Elucidate structure in intermittent demand series

Nikolaos Kourentzes[a,b,*], George Athanasopoulos[c]

[a] *Skövde Artificial Intelligence Lab, School of Informatics, University of Skövde, Högskolevgen, Box 408, 541 28, Skövde, Sweden.*
[b] *Department of Management Science, Lancaster University Management School, Bailrigg, Lancaster, LA1 4YX, UK.*
[c] *Department of Econometrics and Business Statistics, Monash University, 900 Dandenong Road, Caulfield East, Victoria 3145, Australia*

Abstract

Intermittent demand forecasting has been widely researched in the context of spare parts management. However, it is becoming increasingly relevant to many other areas, such as retailing, where at the very disaggregate level time series may be highly intermittent, but at more aggregate levels are likely to exhibit trends and seasonal patterns. The vast majority of intermittent demand forecasting methods are inappropriate for producing forecasts with such features. We propose using temporal hierarchies to produce forecasts that demonstrate these traits at the various aggregation levels, effectively informing the resulting intermittent forecasts of these patterns that are identifiable only at higher levels. We conduct an empirical evaluation on real data and demonstrate statistically significant gains for both point and quantile forecasts.

Keywords: Forecasting, temporal aggregation, temporal hierarchies, forecast combination, forecast reconciliation.

1. Introduction

Intermittent demand forecasting is a challenging problem that relates to many aspects of supply chain (Bacchetti and Saccani, 2012), retailing (Fildes et al., 2019) and predictive maintenance (Van der Auweraer et al., 2018), among other applications. A key characteristic of intermittent demand is that it exhibits several periods of zero demand, therefore being variable both in the demand size, alike conventional demand, but also in terms of demand timing. Croston (1972) proposed a forecasting method to address this complexity by modelling the demand size and interval as two separate entities and dealing with each

*Correspondance: N Kourentzes, School of Informatics, Högskolan i Skövde, Högskolevgen, Box 408, 541 28, Skövde, Sweden.
Email addresses: nikolaos@kourentzes.com (Nikolaos Kourentzes), george.athanasopoulos@monash.edu (George Athanasopoulos)

variability independently. Since, there has been considerable research on alternative forecasting approaches (see, Teunter and Duncan, 2009; Bacchetti and Saccani, 2012; Van der Auweraer et al., 2018).

Many of the subsequently proposed approaches are modifications of the original Croston method or heavily inspired by it (for example, Syntetos and Boylan, 2005; Teunter et al., 2011). Bootstrapping (for example, Willemain et al., 2004; Syntetos et al., 2015), machine learning and neural networks (for example, Kourentzes, 2013; Nikolopoulos et al., 2016), and model based approaches (Snyder et al., 2012; Svetunkov and Boylan, 2017) have also been considered.

A common theme in most of these methods is that they extrapolate the local mean, resulting in constant value forecasts. This has been shown to perform well in a wide variety of empirical evaluations (see, Syntetos and Boylan, 2005; Teunter and Duncan, 2009; Kourentzes, 2014). However, it has some counter-intuitive implications. Consider a retailing example, where we are interested in forecasting daily demand series. Many of these series will be intermittent, as it is quite probable that many items at store level do not exhibit any demand for some days (Fildes et al., 2019; Li and Lim, 2018). Should we however consider demand in monthly time buckets, it is highly likely that the time series display trend and/or seasonal components. Such components are ignored by the intermittent demand methods. The typical constant value forecasts imply that there are omitted sources of variability. Furthermore they force a disconnect between the disaggregate and aggregate views of the time series.

There have been some attempts in the literature to incorporate trend (Altay et al., 2008) and seasonality (Lindsey and Pavur, 2013) in intermittent demand forecasts. The first is based on a modification of the linear trend exponential smoothing method by Wright (1986) that was developed to deal with irregularly sampled data. Altay et al. (2008) argue that this lends itself well for the intermittent demand case and show that it can be beneficial in forecasting aircraft spare parts. They find that it provides more accurate results than the bias corrected variant of Croston's method by Syntetos and Boylan (2005). However, the authors recognise that the evaluation suffers from limitations due to the selected error metrics. The trend method can quickly lead to zero forecasts when it detects a downwards trend. This coupled with the choice of absolute errors for the evaluation can be highly problematic. In intermittent demand absolute errors favour zero forecasts even when this makes limited operational sense (Teunter and Duncan, 2009; Kolassa, 2016).

Moreover, we argue that the main weakness of the proposed method is how the trend itself is modelled. In highly intermittent data an identified negative trend can push forecasts to zero, even when the product is still active in the market. The analogous argument holds for a positive trend. Instead, we would expect a series to remain intermittent, rather

than become zero or continuous, and any changes in the frequency of demand arrivals to appear as a trend at a more aggregate level. Similarly, the seasonality modification by Lindsey and Pavur (2013) is also somewhat unnatural for the intermittent demand setting. The authors adapt Croston's method with multiplicative seasonal indices. The approach is based on the assumption that the seasonal shape has to be mirrored in the most disaggregate view of the data. This means that the there is a canonicity to demand with an implied seasonality, which is against the definition of intermittent demand. In intermittent demand the uncertainty lies in both the demand size and interval. The Lindsey and Pavur (2013) adaptation implies a connection between these, while Croston's method assumes independence. Finally, the provided empirical evidence is relatively weak as the proposed modification is benchmarked only against exponential smoothing on simulated data. We argue that in both cases forcing unobserved or unnatural components upon an intermittent series is not necessary. A natural way for trend and/or seasonality to appear in intermittent data is by increasing or decreasing the rate of demand arrivals that would make such components apparent at more aggregate levels.

The use of temporal aggregation in intermittent demand modelling is not new. Nikolopoulos et al. (2011) proposed the ADIDA methodology that relies on temporally aggregating the time series to a less intermittent view, modelling and generating forecasts there, and disaggregating these to the original data frequency. This was shown to be beneficial, however the authors did not consider the issue of trend or seasonality. Furthermore, this work left the open question of what is the best temporal aggregation level to model at. Rostami-Tabar et al. (2013) explores this question further, but for continuous demand data, leaving this unresolved for intermittent time series. Petropoulos and Kourentzes (2015) attempt to avoid picking a single temporal aggregation level by adapting the multiple temporal aggregation prediction algorithm (Kourentzes et al., 2014) to the intermittent demand case, which uses multiple levels simultaneously. They find this to be beneficial in terms of accuracy, but with relatively small gains. Similarly with previous research, they did not consider the case of trending or seasonal time series, and assumed that single exponential smoothing is adequate as the time series are aggregated and intermittence lessens. Kourentzes et al. (2017) compared using a single temporal aggregation level to multiple and found the latter to be be beneficial. They argue that it is generally very difficult to identify an optimal level of aggregation when the underlying demand generating process is unknown, while using multiple levels hedges the modelling risk.

Building on the findings in the literature, in this paper we rely on using multiple levels of temporal aggregation to bring trend and seasonal information to the intermittent forecasts. We argue that such time series components are identifiable at aggregate levels, while not so at the observed level where the series are highly intermittent. We use the frame-

work of temporal hierarchies (Athanasopoulos et al., 2017) to enforce coherence between aggregate and intermittent forecasts. Coherence, in this context, is the requirement that the sum of the dissagregate intermittent forecasts equals the respective aggregate forecast. Therefore rather than modifying the forecast methods for intermittent series, we modify the intermittent forecasts to reflect trend and seasonal components captured at aggregate levels. We argue that constructing forecasting methods for intermittent data that directly capture such time series components is very challenging. This is among the reasons that there has been limited progress in developing forecasting models for intermittent demand (Shenstone and Hyndman, 2005; Snyder et al., 2012).

Taking into account the nature of intermittent time series, we propose a particular way to set the temporal hierarchies, as well as a correction for potentially negative forecasts that may occur when coherence is enforced upon intermittent forecasts with low values. Using a dataset of aircraft spare parts, we evaluate the proposed approach against an established intermittent demand method and provide evidence of significant gains both in terms of point and quantile forecasts, the latter being closely connected with the inventory decisions that such forecasts support. The contributions of this paper are as follows: (i) we propose the use of temporal hierarchies to elucidate structure in intermittent demand time series that is otherwise very difficult to capture; (ii) we demonstrate how to modify the framework to operate in the intermittent demand context and guarantee non-negative forecasts; and (iii) we provide empirical evidence of the efficacy of temporal hierarchies for intermittent time series.

The rest of the paper is organised as follows. Section 2 introduces forecasting with temporal hierarchies, as well as the necessary considerations for dealing with intermittent demand time series. Section 3 outlines the dataset used and the design of the empirical evaluation. Building on the results presented in Section 4, we discuss our work in the wider context of intermittent demand modelling and provide concluding remarks in Section 6.

2. Forecasting with temporal hierarchies

Forecasting with temporal hierarchies was introduced by Athanasopoulos et al. (2017). In what follows we present a simplified exposition of the methodology. In order to keep it simple and avoid complex notation required to capture all its intricacies we base the presentation and discussion on an example of quarterly time series. In the empirical application that follows the intermittent data are observed at the monthly frequency. We make reference to higher frequency examples where this is beneficial for the exposition. For further details we refer the reader to the above reference. In the following subsections we define the concept of temporal hierarchies and how these are used in forecasting. This is followed

by considering the case were the resulting forecasts are negative for which we propose a remedy.

2.1. Temporal hierarchies

Denote as $\boldsymbol{b} = (y_{Q_1}, y_{Q_2}, y_{Q_3}, y_{Q_4})'$ the $m = 4$-dimensional vector of observations of a quarterly time series y across four consecutive quarters. This is the highest frequency the time series is observed at. These are represented by the nodes at the bottom-level of the quarterly temporal hierarchy as shown in Figure 1. Using non-overlapping temporal aggregation we construct all integer data frequencies up to the annual level. Specifically, we construct semi-annual observations y_{SA_1} and y_{SA_2} and the annual observation y_A represented by the nodes in the middle and the top-level of the temporal hierarchy, respectively. For a monthly series the temporal hierarchy will consist of bi-monthly, quarterly, four-monthly, semi-annual and annual levels of aggregation. We stop at the annual frequency, as at that point higher frequency components of the time series, such as seasonality, are already removed completely.

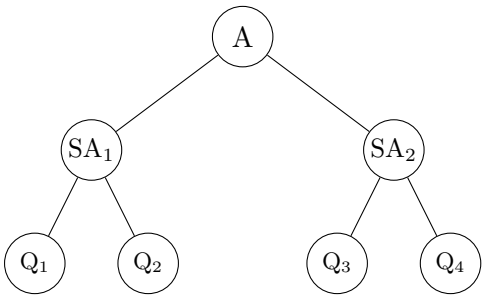

Figure 1: A temporal hierarchy for quarterly data. Q_ℓ with $\ell = 1, \ldots, 4$, denote quarters, SA_ℓ with $\ell = 1, 2$, semi-annual observations, and A the annual observation.

Stacking all observations of the temporal hierarchy in a $n = 7$-dimensional vector $\boldsymbol{y} = (y_A, y_{SA_1}, y_{SA_2}, y_{Q_1}, y_{Q_2}, y_{Q_3}, y_{Q_4})'$, we can write

$$\boldsymbol{y} = \boldsymbol{S}\boldsymbol{b},$$

where

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \boldsymbol{I}_m & \end{bmatrix}$$

has dimensions $n \times m$ and is referred to as the 'summing' matrix. $\boldsymbol{I}_m$ is an m-dimensional identity matrix. $\boldsymbol{S}$ is a map of the temporal hierarchy, where through linear summations

of the observed time series, we can construct all levels of temporal aggregation up to the annual level.

The process of temporal aggregation is implemented in a non-overlapping fashion across the observed sample. Using the temporal hierarchy the time series of different frequencies, corresponding to the various levels of aggregation, are generated. For the quarterly example, we denote these as $\boldsymbol{y}_i = \left(y_i^{[4]}, \boldsymbol{y}_i^{[2]}, \boldsymbol{y}_i^{[1]}\right)'$, where k in the superscript $[k]$ denotes the aggregation level and reflects the number of observations aggregated to obtain the series, and i is a temporal hierarchy index from 1 to $\lfloor T/m \rfloor$, where T is the observed sample size. For simplicity of exposition i is used as a common index across all levels of aggregation. It indicates all the observations within the temporal hierarchy, corresponding to the annual observation $y_i^{[m]}$.

Note that non-overlapping temporal aggregation requires the total number of observations to be an exact multiple of m so that all observations fit within a temporal aggregation structure or bucket. Starting the aggregation from $t^\star = T - \lfloor T/m \rfloor m + 1$ will ensure that any spare observations not fitting into an aggregation bucket are ignored at the beginning rather than the end of the sample, always retaining the most recent information. Alternatively, this can be thought of starting the temporal aggregation from the end of the sample and going backwards.

2.2. Temporal Hierarchical Forecasting - THieF

The concept of THieF is one of forecast reconciliation. Through this the resulting forecasts blend information from the various temporal aggregation levels giving THieF its strength. It begins by first generating a set of forecast for each level of temporal aggregation forming a complete temporal hierarchy of forecasts. These are commonly referred to as 'base' forecasts in a hierarchical forecasting context. For the quarterly series example we denote these as $\hat{\boldsymbol{y}}_\hbar = (\hat{y}_\hbar^{[4]}, \hat{\boldsymbol{y}}_\hbar^{[2]}, \hat{\boldsymbol{y}}_\hbar^{[1]})'$ where $\hbar$ specifies a common forecast index across the temporal hierarchy reflecting the forecast horizon at the annual level. Note that if h is the forecast horizon required for the observed bottom-level series, then $\hbar = 1, \ldots, \lceil h/m \rceil$ at the annual level. For each in between level of the temporal hierarchy we generate $k\lceil h/m \rceil$-steps ahead forecasts. Although temporarily aggregated data are by construction coherent, i.e., they add-up exactly across aggregation levels, in general base forecasts will not be.

Forecast reconciliation of the base forecasts is achieved by

$$\tilde{\boldsymbol{y}}_\hbar = \boldsymbol{SG}\hat{\boldsymbol{y}}_\hbar,$$

where $\boldsymbol{G}$ maps the base forecasts into the bottom-level and $\boldsymbol{S}$ sums these up to a set of coherent forecasts $\tilde{\boldsymbol{y}}_\hbar$. $\boldsymbol{SG}$ can be thought of as a reconciliation matrix, it takes the

incoherent base forecasts across all levels of aggregation, and reconciles them.

It is apparent that as long as $\boldsymbol{G}$ has non-zero columns, it linearly combines all $\hat{\boldsymbol{y}}_{\hbar}$ to bottom-level forecasts, hence these forecasts blend information from all levels, gaining the benefit of forecast combinations. A major drawback of traditional hierarchical forecasting approaches, whether in the temporal or cross-sectional setting, is the fixing of zero-columns in the $\boldsymbol{G}$ matrix. For example, the bottom-up approach only considers information from a the bottom-level zeroing out all other forecasts. There is now ample empirical evidence showing that using the full information set has substantial benefit in forecast accuracy across all levels (see for example Athanasopoulos et al., 2017; Wickramasuriya et al., 2019a, and references therein). Panagiotelis et al. (2019) also present theoretical justifications. Kourentzes et al. (2017) show that the benefits of using multiple levels extend even in the case that an optimal temporal aggregation level could be in theory identified (for example, Rostami-Tabar et al., 2013), as in practice estimation uncertainties creep in.

Wickramasuriya et al. (2019a) introduce an optimal full information approach for forecast reconciliation. They show that

$$\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}_{\hbar}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_{\hbar}^{-1} \tag{1}$$

minimises the $tr[\boldsymbol{S}\boldsymbol{G}\boldsymbol{W}_{\hbar}\boldsymbol{G}'\boldsymbol{S}']$ subject to $\boldsymbol{S}\boldsymbol{G}\boldsymbol{S} = \boldsymbol{S}$, where $\boldsymbol{S}\boldsymbol{G}\boldsymbol{W}_{\hbar}\boldsymbol{G}'\boldsymbol{S}' = \mathrm{Var}(\boldsymbol{y} - \tilde{\boldsymbol{y}}_{\hbar})$, the variance covariance matrix of the h-step ahead coherent forecast errors and $\boldsymbol{W}_{\hbar} = E(\hat{\boldsymbol{e}}_{\hbar}\hat{\boldsymbol{e}}_{\hbar})'$ is a positive definite covariance matrix of the base forecast errors $\hat{\boldsymbol{e}}_{\hbar} = \boldsymbol{y} - \hat{\boldsymbol{y}}_{\hbar}$. The method is referred to as MinT, as it minimises the trace of the covariance of the $\hbar$-step ahead coherent forecast errors. The significance of the $\boldsymbol{S}\boldsymbol{G}\boldsymbol{S} = \boldsymbol{S}$ constraint is that the resulting coherent forecasts are unbiased, as long as the base forecasts that were used are unbiased.

A challenge with the $\boldsymbol{G}$ matrix in MinT as specified in (1), is that it requires an estimate of $\boldsymbol{W}_{\hbar}$. A simplifying assumption imposed by Hyndman et al. (2011) focusing on the cross-sectional hierarchical forecasting setting, and also implemented by Athanasopoulos et al. (2009), was to set $\boldsymbol{W}_{\hbar} = \sigma\boldsymbol{I}_n$ for all $\hbar$, and $\sigma > 0$ is the variance of the forecast errors of the bottom-level series. This simplifying assumption has been shown to work well in practice (as shown in the aforementioned references) and also makes the approach trivial to use, as no further estimation of a covariance matrix is required and $\boldsymbol{G}$ depends only on $\boldsymbol{S}$ that is always known. However, it does ignore valuable information about the obvious scale differences and the interrelations between the observations within the hierarchical structure. In the temporal hierarchy case these are by construction present and therefore $\boldsymbol{W}_{\hbar} = \sigma\boldsymbol{I}_n$ is inappropriate for THieF.

Structural scaling

Athanasopoulos et al. (2017) and Wickramasuriya et al. (2019a) provide alternative approximations and estimators for $\boldsymbol{W}_\hbar$ in an attempt to overcome the above mentioned limitations. A particularly useful approximation for forecasting temporal hierarchies is what is referred to as structural scaling. Set $\boldsymbol{W}_\hbar = \sigma\boldsymbol{\Lambda}$, where $\sigma > 0$, $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{S1})$, and $\mathbf{1}$ is a unit vector of dimension m. This specification assumes that the bottom-level base forecast errors associated with the observed time series, have equal variance σ and are uncorrelated. Hence, it follows that higher level forecast error variances are simply the sum of the bottom-level errors. For the quarterly example $\boldsymbol{\Lambda} = \text{diag}(4, 2, 2, 1, 1, 1, 1)$. Hence, each element of the diagonal matrix contains the number of forecast error variances contributing to each node of the temporal hierarchy.

Structural scaling for THieF has been shown to perform strongly (Athanasopoulos et al., 2017) in various settings, on par with more involved approximations. In this structural scaling is very useful, as it only depends on the structure of the temporal hierarchy and its assumption of proportionally increasing variance to k is reasonable. Furthermore, it involves no variance estimation, which is important for our case, as this is particularly challenging in the intermittent demand setting. Note that σ does not need to be estimated as it cancels out when it enters (1) as part of $\boldsymbol{W}_\hbar$.

2.3. Non-negative reconciled forecasts

Given the intermittent and low count nature of our data an important concern is whether forecasts remain positive after temporal reconciliation is applied. In general, non-negative coherent forecasts are not guaranteed with the forecast reconciliation scheme described above, even when all base forecasts are positive, as reconciliation weights in rows of $\boldsymbol{SG}$ can be negative.

However in applying temporal reconciliation to intermittent data some special conditions hold for which generating negative coherent forecasts has low probability. To demonstrate this in a simple manner we once again assume that we observe our intermittent data at the quarterly level and therefore we have only two levels of aggregation above the bottom-level, the semi-annual and annual, as shown in Figure 1. Using structural scaling the reconciliation weights for each row of $\boldsymbol{SG}$ are shown below.

$$\begin{pmatrix} \tilde{y}_{A} \\ \tilde{y}_{SA_1} \\ \tilde{y}_{SA_2} \\ \tilde{y}_{Q_1} \\ \tilde{y}_{Q_2} \\ \tilde{y}_{Q_3} \\ \tilde{y}_{Q_4} \end{pmatrix} = \begin{pmatrix} 0.333 & 0.333 & 0.333 & 0.333 & 0.333 & 0.333 & 0.333 \\ 0.167 & 0.416 & -0.083 & 0.416 & 0.416 & -0.083 & -0.083 \\ 0.167 & -0.083 & 0.416 & -0.083 & -0.083 & 0.416 & 0.416 \\ 0.083 & 0.208 & -0.042 & 0.708 & -0.292 & -0.042 & -0.042 \\ 0.083 & 0.208 & -0.042 & -0.292 & 0.708 & -0.042 & -0.042 \\ 0.083 & -0.042 & 0.208 & -0.042 & -0.042 & 0.708 & -0.292 \\ 0.083 & -0.042 & 0.208 & -0.042 & -0.042 & -0.292 & 0.708 \end{pmatrix} \begin{pmatrix} \hat{y}_{A} \\ \hat{y}_{SA_1} \\ \hat{y}_{SA_2} \\ \hat{y}_{Q_1} \\ \hat{y}_{Q_2} \\ \hat{y}_{Q_3} \\ \hat{y}_{Q_4} \end{pmatrix} \quad (2)$$

Most intermittent demand methods, as well as many conventional forecasting methods, generate constant forecasts for the required h-steps ahead. Hence, for these cases base forecasts are not only positive but also identical within levels (we provide statistics on this from our empirical evaluation in Table 2). Denoting the constant forecasts at the semi-annual level by $\hat{y}_{SA}$ and forecasts at the quarterly level by $\hat{y}_{Q}$, the reconciled quarterly forecasts are given by $\tilde{y}_{Q} = 0.08\hat{y}_{A} + 0.17\hat{y}_{SA} + 0.33\hat{y}_{Q}$. Hence for these cases, coherent reconciled forecasts are always guaranteed to be positive.

For the rest of the cases, we work on the assumption that only the bottom-level base forecasts are guaranteed to be positive and identical, which are always generated by methods forecasting intermittent data. Denoting the identical bottom-level base forecasts by $\hat{y}_{Q}$ reconciled bottom-level forecasts are given by

$$\begin{pmatrix} \tilde{y}_{Q_1} \\ \tilde{y}_{Q_2} \\ \tilde{y}_{Q_3} \\ \tilde{y}_{Q_4} \end{pmatrix} = \begin{pmatrix} 0.083 & 0.208 & -0.042 & 0.333 \\ 0.083 & 0.208 & -0.042 & 0.333 \\ 0.083 & -0.042 & 0.208 & 0.333 \\ 0.083 & -0.042 & 0.208 & 0.333 \end{pmatrix} \begin{pmatrix} \hat{y}_{A} \\ \hat{y}_{SA_1} \\ \hat{y}_{SA_2} \\ \hat{y}_{Q} \end{pmatrix}. \quad (3)$$

Equation (3) shows one negative reconciliation weight for each reconciled bottom-level forecast. Furthermore, this negative weight is very low in value compared to the positive weights, making it very unlikely that these reconciliation combinations will generate negative forecasts for any "reasonable" set of base forecasts. For example, for $\tilde{y}_{Q_1} < 0$ we must have generated base forecasts so that $0.083\hat{y}_{A} + 0.208\hat{y}_{SA_1} + 0.333\hat{y}_{Q} < 0.042\hat{y}_{SA_1}$. Hence, the definition of reasonable can be loosely interpreted as having base forecasts that adequately represent or capture the scale of the series at each aggregation level. Therefore, we anticipate the probability of negative reconciled forecasts to be minimal. After checking through our empirical results (see Table 3) we find that indeed generating negative forecasts using structural scaling is highly unlikely, resulting in 2 to 7 cases out of the 5,000 time series, depending on the experimental settings.

More generally, when the bottom level forecasts are not expected to be identical, as will be the case for some intermittent demand approaches (for example, Kourentzes, 2013), or often for continuous demand data, we expect the probability of negative reconciled forecasts to increase. Although this is not a concern for this study, we propose below a correction scheme to address the limited cases we face, which is also more generally applicable.

2.4. Negative forecasts correction algorithm

As it is possible to obtain negative forecasts by using temporal hierarchies, albeit with small probability, we propose a correction scheme to guarantee non-negative predictions. Given a vector $\tilde{\boldsymbol{y}}_h$ that contains at least one negative value, we can construct a vector $\boldsymbol{c}_h$ that if added to $\tilde{\boldsymbol{y}}_h$ it will result in non-negative predictions that remain coherent. The coherence restriction implies that we cannot simply replace negative values with zeros. To estimate $\boldsymbol{c}_h$ we propose an iterative process. For this we first set $\breve{\boldsymbol{y}}_h = \tilde{\boldsymbol{y}}_h$ and follow:

Step 1 Form $|\breve{\boldsymbol{y}}_h^-|$ so that all non-negative forecasts are replaced with zeros and negative forecasts with their absolute value.

Step 2 Calculate the correction factor for this iteration as $\boldsymbol{c}_{j,h} = \boldsymbol{SG}|\breve{\boldsymbol{y}}_h^-|$. Note this ensures that the correction factor is itself coherent.

Step 3 If $\breve{\boldsymbol{y}}_h + \boldsymbol{c}_{j,h}$ contains negative forecasts, then update $\breve{\boldsymbol{y}}_h = \breve{\boldsymbol{y}}_h + \boldsymbol{c}_{j,h}$, increase counter j by 1 and return to Step 1. Otherwise, proceed to Step 4.

Step 4 Calculate the total correction $\boldsymbol{c}_h = \sum_j \boldsymbol{c}_{j,h}$.

At the end of this iterative process $\tilde{\boldsymbol{y}}_h + \boldsymbol{c}_h$ provides coherent non-negative forecasts. To ensure quick convergence in Step 3 we use a tolerance margin when evaluating whether there are any negative forecasts. A low value of 10^{-8} is appropriate for most cases.

The rationale behind the proposed iterative approach is that at each iteration the necessary correction to achieve non-negative forecasts is distributed throughout the whole hierarchy so as to retain coherence. This reduces the realised correction, requiring additional, yet smaller, corrections, which are iteratively applied. Table 1 provides an example for a quarterly hierarchy for which alternative quarters (Q_1 and Q_3) have low counts. Base $\hat{\boldsymbol{y}}_h$, coherent $\tilde{\boldsymbol{y}}_h$ and subsequently adjusted $\breve{\boldsymbol{y}}_h$ are provided. Note that the values of the last column are identical to $\tilde{\boldsymbol{y}}_h + \boldsymbol{c}_h$. The table also provides the amount of incoherency. We calculate this as the cumulative sum of the differences between the annual forecasts and the sums of the levels below, i.e., the difference between that annual and the sum of the semi-annual forecasts and the annual and the quarterly forecasts. The proposed iterative approach has the advantage of fast convergence, while being very simple. In our example

fewer than 16 iterations were sufficient and already from the 6^{th} iteration any further adjustments were beyond the third decimal point. Wickramasuriya et al. (2019b) propose an alternative algorithm formulated as a constrained quadratic programming problem which is much more computationally involved. As this is not central to our argument we do not consider it any further.

Table 1: Example of the iterative correction algorithm for a quarterly temporal hierarchy. j indicates the iteration number of the algorithm.

Level	$\hat{\boldsymbol{y}}_h$	$\tilde{\boldsymbol{y}}_h$	$\breve{\boldsymbol{y}}_h$, $j=1$	$\breve{\boldsymbol{y}}_h$, $j=2$	$\breve{\boldsymbol{y}}_h$, $j=3$	$\breve{\boldsymbol{y}}_h$, $j=4$	$\breve{\boldsymbol{y}}_h$, $j=5$	$\breve{\boldsymbol{y}}_h$, $j=6$
A	180	220.67	222.72	223.41	223.64	223.71	223.74	223.75
SA_1	135	129.33	130.49	130.86	130.98	131.02	131.04	131.04
SA_2	100	91.33	92.24	92.55	92.65	92.69	92.70	92.71
Q_1	4	-3.33	-1.09	-0.36	-0.12	-0.04	-0.01	0.00
Q_2	140	132.67	131.58	131.22	131.10	131.06	131.05	131.04
Q_3	3	-2.83	-0.97	-0.33	-0.11	-0.04	-0.01	0.00
Q_4	100	94.17	93.20	92.87	92.76	92.73	92.71	92.71
Incoherency	-122.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

3. Empirical evaluation

3.1. Data

We evaluate the efficacy of the proposed approach using a real data set of aerospace spare parts. The data set has been previously investigated in the literature and is characterised by high intermittence and variability (Syntetos et al., 2009; Teunter and Duncan, 2009; Petropoulos and Kourentzes, 2015). It contains 5,000 monthly time series, with 84 observations each. Figure 2 plots the percentage of zero demand periods against the coefficient of variation of the non-zero periods for each time series. Note that the plotted data include some jittering in the percentage of zeros, to better visualise the number of series in each bucket.

We can observe that the time series exhibit very high intermittence. On average, across all the time series, 89.8% periods have zero demand. Furthermore, most time series have very high demand variability, when that occurs.

Given the relevance of temporal aggregation to this research, we provide in Figure 3 the same information as in Figure 2, but for the temporally aggregated series to the annual level. Observe that both the percentage of zero demand and the coefficient of variation of the non-zero demand drop, as expected. In fact, 11.8% of the annually aggregated series exhibit no periods of zero demand.

For the evaluation we consider three forecast horizons, of 3, 6 and 12 months ahead. We retain the last 24 months for each series as a test set and perform a rolling origin

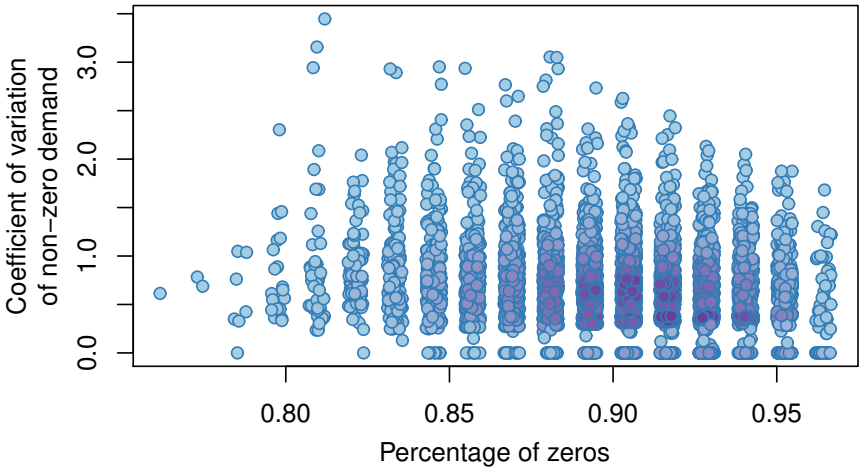

Figure 2: Scatter plot of the percentage of periods of zero demand against the coefficient of variation of non-zero demand of the original data. Each point corresponds to a time series.

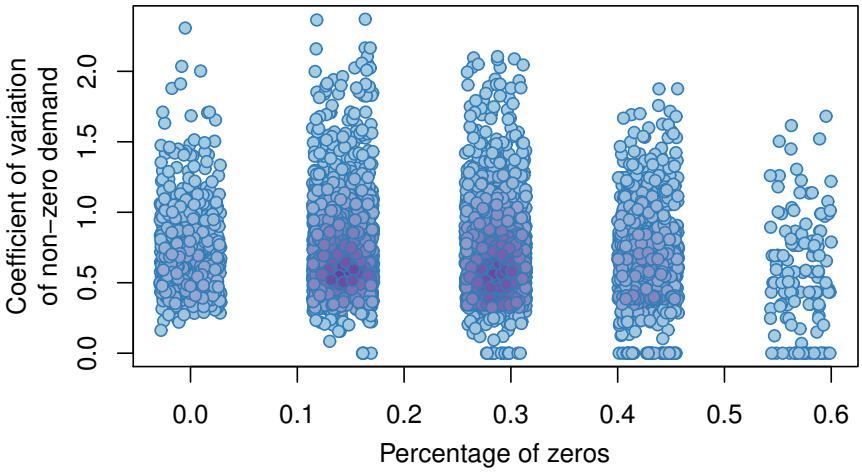

Figure 3: Scatter plot of the percentage of periods of zero demand against the coefficient of variation of non-zero demand of the annually aggregated data. Each point corresponds to a time series.

evaluation (Ord et al., 2017; Hyndman and Athanasopoulos, 2018). This is done as follows. First, we generate a forecast and evaluate its performance. Then we roll the forecast origin forward to include the next data point from the test set in the training set. We re-specify our forecasting method and repeat the process until we have used all available test data. For a given forecasting method and time series we generate 22, 19 and 13 out-of-sample forecasts for the three forecast horizons respectively, providing an adequate measurement

of performance.

3.2. Forecasting methods

To generate the forecasts for intermittent time series we rely on the TSB method (Teunter et al., 2011). The TSB method was introduced to deal with obsolescence issues in slow moving items. The forecast for period $t + h$, for forecast horizon h from period t is calculated as:

$$\hat{y}_{t+h} = \hat{d}_{t+h}\hat{z}_{t+h},$$

where $\hat{d}_{t+h}$ is the forecasted demand event probability and $\hat{z}_{t+h}$ is the forecasted non-zero demand size. The demand event probability is either 0 or 1, depending on whether a historical period contained a demand event or not. The forecast $\hat{d}_{t+h}$ is generated by smoothing the historical probability using single exponential smoothing. For the non-zero demand, we collect only periods with positive demand and construct a new vector removing all zero-demand periods, which is then predicted using single exponential smoothing to obtain $\hat{z}_{t+h}$.

The TSB method requires setting four parameters, two smoothing coefficients and two initial values for each single exponential smoothing process. To obtain these we optimise the method parameters for each time series using the Mean Absolute Rate loss function (MAR, Kourentzes, 2014):

$$MAR = \sum_{i=1}^{t} \left| \hat{y}_i - \frac{1}{i} \sum_{j=1}^{i} y_j \right|,$$

where y_t and $\hat{y}_t$ are the observed and fitted demand for in-sample period t. The idea behind MAR is to calculate the loss at a cumulative level of demand and was shown to perform better than alternative loss functions, such as MSE, for optimising methods for intermittent demand data (Kourentzes, 2014; Kourentzes et al., 2019). We rely on the package `tsintermittent` (Kourentzes and Petropoulos, 2016b) for the R statistical computing language (R Core Team, 2019) for the implementation of the TSB method.

Initial experiments found TSB to perform better than alternative intermittent demand forecasting methods, such as Croston's method and its modification with the SBA approximation (Syntetos and Boylan, 2001, 2005) on this dataset, and therefore we consider only the TSB method for the evaluation. Furthermore, obsolescence is relevant for the type of data and the time scale of our dataset. Past research had found the SBA to perform best (Syntetos et al., 2005; Teunter and Duncan, 2009), however these findings were prior to the proposition of the TSB method.

We use the ExponenTial Smoothing (ETS) family of models for continuous demand time series (Hyndman et al., 2008; Ord et al., 2017; Hyndman and Athanasopoulos, 2018).

It models time series as the total of four components, the time series level, trend, season and error term, which may interact with each other in an additive or multiplicative fashion. Additionally, for the trend component we consider the option of a linear or a damped trend. Combining all options together, the exponential smoothing family of models is conventionally considered to contain 30 members, covering an extensive range of time series. ETS is used widely both in research and practice, due to its relatively good forecasting accuracy, reliability, transparency and ease of implementation (Gardner Jr, 2006; Holt, 2004; Ord et al., 2017). Due to its prominence in the forecasting literature, we do not introduce ETS here, and refer the reader to Hyndman et al. (2008), Ord et al. (2017) or Hyndman and Athanasopoulos (2018) for the details.

In terms of implementation details, ETS model parameters are estimated using maximum likelihood and the selection between alternative model forms is done by Akaike Information Criterion, corrected for sample size, AICc (Burnham and Anderson, 2002; Hyndman et al., 2008). Hyndman and Akram (2006) show that some of the potential component combinations result in unstable forecasts, while Hyndman et al. (2018) recommend eliminating multiplicative trend models. These restrictions results in a reduced set of 15 models, where when there is a multiplicative seasonality only multiplicative errors are permitted, and the trend component may only be absent, additive linear or additive damped. We use the `forecast` package (Hyndman et al., 2018) for the R statistical computing language (R Core Team, 2019) to generate all ETS forecasts. Note that we also trialled ETS with model selection as a benchmark method for the disaggregate intermittent series. We found this to be inferior to the TSB benchmark and therefore did not consider it any further. Teunter and Duncan (2009) have shown that local level exponential smoothing performed poorly on this dataset, which we also found to hold for the ETS family of models.

We produce benchmark forecasts for the intermittent time series using the TSB method. We evaluate these against forecasts resulting from implementing forecast reconciliation via temporal hierarchies which use both TSB and ETS forecasts, as detailed in the following subsection. Note that we do not use a random walk benchmark, as this would result in a zero forecast for most of the time series in the dataset. Furthermore, we do not use a zero forecast as a benchmark, since this has no practical value, following the arguments by Teunter and Duncan (2009).

3.3. Temporal hierarchies

Given that the observed time series are sampled at a monthly frequency, implementing temporal hierarchies means that we consider the original monthly time series, as well as the temporally aggregated bi-monthly, tri-monthly, quarterly, half-yearly and yearly series. At each level of aggregation we generate base forecasts using either the TSB method or ETS.

There is no consensus on how to distinguish between intermittent and continuous demand time series. Syntetos et al. (2005) provide some guidelines for classifying time series that are forecast more accurately using Croston's method, the SBA approximation and single exponential smoothing. However, Kourentzes (2014) demonstrates that this classification scheme does not outperform approaches based on heuristic selection, in particular once methods' parameters are optimised. Furthermore, there is no similar work to help distinguish between other intermittent demand methods, such as the TSB method used here, and continuous forecasting methods, and specifically the ETS family of models. We overcome this issue by considering various alternative *intermittence thresholds*, which are calculated as the percentage of periods of zero-demand over the total number of in-sample periods. We report results for 10%, 20%, 30% and 40%. We expect the performance of ETS to drop as the intermittent threshold increases, as the presence of multiple zero observations makes parameter estimation more challenging (Kourentzes and Petropoulos, 2016a).

In our implementation of THieF we used structural scaling to approximate $\boldsymbol{W}_h$ as discussed in Section 2. From the THieF we retain only the bottom-level forecasts, corresponding to the original intermittent time series. This permits direct comparison with the benchmark forecasts. We use the `thief` package (Hyndman and Kourentzes, 2018) for the R statistical computing language (R Core Team, 2019) to generate the hierarchical forecasts, with the appropriate modifications to accommodate the introduction of TSB method in the temporal hierarchy.

3.4. Evaluation metrics

Given the dataset, in practice, forecasts will be used to support inventory decisions. Therefore we are interested in the cumulative error over the demand lead time, i.e. for a given lead time the total forecast demand has to meet the total realised demand. We track four metrics, the Mean Error (ME), the Root Mean Squared Error (RMSE), the Mean Interval Score (MIS) and the Pinball loss (PIN). The first two focus on the accuracy of point forecasts, while the latter two on quantile forecasts. To calculate these, for each set of 1 to h-step ahead forecasts from forecast origin j, we calculate the cumulative actuals Y_j and cumulative forecasts $\hat{Y}_j$:

$$Y_j = \sum_{i=t+j}^{t+j+h-1} y_i,$$
$$\hat{Y}_j = \sum_{i=t+j}^{t+j+h-1} \hat{y}_i.$$

Using these we get:

$$\text{ME} = \frac{1}{H}\sum_{j=1}^{H}\Big(Y_j - \hat{Y}_j\Big),$$

$$\text{RMSE} = \sqrt{\frac{1}{H}\sum_{j=1}^{H}\Big(Y_j - \hat{Y}_j\Big)^2},$$

$$\text{MIS} = \frac{1}{H}\sum_{j=1}^{H}\left((U-L) + \frac{2}{\alpha}(L-Y_j)\mathbf{1}\{Y_j < L\} + \frac{2}{\alpha}(Y_j - U)\mathbf{1}\{Y_j > U\}\right),$$

$$\text{PIN} = \begin{cases} (Y_j - U)\alpha, & \text{if } Y_j \geq U \\ (U - Y_j)(1-\alpha), & \text{if } Y_j < U \end{cases},$$

where the H is the number of rolling origins for the given forecast horizon h, U and L are the upper and lower quantile forecasts over the lead time demand, α is the target probability and $\mathbf{1}\{\cdot\}$ is an indicator function that takes the value of 1 when its condition is true and 0 otherwise.

The ME reports the point forecast bias and the RMSE reports the magnitude of the point forecast errors. For the measurement of accuracy we rely on quadratic errors as minimising these we get the expectation of the demand distribution. In contrast, picking the forecast that minimises absolute errors will result in a forecast that more closely tracks the median of the demand distribution (Kolassa, 2016). This can be problematic for intermittent demand data, as often the median is zero and explains some of the issues reported in measuring forecasting performance in the literature (Teunter and Duncan, 2009; Kourentzes, 2013).

The MIS shows the performance of the quantiles (L, U) for a given α, the desired interval (Gneiting and Raftery, 2007). Although we often assume the performance of the point forecasts to be a good proxy of the usefulness of the forecasts for the decision maker (Ord et al., 2017), the MIS connects more meaningfully with the use of forecasts to support inventory decisions. The PIN focuses on the upper quantile and directly models the asymmetric cost of over- and under-forecasting (Gneiting, 2011). In a news-vendor setting, the performance of the quantile of the forecasted distribution is connected with the inventory performance. Trapero et al. (2019) shows that the inventory performance of the news-vendor and the order-up-to stock control policy, in the 'ideal case', are linearly connected. In the 'ideal case' we do not consider extraordinary disturbances in the supply chain and back-orders are permitted. The latter is a reasonable assumption for the dataset on hand, that describes the demand of specialised aerospace spare parts that cannot be sourced from alternative sources and have to be back-ordered. Therefore, we report MIS and PIN values, as these

connect better to the decision supported by the forecast. For this evaluation we consider $\alpha = \{90\%, 95\%\}$.

All ME, RMSE, MIS and PIN are scale dependent errors, making them problematic for summarising the performance across time series. We divide the performance of the temporal hierarchies forecast with that of the benchmark TSB for each time series, removing any scaling issues. We summarise across all time series using the geometric mean (Davydenko and Fildes, 2013). Let X = {ME, RMSE, MIS, PIN},

$$\text{AvgRelX} = \sqrt[N]{\prod^{N} \left| \frac{\text{X}_A}{\text{X}_B} \right|},$$

where N is the number of time series and the subscripts A and B correspond to the errors of the temporal hierarchy forecasts and benchmark forecasts respectively, resulting in the AvgRelAME, AvgRelRMSE, AvgRelMIS and AvgRelPIN metrics. Note that we use the absolute values, so as to be able to calculate the geometric mean for the ME. The resulting AvgRelAME reports the magnitude of the bias, disregarding the direction. Relative errors lower than 1 indicate that the evaluated forecast outperform the benchmark, and vice versa.

Beyond the geometric mean, we also report the centred percentage best providing a non-parametric assessment of the forecast performance. For each forecast horizon and error metric, we report the percentage of time series for which THieF outperforms the benchmark (referred to as percentage better in the literature, Armstrong and Collopy, 1992; Makridakis and Hibon, 2000) after subtracting 50% to centre the measure around 0% and multiplying it by 2 to give it a range of $[-100\%, 100\%]$. Positive numbers correspond to the temporal hierarchy forecasts dominating on average, while negative numbers correspond to cases that the TSB benchmark forecasts are best. Note that this centred and scaled percentage best metric corresponds to the difference of mean ranks between the two methods. Beyond providing a non-parametric comparison, this permits to easily calculate whether differences are statistically significant, using the non-parametric Friedman test (Hollander et al., 2015).

3.5. Empirical quantile estimation

The TSB method, like many intermittent demand methods, outputs only point forecasts and therefore we cannot directly calculate desired quantiles of the forecast distribution. These are necessary to support decisions that depend on the forecasts, such as inventory management. We opt to estimate the desired quantiles using the empirical approach proposed by Trapero et al. (2019). The authors recommend using kernel density estimation to achieve a non-parametric modelling of the forecast error distribution, without requiring any assumptions about the underlying distribution. They found that this approach outperformed other parametric and non-parametric alternatives, especially when the error

distribution exhibited asymmetries.

As we are interested in the accuracy of the forecast demand over the lead time, we use the in-sample residuals $e_j = Y_j - \hat{Y}_j$ for the kernel density estimation, where $j = (1, \ldots, t-h)$, for all the in-sample cumulative forecast trace errors. We follow the recommendations by Trapero et al. (2019) and use the Epanechnikov kernel. At a point x,

$$f(x) = \frac{1}{(t-h)b} \sum_{j=1}^{t-h} K\left(\frac{x - e_j}{b}\right),$$

where $K(\cdot)$ is the kernel function:

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right), & \text{if } -\sqrt{(}5) \leq x \leq \sqrt{(}5) \\ 0, & \text{otherwise} \end{cases},$$

and $b = 0.9A(t-h)^{-1/5}$ is the bandwidth of the kernel, with A being an estimate of the spread of the errors that is given by the minimum between their standard deviation and their interquantile range divided by 1.34 (Silverman, 1986).

Using the empirical distribution obtained from the kernel density estimation, we can calculate the desired quantiles. We follow the same approach for both the benchmark TSB forecasts and the temporal hierarchy forecasts. Relying on the same approach for both, allows us to isolate any differences in the reported performance to the use of THieF.

4. Results

The motivating argument of this paper is that temporally aggregated intermittent demand time series can exhibit conventional time series components and accounting for these when generating forecasts can improve forecast accuracy. Table 2 reports the percentage of time series found to display some additional structure when aggregated, such as trend or seasonality, as identified using AICc in selecting an ETS model. The table provides results for different intermittence thresholds and forecast horizons. As the threshold level increases, more time series are considered by ETS and therefore the percentage of time series identified to display some extra structure increases.

The reported percentages change slightly for different horizons, as each involves a different number of rolling origin forecasts. We observe that for a 10% threshold about 20% of the time series are identified to display some additional structure. We note that for thresholds 20% and 30% the percentage stabilises on average to just over 40%. This indicates that the 10% threshold may be overly restrictive. The jump observed for a threshold of 40% is due to more time series being considered by ETS.

Table 2: Percentage of series with identified structure at aggregate levels.

Horizon	Intermittence threshold			
	10%	20%	30%	40%
3	21.58%	43.94%	43.98%	55.62%
6	20.98%	42.40%	42.44%	53.64%
12	19.40%	39.60%	39.64%	50.20%

Figure 4 provides examples of the additional structure modelled by ETS. Two example series are provided, one at each row, at the original monthly sampling frequency and at an aggregate level. Historical demand, base in-sample fit and forecasts for the next year are provided in each panel. In the first case (top panels) a trend becomes apparent at the aggregate level. In the second case (bottom panels) a seasonal pattern emerges. These patterns are easily captured by ETS. Base intermittent demand forecasts, using TSB in our evaluation, are constant. THieF combines these constant forecasts, with the aggregate ETS forecasts, resulting in forecasts that can capture some of these dynamics that become apparent only at the aggregate levels.

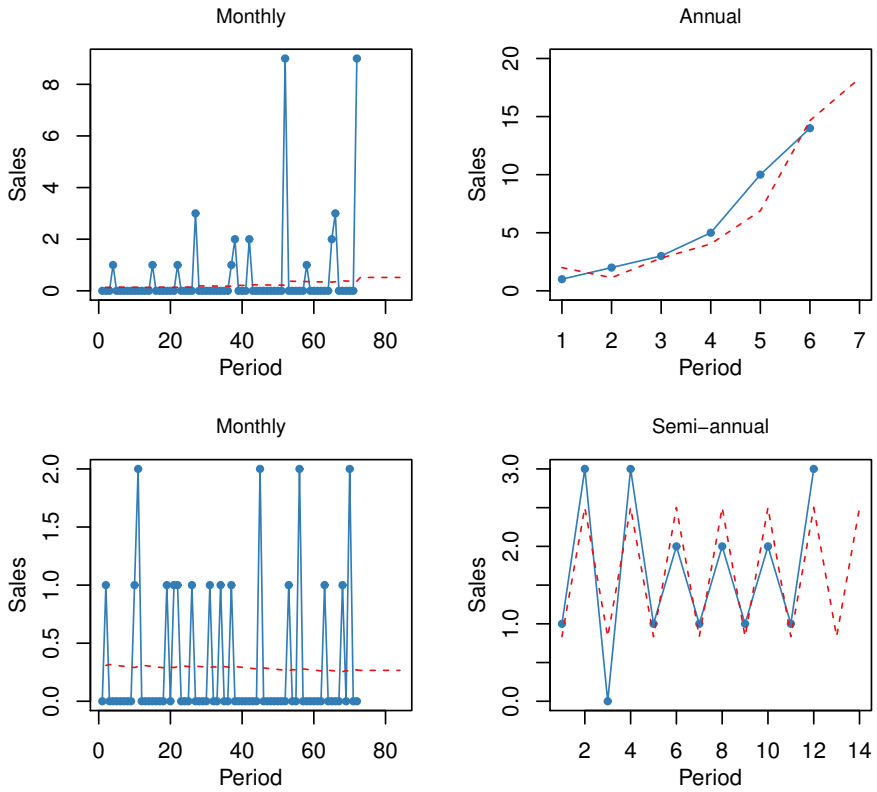

Figure 4: Two example series (131 and 4226) with structure appearing at higher temporal aggregation levels. Forecasts (- - -) are generated using TSB and ETS for the disaggregate and aggregate levels respectively.

Table 3 provides the percentage of cases that THieF resulted in negative values that required correction, and has the same structure as Table 2. The number of time series ranges from 2 to 7 out of 5000, which demonstrates the low probability of negative forecasts when THieF is used for intermittent data, as discussed in Section 2.3. For these cases we rely on the algorithm described in Section 2.4.

Table 3: Percentage of series with negative hierarchical forecasts

Horizon	Intermittence threshold			
	10%	20%	30%	40%
3	0.00%	0.04%	0.04%	0.14%
6	0.00%	0.04%	0.04%	0.12%
12	0.00%	0.04%	0.04%	0.08%

Table 4 summarises the performance of point forecasts in terms of AvgRelAME and AvgRelRMSE. The table is structured as follows: each column corresponds to an intermittence threshold and each row to a different lead time. The left side of the table provides the geometric mean, while the right side provides the centred percentage best with positive numbers suggesting improvements over the benchmark forecast. The p-values of the non-parametric statistical testing are provided in parentheses. The geometric mean retains the information about the magnitude of the errors, while the centred percentage best considers only the ranking of methods and matches the statistical test.

Table 4: Point forecast performance summary.

Horizon	Intermittence threshold							
	Geometric mean				Centred percentage best (p-value)			
	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
	|ME|							
3	0.981	0.971	0.972	0.967	8.160 (0.000)	13.560 (0.000)	13.480 (0.000)	17.160 (0.000)
6	0.988	0.988	0.988	0.977	7.040 (0.000)	11.040 (0.000)	11.000 (0.000)	12.800 (0.000)
12	0.994	0.986	0.988	0.979	7.240 (0.000)	8.880 (0.000)	8.640 (0.000)	8.120 (0.000)
	RMSE							
3	0.991	0.990	0.991	0.992	9.360 (0.000)	5.560 (0.000)	4.840 (0.001)	-2.040 (0.149)
6	0.988	0.989	0.990	0.992	9.840 (0.000)	6.120 (0.000)	5.600 (0.000)	-1.360 (0.336)
12	0.985	0.989	0.990	0.996	10.720 (0.000)	6.360 (0.000)	6.240 (0.000)	0.240 (0.865)

Considering the geometric mean of both the absolute ME and the RMSE, in all cases THieF outperforms the TSB benchmark irrespective of lead time or intermittence threshold. The results for the absolute ME display relatively larger improvements compared to the RMSE. We can observe that as the intermittence threshold increases, the relative bias of THieF improves. As the lead time increases, the relative difference in bias decreases. One might argue that these differences are small, however as the statistical test results

suggest, almost all improvements are statistically significant. The only exception is for an intermittence threshold of 40% when considering RMSE.

Note that gains in bias are generally considered to be more effective in improving the supported inventory decisions compared to the reduction in RMSE. This has been reported multiple times in the literature (for example, Sanders and Graman, 2009; Kourentzes et al., 2020).

Table 5 summarises the quantile performance of THieF and the TSB benchmark. The structure of the table is the same as before, providing the results for MIS and PIN for 90% and 95% targets. The main difference between the MIS and PIN is that the former considers the two-sided performance, while the latter focuses on the upper quantile and matches closely the inventory decision. Overall, we can see larger improvements over the results for the point forecasts (Table 4). In all cases THieF improves upon the TSB benchmark and in all cases the reported differences are statistically significant.

Table 5: Quantile forecast performance summary

Horizon	Intermittence threshold							
	Geometric mean				Centred percentage best (p-value)			
	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
					MIS 90%			
3	0.989	0.990	0.989	0.988	36.537 (0.000)	34.575 (0.000)	36.376 (0.000)	37.830 (0.000)
6	0.986	0.986	0.985	0.984	35.814 (0.000)	34.614 (0.000)	37.495 (0.000)	39.176 (0.000)
12	0.977	0.977	0.975	0.973	32.280 (0.000)	33.080 (0.000)	34.920 (0.000)	38.200 (0.000)
					MIS 95%			
3	0.984	0.983	0.983	0.981	36.844 (0.000)	35.883 (0.000)	37.805 (0.000)	39.808 (0.000)
6	0.984	0.984	0.983	0.982	37.735 (0.000)	37.055 (0.000)	38.856 (0.000)	40.976 (0.000)
12	0.975	0.975	0.974	0.970	33.080 (0.000)	33.360 (0.000)	35.880 (0.000)	40.360 (0.000)
					PIN 90%			
3	0.997	0.998	0.998	0.998	9.280 (0.000)	7.560 (0.000)	8.800 (0.000)	8.840 (0.000)
6	0.996	0.997	0.997	0.996	9.040 (0.000)	7.480 (0.000)	8.520 (0.000)	11.040 (0.000)
12	0.993	0.993	0.992	0.989	9.120 (0.000)	8.960 (0.000)	10.080 (0.000)	13.400 (0.000)
					PIN 95%			
3	0.994	0.993	0.993	0.991	9.440 (0.000)	10.760 (0.000)	11.480 (0.000)	12.400 (0.000)
6	0.993	0.992	0.991	0.989	9.200 (0.000)	10.400 (0.000)	10.960 (0.000)	12.840 (0.000)
12	0.994	0.989	0.987	0.980	8.480 (0.000)	9.200 (0.000)	10.960 (0.000)	14.480 (0.000)

For both MIS and PIN, the results improve further in favour of THieF as we consider more extreme quantiles. In terms of geometric mean of the metrics, as the the lead time increases, so does the gain over the benchmark. This is the case for both MIS and PIN and both quantile levels. Arguably, this is expected, as the forecasts at the higher temporal aggregation levels used within THieF have a stronger effect on the longer term performance (Kourentzes et al., 2014; Athanasopoulos et al., 2017). As we permit higher intermittence thresholds, the performance increases marginally.

For the MIS we observe large gains in terms of the centred percentage best. This provides strong evidence that the THieF based quantiles outperform the benchmark quantiles for the majority of the time series. The differences become more pronounced for MIS at the 95% quantile. The results for PIN show smaller gains for THieF compared to MIS, yet the same picture emerges. Again, we observe an increase in THieF gains when comparing the PIN 90% quantile to the PIN 95% quantile. Given that the PIN performance is directly connected with the inventory decision, this highlights the usefulness of the proposed approach for higher levels of service.

Overall, from both Tables 4 and 5 we observe that THieF outperforms the TSB benchmark. The differences are marginal for RMSE, but increase for the ME where THieF consistently outperforms the benchmark, irrespective of lead time or intermittence threshold. The benefits of THieF are highlighted further when we consider the quantile performance, where the differences increase even further. The geometric mean gains are small, yet consistent across all cases and there is strong evidence of statistical significance. We argue that the practical importance of the gains has two dimensions. First, the size of the gains is connected to the monetary implications of the dataset on hand. If the forecast items are very expensive (either in terms of procuring or storing) then even small differences are beneficial. Second, THieF outperforms the benchmark consistently across multiple conditions, demonstrating its reliability, which when paired with the ease of implementation makes a compelling case for adoption. Athanasopoulos et al. (2017), using simulation experiments, demonstrated that THieF performed equally or better than the base forecasts for continuous demand, under a wide range of design uncertainties, including full knowledge of the underlying demand process. Our results demonstrate that this finding extends to intermittent demand, albeit using real data. Note that merely using temporal aggregation is not as reliable and in that case the literature does not find the same consistent picture with the base forecasts often being better (Petropoulos and Kourentzes, 2015; Kourentzes et al., 2017). The structure of temporal hierarchies is what provides this reliability. Even when no additional time series components are identified at higher levels of temporal aggregation, using multiple levels, as in THieF, allows estimating the level of the forecast multiple times, with a varying degree of intermittency and eventually rely on combining this information to achieve better forecasts, reducing the modelling risk. This is evident in the relatively small impact of the intermittence threshold in the outcome of the empirical evaluation.

5. Discussion

In our evaluation we relied on TSB and ETS for predicting the time series. We further implemented a heuristical approach to distinguish between the two types of forecasts,

and when ETS was chosen we followed the recommended modelling approach to select between the various model forms. It is important to highlight that temporal hierarchies are independent of the approach used to generate the forecasts, and our selections used here can be easily substituted. For example, we can use parametric, non-parametric and model based approaches (Snyder et al., 2012; Svetunkov and Boylan, 2017; Hasni et al., 2019), machine learning and neural networks (Kourentzes, 2013; Nikolopoulos et al., 2016; Salinas et al., 2019). Similarly, the framework allows incorporating any rule for switching between intermittent demand and continuous demand forecasts (Syntetos et al., 2005; Kostenko and Hyndman, 2006; Kourentzes, 2014; Petropoulos and Kourentzes, 2015; Svetunkov and Boylan, 2017). Depending on the setup, the propensity for negative forecasts may change. Irrespectively, the proposed correction algorithm is also model free and therefore applicable in the general case.

Our argument in this paper is not about using specifically TSB or ETS, but rather that temporal hierarchies help the modeller capture the time series components that are not easy or even possible to identify at the disaggregate highly intermittent view that the time series are sampled. Intermittent data are characterised by uncertainty in the demand size and the demand event timing. This has made identifying conventional time series components very challenging, due to the irregularity of demand arrivals. We argue that for an intermittent model to be meaningful it has to provide predictions that when aggregated can give rise to the usual time series components, something that existing methods and models for intermittent demand do not achieve. Although temporal hierarchies do not provide a data generating process for the sampled data, we argue that it is a step in the right direction, as at minimum they offer forecasts that are coherent across temporal aggregation levels and can exhibit the usual time series patterns.

In this work we rely on a spare parts dataset. Our view is that although intermittent demand forecasting has been very essential in supporting supply chain decisions, as we increase the frequency of decision making in organisations, most forecasting challenges are bound to face issues of intermittency. For example, in retailing the horizon of many operational decisions is 1-day ahead, with a current move to even shorter forecast horizons. Once the sampling frequency becomes high enough, demand disaggregated at store/item level will be intermittent with very high probability (Fildes et al., 2019). With sufficiently high decision making frequency, most problems will exhibit elements of intermittency. In order to be able to address these, and support decision making in organisations, it will be essential that any forecasts are aware of the structures that appear at more aggregate levels, so as to result in aligned decision making. It also demonstrates that additional research is necessary in intermittent demand forecasting, as this will become increasingly prevalent.

Here we focused on time series components that become apparent at temporally aggre-

gate views of the data, using temporal hierarchies. The cross-sectional analogous problem, i.e., a hierarchical structure of the different items in the assortment of an organisation, typically grouped along the dimension of product type or market segment, is quite common in practice, and can also lead to different structures appearing at aggregate levels. This has been investigated in the context of intermittent demand, notably because it can substantially reduce intermittency, with mixed results. Moon et al. (2012) find that forecasting at an aggregate level and then disaggregating to the individual item sales offers limited gains compared to forecasting directly each series. Li and Lim (2018) propose a variant for disaggregating forecasts, making use of the forecast demand size and interval, finding gains in performance. The MinT framework outlined in Section 2 provides a common mathematical framework for both temporal and cross-sectional hierarchies. Recent work has demonstrated the connection between temporal and cross-sectional hierarchies (Kourentzes and Athanasopoulos, 2019), resulting in cross-temporally coherent forecasts. We argue that this is particularly relevant for intermittent demand forecasting, informing the most disaggregate intermittent forecasts with structures that appear either temporally or cross-sectionally, which current intermittent demand forecasting methods are unable to do.

6. Conclusions

The vast majority of intermittent demand forecasting methods provide constant forecasts, assuming that there are no trend or seasonal components in the data. Aggregating these forecasts to larger time buckets results in constant forecasts, even though the data start to exhibit such components. To overcome this dissonance we propose using temporal hierarchies to adjust disaggregate intermittent forecasts to account for identified components at higher aggregation views. Furthermore, we discuss the necessary considerations to forecast with temporal hierarchies on intermittent time series.

In our empirical evaluation we find that THieF provides significantly better point and quantile forecasts compared to benchmark intermittent demand forecasts. We rely on a dataset that has been repeatedly explored in the literature. We provide evidence that several time series exhibit trend and seasonality at aggregate levels, which has not been identified in the past when modelling them in their original sampling frequency. This demonstrates the strength of our approach, and validates our intuition that it is simpler to consider adjusting the intermittent forecasts to reflect the additional structure, rather than devise a model that identifies these at the disaggregate level.

We do not claim that the combination of TSB and ETS is the sole way to setup THieF for intermittent demand. On the contrary, we expect that different forecasting approaches

may be needed for different applications, given advances in individual methods and models, as well as data considerations, such as sample size. Similarly, although we rely on a heuristic to distinguish between producing intermittent and continuous forecasts there is no specific requirement and other criteria may be used. This model independence makes THieF powerful. Future research should explore these alternatives in more detail.

Another aspect that is not investigated in this work is the effect of data properties. This work compliments the results in the literature for non-intermittent time series for temporal hierarchies, by investigating THieF on a highly intermittent dataset. Building on previously published results (e.g., Athanasopoulos et al., 2017; Kourentzes and Athanasopoulos, 2019) we anticipate that the quality of the base forecasts is a key determinant of any gains in forecast accuracy achieved by THieF, which provides benefits when the base forecasts are incoherent, implying misspecification issues. Naturally, the quality of the base forecasts dependents on the data. This is particularly relevant for intermittent time series, which are notoriously difficult to model. Aspects such as the degree of intermittency of the time series, or more generally the structure of the data, play an important role in the generation of the base forecasts and in turn on the performance of THieF, and merits further investigation.

In this work we focused on the improvements at the disaggregate intermittent demand level of the time series, to support operational inventory management decisions. However, THieF produces coherent forecasts for all aggregation levels in the hierarchy. These are tied to different decision making problems, with different planning horizons and information bases. Investigating the benefits of coherent temporal hierarchy forecasts in the wider organisational context, as well as the gains on the inventory decision by incorporating the additional information base is interesting for future investigation. Finally, another aspect that is closely connected with decision making is the cost structure of the forecast errors. In many applications over- and under-forecasting have asymmetric costs. Although here we demonstrate empirically that the use of THieF will result in a reduction of forecast bias, the method remains agnostic to such asymmetries. Future research should explore the incorporation and impact of cost asymmetries further.

References

Altay, N., Rudisill, F., Litteral, L. A., 2008. Adapting wright's modification of holt's method to forecasting intermittent demand. International Journal of Production Economics 111 (2), 389–408.

Armstrong, J. S., Collopy, F., 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. International journal of forecasting 8 (1), 69–80.

Athanasopoulos, G., Ahmed, R. A., Hyndman, R. J., 2009. Hierarchical forecasts for australian domestic tourism. International Journal of Forecasting 25 (1), 146–166.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., Petropoulos, F., 2017. Forecasting with temporal hierarchies. European Journal of Operational Research 262, 60–74.

Bacchetti, A., Saccani, N., 2012. Spare parts classification and demand forecasting for stock control: Investigating the gap between research and practice. Omega 40 (6), 722–737.

Burnham, K. P., Anderson, D. R., 2002. Model selection and multi-model inference: a practical information theoretic approach, 2nd Edition. NY: Springer-Verlag.

Croston, J. D., 1972. Forecasting and stock control for intermittent demands. Journal of the Operational Research Society 23 (3), 289–303.

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.

Fildes, R., Ma, S., Kolassa, S., 2019. Retail forecasting: Research and practice. International Journal of Forecasting.

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

Gneiting, T., 2011. Quantiles as optimal point forecasts. International Journal of forecasting 27 (2), 197–207.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Hasni, M., Aguir, M., Babai, M., Jemai, Z., 2019. Spare parts demand forecasting: a review on bootstrapping methods. International Journal of Production Research 57 (15-16), 4791–4804.

Hollander, M., Wolfe, D. A., Chicken, E., 2015. Nonparametric statistical methods, 3rd Edition. John Wiley & Sons.

Holt, C. C., 2004. Author's retrospective on 'forecasting seasonals and trends by exponentially weighted moving averages'. International Journal of Forecasting 20 (1), 11–13.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., 2018. forecast: Forecasting

functions for time series and linear models. R package version 8.4.
URL http://pkg.robjhyndman.com/forecast

Hyndman, R., Koehler, A. B., Ord, J. K., Snyder, R. D., 2008. Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., Shang, H. L., 2011. Optimal combination forecasts for hierarchical time series. Computational Statistics & Data Analysis 55 (9), 2579–2589.

Hyndman, R. J., Akram, M., 2006. Some nonlinear exponential smoothing models are unstable. Tech. rep., Monash University, Department of Econometrics and Business Statistics.

Hyndman, R. J., Athanasopoulos, G., 2018. Forecasting: principles and practice, 2nd Edition. OTexts, Melbourne, Australia.
URL http://otexts.com/fpp2/

Hyndman, R. J., Kourentzes, N., 2018. thief: Temporal HIErarchical Forecasting. R package version 0.3.
URL http://pkg.robjhyndman.com/thief

Kolassa, S., 2016. Evaluating predictive count data distributions in retail sales forecasting. International Journal of Forecasting 32 (3), 788–803.

Kostenko, A. V., Hyndman, R. J., 2006. A note on the categorization of demand patterns. Journal of the Operational Research Society 57 (10), 1256–1257.

Kourentzes, N., 2013. Intermittent demand forecasts with neural networks. International Journal of Production Economics 143 (1), 198–206.

Kourentzes, N., 2014. On intermittent demand model optimisation and selection. International Journal of Production Economics 156, 180–190.

Kourentzes, N., Athanasopoulos, G., 2019. Cross-temporal coherent forecasts for Australian tourism. Annals of Tourism Research 75, 393–409.

Kourentzes, N., Li, D., Strauss, A. K., 2019. Unconstraining methods for revenue management systems under small demand. Journal of Revenue and Pricing Management 18 (1), 27–41.

Kourentzes, N., Petropoulos, F., 2016a. Forecasting with multivariate temporal aggregation: The case of promotional modelling. International Journal of Production Economics 181, 145–153.

Kourentzes, N., Petropoulos, F., 2016b. tsintermittent: Intermittent Time Series Forecasting. R package version 1.9.
URL `https://CRAN.R-project.org/package=tsintermittent`

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., Rostami-Tabar, B., Barrow, D. K., 2017. Demand forecasting by temporal aggregation: using optimal or multiple aggregation levels? Journal of Business Research 78, 1–9.

Kourentzes, N., Trapero, J. R., Barrow, D. K., 2020. Optimising forecasting models for inventory planning. International Journal of Production Economics 225, 107597.

Li, C., Lim, A., 2018. A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing. European Journal of Operational Research 269 (3), 860–869.

Lindsey, M., Pavur, R., 2013. Assessing a modification to Croston's method to incorporate a seasonal component. In: Advances in Business and Management Forecasting. Emerald Group Publishing Limited, pp. 185–195.

Makridakis, S., Hibon, M., 2000. The M3-competition: results, conclusions and implications. International journal of forecasting 16 (4), 451–476.

Moon, S., Hicks, C., Simpson, A., 2012. The development of a hierarchical forecasting method for predicting spare parts demand in the South Korean navy–a case study. International Journal of Production Economics 140 (2), 794–802.

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.

Nikolopoulos, K. I., Babai, M. Z., Bozos, K., 2016. Forecasting supply chain sporadic demand with nearest neighbor approaches. International Journal of Production Economics 177, 139–148.

Ord, J. K., Fildes, R., Kourentzes, N., 2017. Principles of Business Forecasting, 2nd Edition. Wessex Press Publishing Co.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., Hyndman, R. J., 2019. Forecast reconciliation: A geometric view with new insights on bias correction. Monash University, Work Paper 18/19, 1–33.

Petropoulos, F., Kourentzes, N., 2015. Forecast combinations for intermittent demand. Journal of the Operational Research Society 66 (6), 914–924.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

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

Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T., 2019. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting.

Sanders, N. R., Graman, G. A., 2009. Quantifying costs of forecast errors: A case study of the warehouse environment. Omega 37 (1), 116–125.

Shenstone, L., Hyndman, R. J., 2005. Stochastic models underlying Croston's method for intermittent demand forecasting. Journal of Forecasting 24 (6), 389–402.

Silverman, B. W., 1986. Density estimation for statistics and data analysis. Vol. 26. CRC press.

Snyder, R. D., Ord, J. K., Beaumont, A., 2012. Forecasting the intermittent demand for slow-moving inventories: A modelling approach. International Journal of Forecasting 28 (2), 485–496.

Svetunkov, I., Boylan, John, E., 2017. Multiplicative state-space models for intermittent time series. Working Paper of Department of Management Science, Lancaster University.

Syntetos, A. A., Babai, M. Z., Dallery, Y., Teunter, R., 2009. Periodic control of intermittent demand items: theory and empirical analysis. Journal of the Operational Research Society 60 (5), 611–618.

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

Syntetos, A. A., Boylan, J. E., 2001. On the bias of intermittent demand estimates. International Journal of Production Economics 71 (1-3), 457–466.

Syntetos, A. A., Boylan, J. E., 2005. The accuracy of intermittent demand estimates. International Journal of Forecasting 21 (2), 303–314.

Syntetos, A. A., Boylan, J. E., Croston, J., 2005. On the categorization of demand patterns. Journal of the Operational Research Society 56 (5), 495–503.

Teunter, R. H., Duncan, L., 2009. Forecasting intermittent demand: a comparative study. Journal of the Operational Research Society 60 (3), 321–329.

Teunter, R. H., Syntetos, A. A., Babai, M. Z., 2011. Intermittent demand: Linking forecasting to inventory obsolescence. European Journal of Operational Research 214 (3), 606–615.

Trapero, J. R., Cardos, M., Kourentzes, N., 2019. Empirical safety stock estimation based on kernel and GARCH models. Omega 84, 199–211.

Van der Auweraer, S., Boute, R. N., Syntetos, A. A., 2018. Forecasting spare part demand with installed base information: A review. International Journal of Forecasting.

Wickramasuriya, S. L., Athanasopoulos, G., Hyndman, R. J., 2019a. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. Journal of the American Statistical Association 114 (526), 804–819.

Wickramasuriya, S. L., Turlach, B., Hyndman, R. J., 2019b. Optimal non-negative forecast reconciliation. Monash University, Work Paper 15/19.

Willemain, T. R., Smart, C. N., Schwarz, H. F., 2004. A new approach to forecasting intermittent demand for service parts inventories. International Journal of forecasting 20 (3), 375–387.

Wright, D. J., 1986. Forecasting data published at irregular time intervals using an extension of Holt's method. Management science 32 (4), 499–510.