some applications, and requires some appropriate pre-processing. Second, identifying the optimal level assumes knowledge of both the underlying process of the original time series and of the process at the aggregate level, which both come at high modelling uncertainty, especially for business time series that are typically relatively short. Nonetheless, we demonstrate that there are benefits over a simple heuristic to identify the aggregation level.

Using multiple TA levels is theoretically suboptimal, in the strict MSE sense, as the derivations for the optimal levels demonstrate. However, using multiple levels is particularly able at mitigating the modelling uncertainty. This is shown to be very useful for real series. In both the manufacturing and the call centre datasets using multiple temporal aggregation performed consistently more accurate than the benchmarks, across all forecast horizons. The two datasets exhibit substantial differences in terms of structure, given the diverse real application they originate from, but also the augmented Dickey-Fuller test results provided in section 4.1, and naturally the appropriate forecasting model is unknown. By using multiple TA levels to produce the forecasts, we were able to mitigate modelling uncertainty, as compared against the state-of-the-art selection approach for the exponential smoothing benchmark, and provide consistently more accurate forecasts. The accuracy gains are consistent with other application areas reported in the literature.

Irrespectively of which TA approach is used, aggregating a time series filters elements of the data, thus requiring simpler forecasting models. This is desirable in practice, as it is evident by the minimal use of complex forecasting methods (Weller and Crone, 2012), often attributed to the so called ‘algorithm aversion’ (Dietvorst et al., 2015). We provide further evidence that this comes at no forecast accuracy costs. In fact, the opposite is true, as highlighted by our empirical evaluation both on real supply chain data and the simulated examples.

Using multiple TA offers an additional advantage for business forecasting. The forecasts are by construction reconciled at different planning levels, short-term corresponding to low levels of TA and long-term corresponding to higher levels of TA, thus leading to aligned decisions. It is common in organisations that predictions for different objectives and horizons are based on different forecasting methods, which may result in substantial disagreements between them. MAPA avoids this problem, providing reconciled forecasts to support aligned decisions and limiting associated inefficiencies.

We conclude that using TA for demand forecasting is beneficial. There is adequate evidence to support further research in identifying the optimal aggregation level, however effort should be made to address the modelling uncertainty challenge. With regards to MAPA based approaches, we find them to be a practical approach for taking advantage of TA for forecasting time series. Using multiple levels is a novel view of how to implement TA, which has not been investigated in extensively in the forecasting literature that has primarily looked at the effects of using a single level. We found that it was at least as good at forecasting at the original time series, and in many cases substantially more accurate. Therefore additional research should be done to further develop the approach.

References

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., Petropoulos, F., et al., 2017. Forecasting with temporal hierarchies. European Journal of Operational Research.

Babai, M. Z., Ali, M. M., Nikolopoulos, K., 2012. Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis. Omega 40 (6), 713–721.

Barrow, D. K., 2016. Forecasting intraday call arrivals using the seasonal moving average method. Journal of Business Research 69 (12), 6088–6096.

Barrow, D. K., Kourentzes, N., 2016. Distributions of forecasting errors of forecast combinations: implications for inventory management. International Journal of Production Economics 177, 24–33.

Blanc, S. M., Setzer, T., 2016. When to choose the simple average in forecast combination. Journal of Business Research 69 (10), 3951–3962.

Boylan, J. E., Babai, M. Z., 2016. On the performance of overlapping and non-overlapping temporal demand aggregation approaches. International Journal of Production Economics.

Burnham, K. P., Anderson, D., 2002. Model selection and multi-model inference: A practical information-theoric approach. Springer.

Davydenko, A., Fildes, R., 2013. Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting 29 (3), 510–522.

Dietvorst, B. J., Simmons, J. P., Massey, C., 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General 144 (1), 114.

Fildes, R., Nikolopoulos, K., Crone, S. F., Syntetos, A., 2008. Forecasting and operational research: a review. Journal of the Operational Research Society 59 (9), 1150–1172.

Gardner, E. S., 2006. Exponential smoothing: The state of the art - part II. International Journal of Forecasting 22 (4), 637–666.

Green, K. C., Armstrong, J. S., 2015. Simple versus complex forecasting: The evidence. Journal of Business Research 68 (8), 1678–1685.

Hyndman, R. J., 2016. forecast: Forecasting functions for time series and linear models. R package version 7.1.
URL http://github.com/robjhyndman/forecast

Hyndman, R. J., Koehler, A. B., Ord, J. K., Snyder, R. D., 2008. Forecasting with Exponential Smoothing: The State Space Approach. Springer Verlag, Berlin.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., Grose, S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting 18 (3), 439–454.

Jalal, M. E., Hosseini, M., Karlsson, S., 2016. Forecasting incoming call volumes in call centers with recurrent neural networks. Journal of Business Research.

Jin, Y., Williams, B. D., Tokar, T., Waller, M. A., et al., 2015. Forecasting with temporally aggregated demand signals in a retail supply chain. Journal of Business Logistics 36 (2), 199–211.

Kourentzes, N., Petropoulos, F., 2015. Forecasting with multivariate temporal aggregation: The case of promotional modelling. International Journal of Production Economics.

Kourentzes, N., Petropoulos, F., 2016. MAPA: Multiple Aggregation Prediction Algorithm. R package version 1.9.1.
URL `https://CRAN.R-project.org/package=MAPA`

Kourentzes, N., Petropoulos, F., Trapero, J. R., 2014. Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30 (2), 291–302.

Kourentzes, N., Svetunkov, I., 2016. TStools: Time Series Analysis Tools and Functions. R package version 2.1.0.
URL `https://github.com/trnnick/TStools`

Lapide, L., 2004. Sales and operations planning part I: the process. The Journal of business forecasting 23 (3).

Luna, I., Ballini, R., 2011. Top-down strategies based on adaptive fuzzy rule-based systems for daily time series forecasting. International Journal of Forecasting 27 (3), 708–724.

Ma, S., Fildes, R., Huang, T., 2016. Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra-and inter-category promotional information. European Journal of Operational Research 249 (1), 245–257.

Merino, M., Ramirez-Nafarrate, A., 2016. Estimation of retail sales under competitive location in mexico. Journal of Business Research 69 (2), 445–451.

Miyaoka, J., Hausman, W. H., 2008. How improved forecasts can degrade decentralized supply chains. Manufacturing & Service Operations Management 10 (3), 547–562.

Mohammadipour, M., Boylan, J. E., 2012. Forecast horizon aggregation in integer autoregressive moving average (INARMA) models. Omega 40 (6), 703–712.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., Assimakopoulos, V., 2011. An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis. Journal of the Operational Research Society 62 (3), 544–554.

Pedregal, D. J., Trapero, J. R., 2010. Mid-term hourly electricity forecasting based on a multi-rate approach. Energy Conversion and Management 51 (1), 105–111.

Petropoulos, F., Kourentzes, N., 2014a. Forecast combinations for intermittent demand. Journal of the Operational Research Society 66 (6), 914–924.

Petropoulos, F., Kourentzes, N., 2014b. Improving forecasting via multiple temporal aggregation. Foresight: The International Journal of Applied Forecasting 2014 (34), 12–17.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., 2016. Another look at estimators for intermittent demand. International Journal of Production Economics.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Rossana, R. J., Seater, J. J., 1995. Temporal aggregation and economic time series. Journal of Business & Economic Statistics 13 (4), 441–451.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A., Ducq, Y., 2013. Demand forecasting by temporal aggregation. Naval Research Logistics (NRL) 60 (6), 479–498.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A., Ducq, Y., 2014. A note on the forecast performance of temporal aggregation. Naval Research Logistics (NRL) 61 (7), 489–500.

Sagaert, Y. R., Aghezzaf, E.-H., Kourentzes, N., Desmet, B., 2017. Temporal big data for tire industry tactical sales forecasting. Interfaces.

Silvestrini, A., Veredas, D., 2008. Temporal aggregation of univariate and multivariate time series models: a survey. Journal of Economic Surveys 22 (3), 458–497.

Spithourakis, G. P., Petropoulos, F., Babai, M. Z., Nikolopoulos, K., Assimakopoulos, V., 2011. Improving the performance of popular supply chain forecasting techniques. Supply Chain Forum: an international journal 12 (4), 16–25.

Syntetos, A. A., Babai, M. Z., Gardner, E. S., 2015. Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping. Journal of Business Research 68 (8), 1746–1752.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., Nikolopoulos, K., 2016. Supply chain forecasting: Theory, practice, their gap and the future. European Journal of Operational Research 252 (1), 1–26.

Tashman, L. J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. International journal of forecasting 16 (4), 437–450.

Trapero, J. R., Kourentzes, N., Fildes, R., 2012. Impact of information exchange on supplier forecasting performance. Omega 40 (6), 738–747.

Trapero, J. R., Kourentzes, N., Fildes, R., 2014. On the identification of sales forecasting models in the presence of promotions. Journal of the Operational Research Society 66 (2), 299–307.

Wei, W. W., 1978. Some consequences of temporal aggregation in seasonal time series models. In: Seasonal analysis of economic time series. NBER, pp. 433–448.

Weller, M., Crone, S. F., November 2012. Supply chain forecasting: Best practices & benchmarking study. Tech. rep., Lancaster Centre for Forecasting.