Forecast Combinations for Intermittent Demand

Fotios Petropoulos[a,*], Nikolaos Kourentzes[a]

[a]*Lancaster Centre for Forecasting, Department of Management Science*
Lancaster University Management School, Lancaster University, Lancaster, UK

Abstract

Intermittent demand is characterised by infrequent demand arrivals, where many periods have zero demand, coupled with varied demand sizes. The dual source of variation renders forecasting for intermittent demand a very challenging task. Many researchers have focused on the development of specialised methods for intermittent demand. However, apart from a case study on hierarchical forecasting, the effects of combining, which is a standard practice for regular demand, have not been investigated. This paper empirically explores the efficiency of forecast combinations in the intermittent demand context. We examine both method and temporal combinations of forecasts. The first are based on combinations of different methods on the same time series, while the latter use combinations of forecasts produced on different views of the time series, based on temporal aggregation. Temporal combinations of single or multiple methods are investigated, leading to a new time-series classification, which leads to model selection and combination. Results suggest that appropriate combinations lead to improved forecasting performance over single methods, as well as simplifying the forecasting process by limiting the need for manual selection of methods or hyper-parameters of good performing benchmarks. This has direct implications for intermittent demand forecasting in practice.

Keywords: Intermittent demand, Parametric Methods, Combining, Temporal Aggregation, Classification, Forecasting

1. Introduction

Demand forecasts are necessary in every aspect of decision making and operations, from short-term inventory planning to long-term strategic estimates. Minimum variance, unbiased forecasts are of great practical significance, as they affect both inventory costs and customer service levels. However, this task is not always straightforward. According to Johnston et al. (2003), 60% of the stock keeping units (SKUs) in a standard industrial setting are characterised as intermittent, translating to infrequent demand arrivals coupled with variable demand sizes, whenever demand occurs. These irregular patterns render the problem of accurately estimating the demand especially challenging.

*Correspondance: F Petropoulos, Department of Management Science, Lancaster University Management School, Lancaster, Lancashire, LA1 4YX, UK.

Email addresses: f.petropoulos@lancaster.ac.uk (Fotios Petropoulos), n.kourentzes@lancaster.ac.uk (Nikolaos Kourentzes)

The most widely used approach for forecasting intermittent demand is Croston's method (Croston, 1972). Almost 30 years have passed before this method was proved to be biased (Syntetos and Boylan, 2001) and alternatives have been considered (for example see: Syntetos and Boylan, 2005; Teunter et al., 2011). Despite the developments over the recent years, many industries are still relying on even simpler approaches, such as Moving Average (MA) and Simple Exponential Smoothing (SES) directly applied on the original data.

Forecast combination has been widely regarded as beneficial (Clemen, 1989), producing more accurate forecasts, while reducing the variance of the residuals. The most common approach is the weighted or unweighted combination of the outputs of different forecasting methods fitted on the same data. Alternatively, one could combine forecasts derived from the same information, but with different sampling properties, such as frequency. Nikolopoulos et al. (2011) showed that using non-overlapping temporal aggregation and forecasting the aggregated time series can lead to substantial performance improvements against forecasts created from the original time series. For intermittent time series, temporal aggregation results in a reduction of the intermittence of the demand, thus simplifying the forecasting problem. Although one may use the same method to forecast at the different levels of aggregation, different forecasting methods are able to capture the dynamics that are highlighted or attenuated via temporal aggregation. Kourentzes et al. (2014) showed that combining such forecasts from multiple aggregation levels resulted in improved accuracy for fast moving items.

With the exception of hierarchical forecasts (Moon et al., 2012), to the best of our knowledge, forecast combinations have not been examined in the context of intermittent demand. The current paper investigates the efficiency of combining forecasts produced from widely used parametric methods when dealing with intermittent demand. Forecasting combination is examined in the cases of: (i) single sampling frequency and multiple methods, (ii) multiple frequencies and single method, and (iii) multiple frequencies and multiple methods. In the latter case, classification schemes may be considered (Syntetos et al., 2005; Kostenko and Hyndman, 2006) in order to select which is the most appropriate method as the time-series characteristics change along with the level of aggregation.

One aspect of this study is to investigate the forecast accuracy and robustness benefits, if any, that the different combination schemes offer. Another aspect is associated with removing the need to select a single method. By combining forecasts from several methods it is implied that no single method is preferable. This may be particularly interesting for intermittent demand time series. Although there are classification schemes to support our choice between some alternative methods (Syntetos et al., 2005), the literature does not offer adequate guidelines how to select between the various methods that are applicable to intermittent time series.

An empirical evaluation is conducted using real time series from the Royal Air Force (RAF), providing interesting results with regard to the forecasting performance, while reducing the need for model selection. The rest of the paper is organised as follows: in the next section we present approaches for intermittent demand that are widely used in the literature, followed by a short review of forecast combinations. Section 3 outlines the experimental set-up of the current research, while Section 4 discusses the evaluation of the results. Concluding remarks, managerial implications and paths for future research are provided in Section 5.

2. Background Research

2.1. Forecasting for Intermittent Demand

The irregular nature of intermittent demand data makes conventional extrapolative forecasting methods unsuitable. Croston (1972) proposed an innovative approach to handle data containing periods of zero demand, later corrected by Rao (1973). A decomposition takes place prior to forecasting so that the original series is divided in two components: i) the non-zero demands sizes, and ii) the time intervals between consecutive non-zero demands. Each series is forecasted separately and the final forecast is calculated as the ratio of the two forecasts, thus giving a *demand rate*. If we let $\hat{z}_t$ and $\hat{p}_t$ denote the forecasts for the demand sizes and the intervals respectively for period t, the final forecast $\hat{y}_t$ is simply derived as:

$$\hat{y}_t = \frac{\hat{z}_t}{\hat{p}_t}. \tag{1}$$

Both demand sizes and intervals are estimated using SES:

$$\hat{z}_t = \hat{z}_{t-1} + \alpha_z \left(z_{t-1} - \hat{z}_{t-1}\right), \tag{2}$$

$$\hat{p}_t = \hat{p}_{t-1} + \alpha_p \left(p_{t-1} - \hat{p}_{t-1}\right), \tag{3}$$

where z and p are the vectors containing the actual values for the non-zero demand sizes and the inter-demand internals respectively and α_z and α_p are smoothing factors. Demand sizes and intervals are updated only when a non-zero demand occurs.

The forecasting accuracy and inventory performance of this method has been widely verified (for examples see: Willemain et al., 1994; Johnston and Boylan, 1996; Syntetos and Boylan, 2006; Gardner, 2006). At the same time, Croston's method has been criticised for the inconsistency between the underlying model and his method (Shenstone and Hyndman, 2005), while its assumption of the independence between demands and intervals has been challenged (Willemain et al., 1994; Kourentzes, 2013). Syntetos and Boylan (2001) proved that the method is biased, with the magnitude of bias being positively correlated with the value of the α_p smoothing parameter. To overcome this problem, Syntetos and Boylan (2005) suggested a damping factor which is directly multiplied to the demand rate as estimated by Croston's method, thus:

$$\hat{y}_t = \left(1 - \frac{\alpha_p}{2}\right) \frac{\hat{z}_t}{\hat{p}_t}. \tag{4}$$

This modified version is known as the Syntetos-Boylan Approximation (SBA). Even if some other modifications of the Croston's method have been proposed (for examples see: Syntetos, 2001; Levén and Segerstedt, 2004; Teunter et al., 2011), the SBA is the only alternative approach with substantial empirical support, demonstrating superior performance compared to the original method (Eaves and Kingsman, 2004; Syntetos and Boylan, 2005, 2006; Teunter and Sani, 2009; Petropoulos et al., 2013).

For both Croston's method and SBA it has been suggested that α_z, $\alpha_p \in [0.1, 0.3]$ (Croston, 1972) or α_z, $\alpha_p \in [0.05, 0.2]$ (Syntetos and Boylan, 2005). Even if there is some empirical evidence that per series optimisation can lead to improvements in terms of bias (Petropoulos et al., 2013), it is common practice that the values of the smoothing parameters are selected in an ad-hoc way (Syntetos and Boylan, 2005; Teunter and Duncan, 2009; Romeijnders et al., 2012). Lastly, while Snyder (2002) and Teunter et al. (2010) have examined different parameters for the numerator

and the denominator, it is common that both $\hat{z}$ and $\hat{p}$ are smoothed using the same value, thus $\alpha_z = \alpha_p$.

Despite the advances in methods specialised in intermittent demand, simpler methods, such as MA and SES, are often used in practice (Syntetos and Boylan, 2005; Teunter and Duncan, 2009; Willemain et al., 1994), with the latter one demonstrating good performance in some cases (Wallström and Segerstedt, 2010). Moreover, the Naive method is usually considered as a benchmark (Babai et al., 2011; Petropoulos et al., 2014). This is of particular interest when examining series with a high degree of intermittence. For series where the last observation is zero, the Naive method will give zero forecasts. This will result in zero forecast error for most of the out-of-sample periods, but also implies zero orders and stock keeping that makes no business sense.

On the antipode of parametric approaches, researchers have focused also in bootstrapping techniques (Willemain et al., 2004; Teunter and Duncan, 2009; Snyder et al., 2012) and neural networks (Gutierrez et al., 2008; Kourentzes, 2013). However, these approaches have not been widely applied, being more complex for practitioners, and in some cases there is no clear evidence of their superiority against simpler alternatives (Teunter and Duncan, 2009).

2.2. Temporal Aggregation and Classification

A very appealing strategy for dealing with the intermittence is temporal non-overlapping aggregation. This approach refers to the transformation of the original series to alternative frequencies (for example, from monthly to quarterly frequency). Willemain et al. (1994) were the first to examine temporal aggregation on slow moving items, a research that was limited to a very small number of series. Nikolopoulos et al. (2011) presented more empirical evidence on its effectiveness, while at the same time proposing alternatives for reverting the aggregated forecasts back to the original level, if cumulative forecasts are not appropriate. They proposed the *Aggregate-Disaggregate Intermittent Demand Approach* (ADIDA). ADIDA is a forecasting framework where one should select an appropriate aggregation level, a suitable extrapolation method and a disaggregation mechanism. On top of its good forecasting performance, Babai et al. (2012) demonstrated that ADIDA can also enhance stock control performance, resulting in higher service levels while being more cost-efficient.

However, the selection of the appropriate aggregation level is not trivial. Nikolopoulos et al. (2011) identified empirically that on the dataset they explored a level around 8 periods gave the best forecasting performance, but did not prove that this is optimal. At the same time, Spithourakis et al. (2014), who provided some mathematical insights into the ADIDA framework, found that high aggregation levels suffer from excessive smoothing. Rostami-Tabar et al. (2013) proved theoretically and demonstrated empirically that, under the assumption of some stationary processes, temporal aggregation leads to performance improvements positively correlated to the aggregation level. However, they argued that one may not be able to consider very high aggregation levels due to data availability limitations and the requirements of forecasting methods. Therefore, the optimal aggregation level for ADIDA is still an open research question.

One of the greatest advantages of temporal aggregation is that different series characteristics are highlighted in each aggregation level (Andrawis et al., 2011), suggesting the application of methods with different features (Kourentzes et al., 2014). In the case of intermittent demand data, the average inter-demand interval and the coefficient of variation of the demands will change. The same concept was also discussed by Nikolopoulos et al. (2011). They proposed that methods originally designed for fast-moving data may be applied in higher aggregation levels of intermittent demand data, as the zero demands will practically have been removed.

Disadvantages linked with temporal aggregation include the excessive smoothing of the data (Spithourakis et al., 2014) and loss of information, as a direct result of the decreased number of available observations. The latter is of particularly importance when intermittent demand data are examined, where typically short series are available.

2.3. Combining Forecasts

Combining forecasts derived from alternative extrapolation methods is widely considered as beneficial (Bates and Granger, 1969; Makridakis and Winkler, 1983; Clemen, 1989). Averaging leads to more accurate and robust forecasts by reducing the uncertainty and the variance of forecasting errors (Hibon and Evgeniou, 2005). Complex weighting approaches have been recently examined by many researchers (for examples see: Kolassa, 2011). However, simple averages that use equal weights across considered methods have shown to be robust and reasonably accurate (for examples see: Timmermann, 2006; Jose and Winkler, 2008), compared to their more complex counterparts.

Even though the majority of the literature has focused on combining point forecasts derived from different forecasting approaches using the same data, another strategy would include the combination of forecasts calculated at different frequencies (Cholette, 1982; Trabelsi and Hillmer, 1989; Casals et al., 2009). This approach is intuitively appealing, aiming at capturing different dynamics of the data. Kourentzes et al. (2014) proposed the *Multiple Aggregation Prediction Algorithm* (MAPA) which combines the states of the series' components (level, trend and seasonality) across multiple levels of aggregation using the Exponential Smoothing family of methods. The application of MAPA resulted in improved forecasting accuracy, especially for the longer horizons. Combining the states of the components rather than the point forecasts is one of the main advantages of the MAPA approach. Whereas the level and the trend component are modelled in each level, seasonality is modelled only on the permitted aggregation levels (i.e. where the result of the division of periods per year and the aggregation level is integer and greater than one or, in other words, where non-fractional seasonality may be observed). As seasonality is generally not the case in the intermittent demand context, a simple combination of the derived point forecasts would be the direct equivalent of MAPA approach for the intermittent demand.

In order to produce point forecasts at each aggregation level, the same extrapolation method may be used. Nikolopoulos et al. (2011) and Spithourakis et al. (2011) showed that temporal aggregation improves the average performance of a single method, referring to this strategy as a *self improving mechanism*. Alternatively, as discussed in subsection 2.2, in the case of intermittent demand data, intermittence and variance of data will change along with the frequency. This motivates appropriate model selection for each aggregation level.

3. Experimental Design

3.1. Empirical Data

The forecasting performance of combining approaches versus single methods will be tested using a data set containing 5,000 SKUs coming from the RAF. The same data set was used in earlier studies (Teunter and Duncan, 2009; Syntetos et al., 2009; Nikolopoulos et al., 2011). The original data cover a length of 7 years in monthly frequency, thus 84 observations per series. We split each time series into two sets of observations. The first 6 years (72 periods) are treated as the in-sample set, which is used for initialising the different methods and fitting the models. The 72^{th} period is

the forecast origin, from which 12-step-ahead point forecasts are produced. The performance of the different methods and combination approaches is measured by comparing the last year of data (periods 73 through 84), which was treated as the withheld out-of-sample set.

The relatively long available history of the in-sample (6 years of monthly observations, or 72 data points) allows for transformation through temporal aggregation, as sufficient data points will still be available in lower frequencies (i.e., higher aggregation levels). While the empirical optimal aggregation level across this data set was found to be 8 or 9 (Nikolopoulos et al., 2011), depending on the extrapolation method, this varies for each series. Based on these findings, we set the maximum aggregation level for this study equal to 12 periods. All frequencies from monthly down to yearly are considered, thus the available in-sample data points range from 72 points (for aggregation level equal to 1, or monthly frequency) down to 6 points (for aggregation level equal to 12, or yearly frequency). Finally, we excluded the series that contain less than 4 non-zero demands at any aggregation level, as to have sufficient data points to fit the sub-series for both the demands and the intervals. Therefore, only 3,810 SKUs are considered. Their descriptive statistics are presented in table 1. More details on how the temporal aggregation (frequency transformation) is performed in practice are given in Section 3.3.

Table 1: Descriptive statistics of the of the examined data set.

3,810 SKUs	Demand sizes (units)		Demand intervals (months)		Demand per period (units/month)	
	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
Min	1.00	0.00	3.79	1.14	0.05	0.21
25%ile	1.58	0.83	6.82	4.96	0.18	0.57
Median	4.00	3.19	8.29	6.25	0.40	1.57
75%ile	11.56	9.78	10.20	7.78	1.25	4.82
Max	483.56	655.13	21.33	14.08	65.08	252.99

On top of the experimental analysis performed using a static forecast origin and multiple steps ahead, we replicated the experiment using a rolling origin set up with lead time (forecasting horizon) equal to 3 periods. Moreover, as short series may be the case in many intermittent demand data sets, we also considered the use of just the first 36 observations as the in-sample set (three years of available information), producing again 12 out-of-sample point forecasts for each series. In both cases, the results provided the same insights, suggesting robustness of the analysis and the relative performance of the considered approaches. Thus, we present only the results obtained using the 7-year in-sample length and static origin.

3.2. Measuring Performance

Five different errors measures are used to assess the performance of the alternative approaches considered. First, the average signed error or mean error of the forecasts is measured, as to determine if an examined approach is consistently positively or negatively biased. In order to be able to average across all series, we scale the out-of-sample mean error of each series using the mean value of all in-sample periods. Thus, the scaled Error (sE) for the series i and horizon h is given by:

$$sE_{i,h} = \frac{y_{N+h} - \hat{y}_h}{\frac{1}{N}\sum_{t=1}^{N} y_t}, \tag{5}$$

where N is the number of in-sample observations, y_{N+h} is the actual value of the h^{th} out-of-sample period and $\hat{y}_h$ is the h-steps-ahead forecast. The sE are averaged across all horizons and series as to provide the reported scaled Mean Error (sME).

Accordingly, we measure the accuracy of the estimates using absolute errors instead of signed ones. The scaled Absolute Error (sAE) for the series i and horizon h is calculated as:

$$sAE_{i,h} = \frac{\mid y_{N+h} - \hat{y}_h \mid}{\frac{1}{N}\sum_{t=1}^{N} y_t}. \tag{6}$$

Once again, the scaled Mean Absolute Error (sMAE) is calculated as the average of sAE across all series and horizons.

In order to be able to evaluate the performance in terms of variance, we calculate a scaled variant of the Mean Squared Error, where the denominator is calculated as the squared average of the actual demands. The scaled Squared Error (sSE) for the series i and horizon h is defined as:

$$sSE_{i,h} = \left(\frac{y_{N+h} - \hat{y}_h}{\frac{1}{N}\sum_{t=1}^{N} y_t}\right)^2. \tag{7}$$

The scaled Mean Squared Error (sMSE) is calculated as the average of sSE across all series and horizons.

However, traditional error metrics have been criticised in the intermittent demand context (Teunter and Duncan, 2009; Wallström and Segerstedt, 2010). The bias measure may collapse under certain circumstances, while Mean Squared Error and, most prominently, Mean Absolute Error focus on periods with zero demand, favouring distorted forecasts (Wallström and Segerstedt, 2010). This is especially relevant for data with high degree of intermittence.

To address these limitations, we use a newly introduced bias measure called *Periods in Stock* (PIS) (Wallström and Segerstedt, 2010). PIS measures the number of periods a single unit of a SKU has spent in the stock or the number of stock-out periods. Thus, the measurement of error is enhanced by the dimension of time, via a double cumulative summation. Originally, Wallström and Segerstedt (2010) suggested a signed version of this metric, which for series i is defined as follows:

$$PIS_i = -\sum_{h=1}^{H}\sum_{j=1}^{h}(y_{N+j} - \hat{y}_j), \tag{8}$$

where H is the length of the required forecasting horizon. Positive values imply that stock is left over, while negative ones refer to stock-outs. In order to make the measure scale independent, it is scaled by $\frac{1}{N}\sum_{t=1}^{N} y_t$. So, the scaled *Periods in Stock* (sPIS) for series i is given by:

$$sPIS_i = \frac{PIS_i}{\frac{1}{N}\sum_{t=1}^{N} y_t}. \tag{9}$$

sPIS acts complementary to sE in providing evidence for systematic behavior with regards to the direction of the forecast errors. The absolute value of sPIS is also considered as a measure of the magnitude of the bias. The scaled Absolute *Periods in Stock* (sAPIS) for series i is given by:

$$sAPIS_i = \frac{\mid PIS_i \mid}{\frac{1}{N}\sum_{t=1}^{N} y_t}. \tag{10}$$

The scaled Mean *Periods in Stock* (sMPIS) and the scaled Mean Absolute *Periods in Stock* (sMAPIS) across all series are calculated as the simple average of sPIS and sAPIS respectively.

3.3. Forecasting and Combining

Three combination approaches are considered in the current study. First, we examine the combination of forecasts derived from different forecasting methods using the original sampling frequency of the time series. This is the simplest and most widely applied approach for forecast combinations and has provided good results for fast-moving data across different studies. Second, we create combinations of forecasts calculated using a single method and multiple sampling frequencies for each series. Third, we combine forecasts produced using multiple forecasting methods and multiple frequencies of the data. Especially for the latter case, apart from the obvious approach of combining multiple methods applied on multiple frequencies (each method applied on each frequency), we also examine the use of method selection schemes, so that only one (the most suitable) method is applied at each aggregation level. In all cases, equal weights for combining the point forecasts have been considered.

For combining across multiple frequencies, temporal non-overlapping aggregation is used in order to transform the original monthly frequency of the data into alternative ones. We consider all aggregation levels up to 12 periods. The very first observations of the series may have to be truncated appropriately for some aggregation levels, as to form equal time buckets. For example, given that the in-sample consists of 72 periods, when the aggregation level is set to 5 the first 2 data points are dropped and the remaining 70 are divided in equal buckets of 5 observations each in order to create 14 aggregated observations. Figure 1 provides a visual presentation of the non-overlapping temporal aggregation process. The gray shaded areas highlight the observations from the beginning of the series that have been dropped from some levels. The extrapolation takes place at the aggregated level and the cumulative forecast is disaggregated using equal weights which has proved to be a good strategy in previous studies (Nikolopoulos et al., 2011; Spithourakis et al., 2011). In fact, such an approach assumes the absence of any repeating patterns across the time buckets, such as seasonality. So, the demand pattern of each time bucket is regarded as independent.

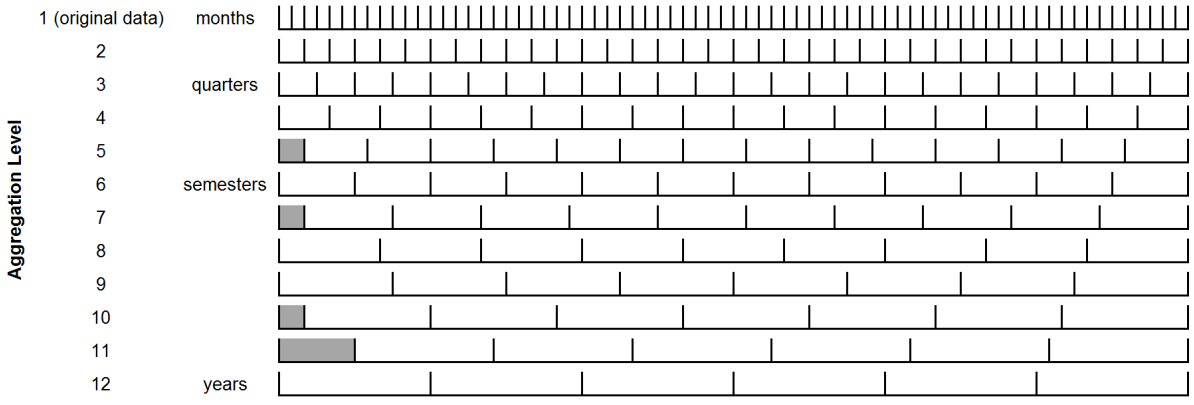

Figure 1: Temporal non-overlapping aggregation for multiple levels. The gray areas highlight observations that are not used, in order to have complete aggregation buckets.

To select the most appropriate method in each aggregation level, there are established classification schemes for intermittent time series. Syntetos et al. (2005) offer such a scheme, which

selects the best method amongst Croston's method and SBA based on the intermittence of the data and the coefficient of variation of the demands. This scheme was later revised by Kostenko and Hyndman (2006). An empirical analysis of the original and the revised schemes suggested that, despite of the former being simpler, the latter results in an overall superior forecasting performance (Heinecke et al., 2013). Hereafter, the revised classification scheme will be considered and is referred to as SBC-KH.

As the aggregation level increases, the intermittence of the data decreases. There is a good chance this will result in series containing no zero demands for the higher aggregation levels (where the mean intermittent demand interval will be equal to unity). In this special case, we propose that SES shall be used instead of approaches specifically designed for intermittent demand. One could argue that this refinement is not necessary, as Croston's method is equivalent to SES in the case where all periods have non-zero demands. However, this is not true for higher coefficients of variation, where under this classification scheme SBA is selected. With SBA, a damping factor is multiplied on all point forecasts and, therefore, it is not equivalent to SES. Moreover, we propose the use of an optimised smoothing parameter for the level of SES, rather than a static value as widely suggested for Croston's method and SBA. This is the norm for fast-moving consumer goods that typically have continuous (non-zero) demand (Hyndman et al., 2008). Hereafter, we will refer to the proposed classification as SBC-KH-SES.

In order to examine the efficiency of combining in the case of intermittent demand, widely used parametric approaches will be considered as benchmarks. These include Croston's method and SBA, as well as simpler approaches such as Naive, MA and SES. In terms of parameters for Croston's method and SBA, the smoothing parameters are set to 0.1 for updating both the demands and intervals, while initial estimates for the levels are calculated through averaging across all available points. The number of periods to be considered in the case of the moving average is set to 6 for all aggregation levels, corresponding to the maximum possible selection across all aggregation levels. The best SES method for each series was estimated using the `ets()` function available in the *forecast* package by Hyndman and Khandakar (2008) for the R Project (R Core Team, 2012). Lastly, the ADIDA framework is considered, using a pre-defined aggregation level of 8 periods applied for all series. This aggregation level has led to the best performance improvements for the SBA estimator for the same data set in a previous study by Nikolopoulos et al. (2011). Note that the best aggregation level is typically not known in advance, however in this case we make use of this information to ensure the good performance of the benchmarks.

4. Empirical Evaluation and Discussion

In this section, the results on the empirical performance of the alternative approaches will be presented and discussed. Table 2 provides the summarised results for all six approaches considered in this paper. The first column provides a quick description of the approach. The first two approaches refer to the application of a single method on a single frequency. The first one applies an extrapolation method directly on the original (observed) frequency, being the most widely implemented approach in theory and practice. The second one is the ADIDA framework (Nikolopoulos et al., 2011; Babai et al., 2012), using a fixed level of aggregation across all series. The rest refer to the possible combination strategies: simple combination of methods applied on the original frequency, combination of forecasts derived from multiple frequencies by applying a single method, combination of forecasts derived from multiple frequencies and methods, and combination of forecasts derived from multiple frequencies by applying appropriate forecasting methods

for each aggregation level as indicated by the classification schemes. The second column presents the forecasting methods or classification schemes considered in each approach. The third column refers to the aggregation levels taken into account, which are directly linked to the frequency transformation. Lastly, columns 4 to 8 provide the forecasting performance as measured by five metrics, sME, sMAE, sMSE, sMPIS and sMAPIS.

Focusing on the performance of the single methods, the results are mixed across the different metrics considered. Methods specifically designed for intermittent demand (Croston and SBA) perform worse than simpler benchmarks, such as MA and SES. However, SBA succeeds in improving over Croston's methods for all metrics considered, especially in terms of bias. MA provides the least biased forecasts (based on the sME), while it fails according to the sMSE and sMAPIS metrics. Naive's performance seems also adequate, according to sMAE and sMPIS. However, this is driven from the fact that this method provides zero estimates for the majority of the series. This leads to perfect forecasts, as the RAF data set is overwhelmed by zero demands (the median value of demand intervals is 8.29, according to table 1). On the other hand, the positive value of sME indicates under-stocks, as $error = actual - forecast$, while Naive has the worst performance as measured by the sMAPIS. If we consider the average rank of the methods across all metrics, MA is regarded as the best single approach, closely followed by Naive, SES and SBA. On the other hand, Croston's method has overall the worst performance.

Aggregating using only a single frequency leads to improvements for almost all methods and all metrics considered. This result confirms previous studies (Nikolopoulos et al., 2011). In more detail, Croston's method and SBA are improved for all five metrics, while Naive and MA are improved for three metrics each. Furthermore, the value of sME for Naive is now negative, as the aggregation level of 8 periods removes intermittence for almost half of the series. On the other hand, the ADIDA framework does not lead in improvements for the case of SES, apart from a small decrease in the sMSE.

Next we consider the simple combinations of methods applied on the original frequency. As observed in almost all cases, the forecasting performance of the combinations is worse than the best performance among the individual methods considered. The same is true for all two-method combinations which do not include Croston's method or SBA and are not reported on table 2. As a result, simply combining the outputs of the methods applied on the original data does not lead to improved forecasting performance in the context of intermittent demand. However, combinations that include the Naive method produce less biased estimates, as Naive is the only method that gives negatively biased estimates.

It is worthwhile to try to provide some insights why the combination of point forecasts derived from different methods does not work in the intermittent demand context. Let us consider the simplest case where the point forecasts of two methods are combined with equal weights. In order for this to improve forecasting performance, the point forecasts of the two methods should lie in the opposite sides of the future actual values. Otherwise, if the point forecasts are both lower or greater than the actual, then the final forecast error is equal to the simple average of the forecast error of the two methods. While in the case of fast-moving demand, methods with different modelling features (level, trend, seasonality) will have sufficiently different point forecasts, forecasts produced by intermittent demand methods are generally close in value, as these methods model only the level of the demand rate. In the majority of cases, these demand rate forecasts are collectively either underestimating the observed demand, when a non-zero demand is recorded, or overestimating it, when the observed demand is zero due to the intermittent nature of the data. Therefore, combining

Table 2: Empirical performance results for the different approaches considered.

Approach	Method(s) or Classification Scheme	Aggregation Level(s)	sME	sMAE	sMSE	sMPIS	sMAPIS
Single Method, Original Frequency	Naive	1	0.134	1.511	77.53	-8.90	113.20
	MA	1	-0.118	1.697	67.97	10.75	98.07
	SES	1	-0.161	1.713	66.14	14.08	78.87
	Croston	1	-0.232	1.770	65.82	19.58	80.63
	SBA	1	-0.177	1.724	65.80	15.33	78.48
Single Method, Transformed Frequency (ADIDA)	Naive	8	-0.119	1.694	67.27	10.84	93.17
	MA	8	-0.129	1.685	65.86	11.56	77.95
	SES	8	-0.188	1.738	66.03	16.51	81.01
	Croston	8	-0.197	1.741	65.81	16.89	79.38
	SBA	8	-0.144	1.697	65.79	12.78	77.37
Combination of Methods, Original Frequency	Croston, SBA	1	-0.204	1.747	65.81	17.46	79.54
	Naive, Croston, SBA	1	-0.092	1.661	67.09	8.67	82.24
	MA, Croston, SBA	1	-0.176	1.728	66.05	15.22	80.32
	SES, Croston, SBA	1	-0.190	1.735	65.84	16.33	79.13
	Naive, MA, SES, Croston, SBA	1	-0.111	1.677	66.56	10.17	80.64
Single Method, Multiple Frequencies	Naive	1-12	-0.091	1.672	67.55	8.59	92.54
	MA	1-12	-0.121	1.681	65.94	10.96	77.87
	SES	1-12	-0.179	1.730	65.93	15.69	79.68
	Croston	1-12	-0.203	1.746	65.80	17.37	79.45
	SBA	1-12	-0.150	1.701	65.79	13.23	77.41
Combination of Methods, Multiple Frequencies	Croston, SBA	1-12	-1.777	1.723	65.79	15.30	78.42
	Naive, Croston, SBA	1-12	-0.148	1.704	66.00	13.06	78.91
	MA, Croston, SBA	1-12	-0.158	1.709	65.81	13.85	77.77
	SES, Croston, SBA	1-12	-0.178	1.725	65.81	15.43	78.62
	Naive, MA, SES, Croston, SBA	1-12	-0.149	1.705	65.94	13.17	78.58
Selection Scheme, Original Frequency	SBC-KH	1	-0.177	1.724	65.80	15.33	78.48
Selection Scheme, Multiple Frequencies	SBC-KH	1-12	-0.154	1.704	65.79	13.56	77.56
	SBC-KH-SES	1-12	-0.144	1.696	65.79	12.78	77.32

such forecasts does not allow for the benefits of combination to materialise, in contrast to empirical studies on regular data.

On the other hand, combining forecasts derived from the same single method applied on multiple frequencies of the same data proves to be more promising. We observe that Naive, MA and SES on multiple levels perform better than the respective ADIDA applications for most of the metrics considered. In fact, simple combination of the Naive method applied across multiple instances of the same data offers the least biased estimates compared to any other approach in the current study. Croston's method and SBA performance is slightly worse but very close to the one achieved by the respective methods using the ADIDA framework. This is a very useful result. To achieve ADIDA's performance a pre-defined aggregation level was used, which had been empirically determined in an ex-post manner by a separate study (Nikolopoulos et al., 2011). Here, comparable, if not better, performance is offered by the combined forecasts, without the need to identify and select the aggregation level hyper-parameter. In contrast to the case of combining different methods on the original frequency, combinations of the forecasts derived from a single method applied on multiple frequencies can provide improvements, as temporal aggregation results to sufficiently different forecasts.

Following the insights from the study of Nikolopoulos et al. (2011), temporal aggregation is a very promising approach for intermittent demand, as higher aggregation levels will have lower degree of intermittence, thus their behaviour will be more predictable. Theoretical insights on the improvements achieved by the use of temporal aggregation have been demonstrated by Rostami-Tabar et al. (2013). Managerial-driven (Nikolopoulos et al., 2011) and empirical-driven (Spithourakis et al., 2011) strategies for selecting the optimal aggregation level per series have been proposed. However, the advantage of considering multiple aggregation levels is that the need for selecting a single aggregation level is removed.

The use of combinations of methods applied on multiple frequencies improves the forecasting performance. In many cases, forecasting bias, accuracy and variance lies below the respective performance of single methods applied on the original frequency and are comparable to ADIDA's. Combination across both methods and frequencies also leads to improvements compared to the respective combinations of methods applied only on the original frequency. This indicates that the problem of selecting either forecasting method or aggregation level of the time series can be effectively removed through forecast combination. Given the sparsity of guidelines from the literature how to select the appropriate method for intermittent demand time series, this finding motivates further the use of combination in practice.

Lastly, the case of applying classification schemes is considered. We first provide the performance of SBC-KH selection scheme when only the original frequency of the data is considered, as a benchmark. Due to the high intermittence of the specific data set employed for this study, the SBA is always selected over the Croston's method. As a result, the figures for SBC-KH applied on the original data are identical to those of the SBA. We also consider selection schemes using data that are derived from multiple aggregation levels. Temporal aggregation affects the intermittence and the variance of the series, so an appropriate model is selected for each level of aggregation separately. Then, individual estimates across all frequencies are combined. Model selection is performed using SBC-KH and SBC-KH-SES. Both schemes give very good forecasting performance across all metrics considered. However, using the SES in the cases when all zero demands have been removed from the series provides the best results. In fact, the SBC-KH-SES scheme delivers the best forecasts across all approaches when Croston's method, SBA or SES are considered. It

is useful here to contrast the results with the findings from the combination of multiple methods over multiple frequencies. The results of the refined classification scheme are overall better than the best combination's result across randomly grouped methods, which in turn were better than the benchmarks'. This gives support to the proposed refinement of the classification scheme that selects across models intelligently, while still not requiring from the users to make ad-hoc selections on either between methods or hyper-parameters, such as the optimal aggregation level. Furthermore, this result gives more evidence on the validity of the original SBC-KH scheme. Note that this approach has many parallels with MAPA that was proposed for fast-moving items by Kourentzes et al. (2014).

Figure 2 presents the distributions of the selected methods for several aggregation levels. Graphs (a) through (d) refer to the SBC-KH classification scheme, where a cut-off line is used as the threshold for selecting amongst SBA and Croston's method. According to Kostenko and Hyndman (2006), an approximate rule for selecting Croston's method over SBA is given by $v \leq 2 - (3/2)p$, where v denotes the square of the coefficient of variation of the demands and p refers to the mean value of the intervals. Graphs (e) through (h) refer to the SBC-KH-SES, where we propose the use of SES in the case $p = 1$, irrelevant of the value of v. In lower aggregation levels, the high degree of intermittence leads to the selection of the SBA for the total number of series considered. However, as the aggregation level increases, intermittence decreases, thus Croston's method or SES are preferred. In the case of SBC-KH classification scheme, about 27% of the series are gradually modelled with Croston's method. On the other hand, in the case of SBC-KH-SES scheme, SES is considered for more than 40% of the series in the maximum aggregation level, with the role of Croston's method being limited. The differences in bias and accuracy between SBC-KH and the proposed SBC-KH-SES are explained in this plot. The former uses SBA even when $p = 1$, for the high aggregation levels, leading to biased forecasts. By allowing the time series to be modelled by the simpler SES method in the case $p = 1$ better results are achieved.

The box-plots of sAE for Croston's Method and SBA across all different approaches considered in this study are presented in figure 3. $ADIDA(x, y)$ refers to the application of a single x method on a single aggregation level y. $Comb(x_1, x_2, ..., x_n, y)$ denotes the simple combination of estimates derived by methods or classification schemes x_i using the frequency (or frequencies) of the data that correspond to the y^{th} aggregation level(s). We have indicated alternative approaches with different background colours. It is obvious that the enhanced forecasting performance achieved by combining forecasts leads to reduced variance of the scaled absolute errors, thus turning the estimates more robust. In fact, the interquartile range of sAE when using SBC-KH-SES is 22% and 18% lower than this of Croston's method and SBA respectively.

5. Conclusions and Implications

Forecasting for intermittent demand data has proved challenging in the past. The motivation of this study was to explore the application of combination schemes on the intermittent demand forecasting problem, a strategy that has been proved very valuable when forecasting fast-moving data. Apart from a single research where forecasts from different levels of hierarchies are combined (Moon et al., 2012), this has been overlooked in the literature. We explored the efficiency of forecast combinations in the intermittent demand context considering four approaches. First, estimates provided by different methods applied on the original data were combined. Second, we combined estimates calculated using a single method applied on different frequencies of the same data. Third, we investigated the combination of methods applied on multiple frequencies of the

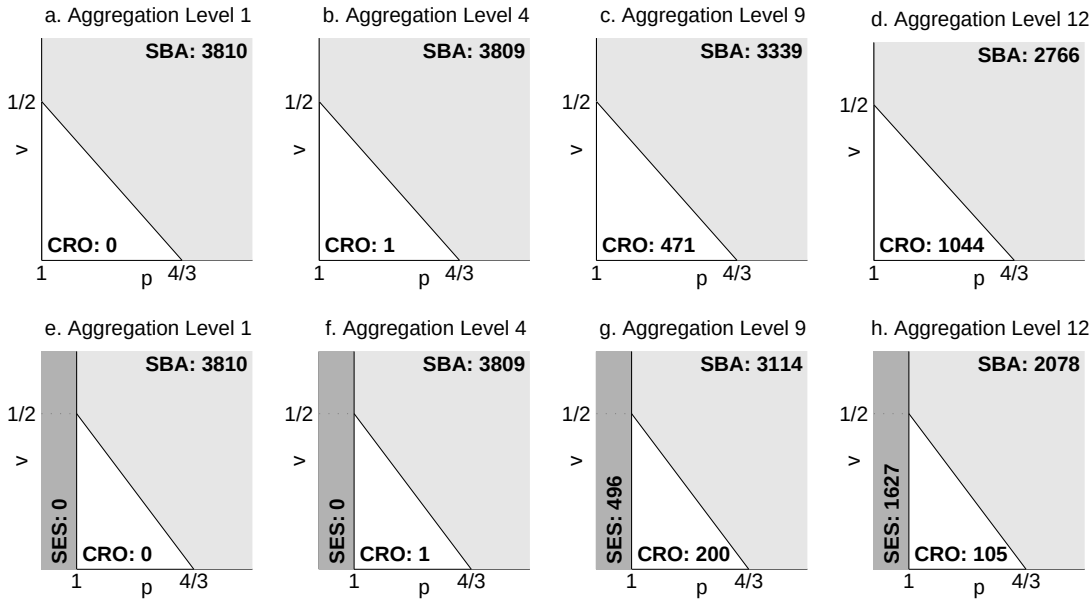

Figure 2: Selected models for the SBC-KH scheme (figures (a) to (d)) and the SBC-KH-SES scheme (figures (e) to (h)) across various levels of aggregation.

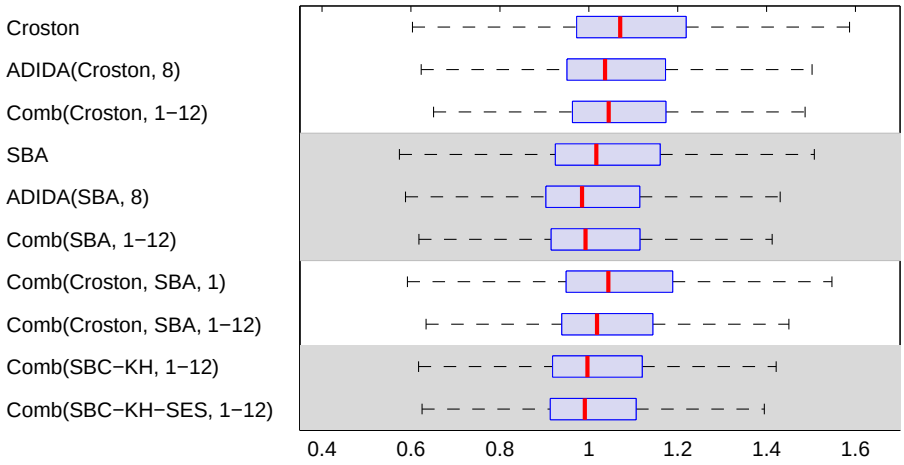

Figure 3: Box-plots of sAE for different approaches.

data. Fourth, forecasts derived by multiple aggregation levels using appropriately selected methods were averaged.

Empirical evidence on a large number of spare parts suggested that simply combining the outputs of multiple methods does not lead to improvements in terms of forecasting performance. Combination across forecasts derived from transformed frequencies using the same single method or multiple methods is efficient and improves the forecasting performance. The same is true when classification schemes are employed. At the same time, estimates based on multiple frequencies proved to be more robust, with lower interquartile ranges. Results were analysed in terms of different alternative and complementary metrics, corresponding to the bias, the accuracy, the variance and a proxy of the inventory performance. Benchmarks included simple methods (Naive, MA, SES), methods specifically designed for intermittent demand (Croston's method and SBA) and the ADIDA framework, which performed very well in previous studies (Nikolopoulos et al., 2011; Babai et al., 2012). Our overall finding is that combinations of forecasts derived from different frequencies are beneficial for intermittent demand.

A very interesting finding lies in the fact that the use of multiple aggregation levels removes the problem of appropriately selecting the hyper-parameter for the ADIDA framework, without harming accuracy. The good forecasting performance of ADIDA is based on correctly pre-selecting the aggregation level. Combinations across frequencies offer similar, if not better, performance without the need to select such a parameter, thus simplifying its implementation in real applications. A potential new problem is that of selecting a meaningful maximum aggregation level. Following the discussions in Spithourakis et al. (2014), Rostami-Tabar et al. (2013), and Kourentzes et al. (2014), we limit the maximum aggregation level so as to prevent over-smoothing and to allow enough data points to be available for model fitting purposes. Therefore, the maximum aggregation level is effectively constrained by the data features of the problem at hand.

Similarly, the problem of model selection for intermittent demand can be side-stepped through combination of methods. The SBC-KH classification provides guidance on how to select between Croston's and SBA methods. We proposed an extension to this classification as to include the special case where intermittence is removed when temporal aggregation is used. We suggest that in such cases data should be modelled with SES, using an optimised smoothing parameter, instead of Croston's method or SBA, regardless the coefficient of variation of the demands. The proposed SBC-KH-SES classification scheme provided the best overall forecasting performance in the current research and is our recommended approach for using method combination when dealing with intermittent demand problems. Crucially, this approach eliminates the manual choices of either method or aggregation level that a user has to consider in practice.

The results of this study are of direct interest for practitioners, as we demonstrate how to obtain in a straightforward manner better forecasts for intermittent demand data using established methods (Croston's method, SBA, and SES). Simple data manipulations (temporal aggregation) and equal weighted combinations of the output point forecasts can easily be employed in standard spreadsheets. Moreover this research has practical implications for the effective design of specialised forecasting support systems. We propose that such packages should introduce features with regards to the transformation of data using multiple aggregation levels, as well as integrate classification schemes for the intermittent demand context (SBC-KH and SBC-KH-SES).

Irrespective of the gains in accuracy and simplifying the forecasting process, combining forecasts demonstrated reduced error variance, which is expected to have positive implications for inventory management. This will be explored in more detail in future work. This paper focused

on the empirical evaluation of forecast combinations for intermittent demand data. However, the theoretical underpinnings of using combination approaches should be considered as to enable the identification of the most appropriate combination strategy for each case. In our evaluation we used unweighted combinations, motivated by their documented good performance. In future work we plan to examine complex weighting approaches for combining the forecasts. Lastly, a further extension on the classification scheme could examine the efficacy of alternative forecasting methods originally designed for fast-moving data. Once the intermittence of the data is removed, more complex time-series patterns may appear. This is turn may require forecasting methods that can deal with trend and seasonality, instead of SES that was found adequate for this case.

References

Andrawis, R. R., Atiya, A. F., El-Shishiny, H., 2011. Combination of long term and short term forecasts, with application to tourism demand forecasting. International Journal of Forecasting 27 (3), 870–886.

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.

Babai, M. Z., Syntetos, A. A., Teunter, R., 2011. Intermittent demand estimators: Empirical performance and sensitivity to the smoothing constants used. Working paper N°138-11, Centre de recherche de BEM.

Bates, J. M., Granger, C. W. J., 1969. The combination of forecasts. Operational Research Society 20 (4), 451–468.

Casals, J., Jerez, M., Sotoca, S., 2009. Modelling and forecasting time series sampled at different frequencies. Journal of Forecasting 28, 316–342.

Cholette, P. A., 1982. Prior information and ARIMA forecasting. Journal of Forecasting 1, 375–384.

Clemen, R. T., 1989. Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5, 559–583.

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

Eaves, A. H. C., Kingsman, B. G., 2004. Forecasting for the ordering and stock-holding of spare parts. The Journal of the Operational Research Society 55 (4), 431–437.

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

Gutierrez, R. S., Solis, A. O., Mukhopadhyay, S., 2008. Lumpy demand forecasting using neural networks. International Journal of Production Economics 111 (2), 409–420.

Heinecke, G., Syntetos, A., Wang, W., 2013. Forecasting-based SKU classification. International Journal of Production Economics 143 (2), 455–462.

Hibon, M., Evgeniou, T., 2005. To combine or not to combine: selecting among forecasts and their combinations. International Journal of Forecasting 21, 15–24.

Hyndman, R., Akram, M., Archibald, B., 2008. The admissible parameter space for exponential smoothing models. Annals of the Institute of Statistical Mathematics 60, 407–426.

Hyndman, R. J., Khandakar, Y., 7 2008. Automatic time series forecasting: The forecast package for R. Journal of Statistical Software 27 (3), 1–22.

Johnston, F. R., Boylan, J. E., 1996. Forecasting for items with intermittent demand. The Journal of the Operational Research Society 47 (1), 113–121.

Johnston, F. R., Boylan, J. E., Shale, E. A., 2003. An examination of the size of orders from customers, their characterisation and the implications for inventory control of slow moving items. The Journal of the Operational Research Society 54 (8), 833–837.

Jose, V. R. R., Winkler, R. L., 2008. Simple robust averages of forecasts: Some empirical results. International Journal of Forecasting 24, 163–169.

Kolassa, S., 2011. Combining exponential smoothing forecasts using Akaike weights. International Journal of Forecasting 27 (2), 238–251.

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

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

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.

Levén, E., Segerstedt, A., 2004. Inventory control with a modified Croston procedure and Erlang distribution. International Journal of Production Economics 90 (3), 361–367.

Makridakis, S., Winkler, R., 1983. Average of forecasts: Some empirical results. Management Science 29, 987–996.

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, 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.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., Nikolopoulos, K., 2014. 'Horses for Courses' in demand forecasting. European Journal of Operational Research 237, 152–163.

Petropoulos, F., Nikolopoulos, K., Spithourakis, G., Assimakopoulos, V., 2013. Empirical heuristics for improving intermittent demand forecasting. Industrial Management and Data Systems 113 (5), 683–696.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org/`

Rao, A., 1973. A comment on: Forecasting and stock control for intermittent demands. Operational Research Quarterly 24, 639–640.

Romeijnders, W., Teunter, R., van Jaarsveld, W., 2012. A two-step method for forecasting spare parts demand using information on component repairs. European Journal of Operational Research 220 (2), 386–393.

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

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

Snyder, R., 2002. Forecasting sales of slow and fast moving inventories. European Journal of Operational Research 140 (3), 684–699.

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, 485–496.

Spithourakis, G., 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.

Spithourakis, G., Petropoulos, F., Nikolopoulos, K., Assimakopoulos, V., 2014. A systemic view of ADIDA framework. IMA Management Mathematics 25, 125–137.

Syntetos, A., 2001. Forecasting for intermittent demand. Ph.D. thesis, Buckinghamshire Chilterns University College, Brunel University.

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

Syntetos, A. A., Boylan, J. E., May 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., 2006. On the stock control performance of intermittent demand estimators. International Journal of Production Economics 103 (1), 36–47.

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

Teunter, R., Sani, B., April 2009. On the bias of croston's forecasting method. European Journal of Operational Research 194 (1), 177–183.

Teunter, R., Syntetos, A., Babai, M., 2010. Determining order-up-to levels under periodic review for compound binomial (intermittent) demand. European Journal of Operational Research 203 (3), 619–624.

Teunter, R. H., Duncan, L., 2009. Forecasting intermittent demand: a comparative study. The Journal of the Operational Research Society 60, 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.

Timmermann, A., 2006. Forecast combinations. In: G. Elliott, C. G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Vol. 1. Elsevier, pp. 135–196.

Trabelsi, A., Hillmer, S., 1989. A benchmarking approach to forecast combination. Journal of Business and Economic Statistics 7, 353–362.

Wallström, P., Segerstedt, A., 2010. Evaluation of forecasting error measurements and techniques for intermittent demand. International Journal of Production Economics 128 (2), 625–636.

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.

Willemain, T. R., Smart, C. N., Shockor, J. H., DeSautels, P. A., 1994. Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. International Journal of Forecasting 10 (4), 529–538.