Distributions of forecasting errors of forecast combinations: implications for inventory management

Devon Barrow[a,*], Nikolaos Kourentzes[b]

[a]*Faculty of Business, Environment and Society, Coventry University*
Coventry University, Coventry, West Midlands, CV1 5FB, UK
[b]*Lancaster University Management School*
Department of Management Science, Lancaster, LA1 4YX, UK

Abstract

Inventory control systems rely on accurate and robust forecasts of future demand to support decisions such as setting of safety stocks. Combining forecasts is shown to be effective not only in reducing forecast errors, but also in being less sensitive to limitations of a single model. Research on forecast combination has primarily focused on improving accuracy, largely ignoring the overall shape and distribution of forecast errors. Nonetheless, these are essential for managing the level of aversion to risk and uncertainty for companies. This study examines the forecast error distributions of base and combination forecasts and their implications for inventory performance. It explores whether forecast combinations transform the forecast error distribution towards desired properties for safety stock calculations, typically based on the assumption of normally distributed errors and unbiased forecasts. In addition, it considers the similarity between in- and out-of-sample characteristics of such errors and the impact of different lead times. The effects of established combination methods are explored empirically using a representative set of forecasting methods and a dataset of 229 weekly demand series from a household and personal care leading UK manufacturer. Findings suggest that forecast combinations make the in- and out-of-sample behaviour more consistent, requiring less safety stock on average than base

[*]Correspondance: D Barrow, Strategy and Applied Management, Faculty of Business, Environment and Society, Coventry University, Coventry, West Midlands, CV1 5FB, UK. Tel.: +44-024-7765-7413

Email address: devon.barrow@coventry.ac.uk (Devon Barrow)

forecasts. Furthermore we find that using in-sample empirical error distributions of combined forecasts approximates well the out-of-sample ones, in contrast to base forecasts.

Keywords: Time Series, Forecasting, Combination, Inventory, Safety Stock

1. Introduction

The combination of multiple forecasts is a well-established procedure for improving forecast accuracy. The two key reported advantages are the reduction of both forecast error variance and reliance on a single forecast method (Clemen and Winkler, 1986; Timmermann, 2006). While there is general acceptance that forecast combination improves accuracy, there is limited research on its impact on the distribution of forecast errors (de Menezes et al., 2000), and even less on the impact this has on inventory (Chan et al., 1999).

In this work we examine the key properties of the forecast error distributions of both base and combined forecasts, and the impact combination has on setting the safety stock. We focus on safety stock as this affects the entire control system including total inventory, reorder levels, backorders, and stockouts. Many inventory control systems assume that forecast errors are normally distributed and unbiased. This is often violated in practice. We investigate whether the combination of forecasts leads to any improvement in the shape of the error distribution towards desired properties of normality and unbiasedness.

Specifically the contributions of this paper are as follows: (a) investigate the effect that forecast combinations have on the shape of the forecast error distributions, as measured in terms of bias, variance and deviation from normality; (b) compare base and combined forecasts error characteristics between in-sample, where inventory variables are estimated, and out-of-sample, where these are utilised to support decisions; and (c) evaluate the impact of combinations on safety stock.

The rest of the paper is organised as follows. Section 2 provides the research background of this work, motivating its research questions. Section 3 describes the various forecast combination methods that are considered. Section 4 presents the setup of the empirical evaluation, and section 5 discusses the results, followed by concluding remarks.

2. Background research

Forecast combination has been used successfully in many areas of research and practice including economics, meteorology, insurance and retail forecasting (Clemen and Winkler, 1986; Timmermann, 2006). Evidence from empirical several studies (Aksu and Gunter, 1992; Macdonald and Marsh, 1994; Elliott and Timmermann, 2004; Stock and Watson, 2004; Dekker et al., 2004; Clements and Hendry, 2007; Jose and Winkler, 2008; Guidolin and Timmermann, 2009; Andrawis et al., 2011; Kourentzes et al., 2014a), and forecasting competitions (Makridakis et al., 1982; Makridakis and Hibon, 2000) have been almost unanimous in concluding that combining forecasts improves forecasting accuracy.

Since the seminal paper of Bates and Granger (1969) there has been extensive research in combining forecasts. The main focus is proposing better combination methods and the evaluation has focused on improving forecasting accuracy. However evaluating the benefits of forecast combination only in terms of better point forecasts can be rather misleading (Chatfield, 1995, 1996; Fildes and Howell, 1979; Fildes, 1989) as it ignores all other information contained in the entire distribution of the forecast errors.

This falls short of what is required of inventory control systems, where the decision maker needs to take an explicit account of the risk and uncertainty associated with such forecasts (Chen et al., 2007; Gerchak and Mossman, 1992). The safety stock (SS) required for a given item is defined as:

$$SS = k\hat{\sigma}_L, \tag{1}$$

where k is the safety factor for achieving the target service level, typically calculated based on reference to the normal distribution, and σ_L is the standard deviation of forecast errors for the respective lead time L. The $\hat{\sigma}_L$ is typically estimated by calculating the respective Root Mean Squared Error (RMSE) of the forecasts. This is a reasonable estimate of the true standard deviation when the forecasts are unbiased. Hence, it is obvious that a good forecast for inventory purposes should be unbiased and its errors having minimum variance and deviation from normality.

The majority of the forecast combination literature has focused mainly on accuracy evaluation and there is limited research on the impact of combining on the overall distribution of forecast errors (the most notable papers being those by de Menezes and Bunn, 1993, 1998; de Menezes et al., 2000), and even fewer studies on the impact this has on inventory. Chan et al. (1999) find that combined forecasts outperform base forecasts for inventory management applications. However the evaluation is again focused on RMSE and does not consider the entire error distribution. Thus the aspects of bias and deviation from the assumed normality are not explored. Furthermore de Menezes et al. (2000) warn that forecast error variances should not be the single focus of attention when evaluating forecast combination for decision making under uncertainty. In this study we therefore consider the entire forecast error distribution, and in- and out-of-sample behaviour of different combination methods to understand the impact on inventory decisions.

This study expands on the work by Chan et al. (1999) by considering several alternative forecast combination schemes, making use of different aspects of the forecast errors of the individual forecasts. In that paper the authors evaluated the constrained OLS optimal method by Newbold and Granger (1974), albeit with different weight procedures based on fixed and rolling windows. Here we valuate combinations methods which are qualitatively different in their approach to estimating model weights including the Outperformance method by Bunn (1975) based on probabilities, several methods based on OLS regression, and multiple variants based on minimisation of the covariance matrix of forecast errors including the Optimal method. This extension is useful as the Optimal method is known to suffer from poor performance under certain conditions, but also as we directly evaluate the need for more complex combination methods over simpler ones.

The out-of-sample errors are of essence for inventory management. Since these are not available the in-sample errors are used instead as an approximation. The accuracy of this approximation is crucial for the quality of the inventory decisions. Makridakis and Winkler (1989) find differences between the properties of in- and out-of-sample errors to be quite large and variable. Their results suggest that even when in-sample one-step ahead errors satisfy the usual conditions of normality, independence and bias, the out-of-sample errors do not. This may result in over optimism with regards to the accuracy and uncertainty of forecasts, for example making confidence intervals too narrow (Makridakis et al., 1987). Makridakis (1986) and Makridakis and Winkler (1989) find little correlation, on average 0.2, between in-sample one-step ahead forecast accuracy and out-of-sample accuracy for lead times one to three steps ahead. Pant and Starbuck (1990) obtained similar results using the M-Competition data. The implications for calculating inventory safety stocks are obvious. Considering forecast combinations, a relevant question is whether they increase the quality of approximation of the out-of-sample error behaviour over base forecasts.

So far we ignored the additional uncertainty introduced by the lead time. A traditional approach is to assume that forecast errors are independent over time (Silver et al., 1998), and to approximate the lead time standard deviation in (1) by multiplying the lead time by the standard deviation of the one-step ahead forecast errors $\hat{\sigma}_1$ (Axsäter, 2006):

$$\hat{\sigma}_L = \sqrt{L}\hat{\sigma}_1. \tag{2}$$

This approach assumes rather importantly that forecast errors are uncorrelated with constant variance over time, both often violated in practice. In doing so it ignores potential covariance between errors of the different h-step-ahead forecasts and covariances due to the cumulative demand across lead times. While we do not claim that (2) is the only or best way to calculate $\hat{\sigma}_L$, it does illustrate the importance of understanding how the forecast error distribution changes over different lead times. Furthermore, if the assumptions of homoscedasticity and normality are lifted one would expect that the empirically calculated safety stock will differ from the one prescribed by the theoretical formulas. Given that forecasts combination alters the error distribution it is important to understand, when compared to base forecasts, how it performs over increasing lead time horizons and how empirical safety stock diverge from theoretical ones.

Understanding these aspects of forecast combinations will enable us to help managers decide whether combinations lead to better inventory decisions over base forecasts and to what extent.

3. Forecast combination methods

Forecast combination methods are based on in-sample forecast error variance minimization (Newbold and Granger, 1974; Min and Zellner, 1993), ordinary least squares (OLS) regression (Granger and Ramanathan, 1984; Macdonald and Marsh, 1994), Bayesian probability theory (Bunn, 1975; Bordley, 1982; Clemen and Winkler, 1986; Diebold and Pauly, 1990), regime switching and time varying weights (Diebold and Pauly, 1987; Elliott and Timmermann, 2005; Lütkepohl, 2011; Tian and Anderson, 2014), Akaike weights (Kolassa, 2011), meta-learning (Lemke and Gabrys, 2010), computational intelligence methods e.g. artificial neural networks (Donaldson and Kamstra, 1996), and countless other innovations.

Following Newbold and Granger (1974), all methods can be expressed as a linear combination such that:

$$\hat{y}^c_{mt} = \sum_{m=1}^{M} w_{mt}\hat{y}_{mt} = \mathbf{w}'_t\hat{\mathbf{y}}_t, \tag{3}$$

where $\hat{\mathbf{y}}_t$ is the column vector of one-step-ahead forecasts $(\hat{y}_{1t}, \hat{y}_{2t}, \hat{y}_{3t}, \ldots, \hat{y}_{Mt})$ at time t produced by the m^{th} forecasting method, and $\mathbf{w}_t$ is the column vector of weights for the set of M forecasting methods $(w_{1t}, w_{2t}, w_{3t}, \ldots, w_{Mt})$.The weights w_{mt} will generally depend on the historical accuracy of base forecasts. Therefore when forecasting at time t, we use all observations prior to t to estimate both base forecast model parameters and forecast weights. In the methods described below, weights obtained in forecasting at time t are also utilised in forecasting multiple steps ahead.

In this paper we focus on a number of methods outlined below that are a good representation of different degrees of sophistication and common practice.

3.1. Simple average and median

This is the simplest of all the forecast combination methods. It is popular due to its ease of implementation, robustness, and a good record in economic and business forecasting (Jose and Winkler, 2008; Timmermann, 2006). The simple average forecast is obtained by setting all weights $w_m = M^{-1}$. This will be referred to in the empirical evaluation as *Mean*. The simple average is sensitive to outliers and assumes symmetric distributions. Alternative combination operators such as the median and the mode can be used. The former is less sensitive to outliers. We will refer to this as *Median*. The mode is insensitive to either outliers or lack of distribution symmetry, but has been shown to require about 30 or more forecasts to function well (Kourentzes et al., 2014a) and therefore will not be used here.

3.2. The Optimal method

This method provides optimal weights in the sense that the variance of the combined forecast error is minimised, while producing an unbiased combined forecast. The error variance at time t is minimized with weights $\mathbf{w}_t$ determined according to the formula:

$$\mathbf{w}_t = \frac{S^{-1}I}{I'S^{-1}I}, \tag{4}$$

where I is an m dimensional column vector of ones, and S is the covariance matrix of the one-step-ahead forecast errors. We will refer to this as *Optimal*.

3.3. Optimal with independence assumption

When forecasts are (assumed) independent the diagonal of S is sufficient. This mitigates estimation issues when only short time series are available (Bates and Granger, 1969; Newbold and Granger, 1974). We will refer to this in the empirical evaluation as *Optimal adaptive*.

3.4. Optimal with restricted weights

Another variation on the optimal method is that weights must belong to the interval $[0, 1]$. We will refer to this as *Optimal adaptive RW*.

3.5. Regression

In this method, actual values of the time series are regressed on the base forecasts with the inclusion of an intercept

$$y_t = w_0 + \mathbf{w}'_t \hat{\mathbf{y}}_t + \varepsilon_t. \tag{5}$$

The coefficients are used as combination weights. Granger and Ramanathan (1984) argued that this method guarantees unbiased combined forecast. This will be referred to as *Regression*.

3.6. Regression with restricted weights

Granger and Ramanathan (1984) shows that a constrained regression (weights restricted to sum to one) with the constant suppressed is equivalent to the optimal method. Granger and Ramanathan (1984) suggests the variant of employing a constrained least squares regression with the inclusion of a constant

$$y_t = w_0 + \mathbf{w}'_t \hat{\mathbf{y}}_t + \varepsilon_t, \quad \text{s.}t. \;\; \mathbf{w}'_t I = 1. \tag{6}$$

We will refer to this as *Regression RW*.

3.7. Outperformance

One of the first attempts at using Bayesian analysis for forecast combination was by Bunn (1975). In this method each weight is interpreted as being the probability that the corresponding one-step-ahead forecast outperforms all others as measured by the absolute error. This easy to implement robust nonparametric method is attractive due to its intuitive interpretation, the ability to incorporate expert judgement through priors, and its robust performance particularly when there is relatively little past data. We will refer to this as *Outperformance*.

3.8. Bates methods

In our analysis we include the five methods of Bates and Granger (1969). For each forecast m: $E_{mt} = \sum_{i=t-\nu}^{t-1}(e_{mt})^2$, where e_{mt} are the forecasts errors at time t. The weight of each forecast is calculated as

$$w_{mt} = E_{mt} / \sum_{m=1}^{M} E_{mt}.$$

This constitutes method *Bates I*. In the second variant, *Bates II*, the weights are generated using

$$w_{mt} = \alpha w_{mt-1} + (1-\alpha)\frac{E_{mt}}{\sum_{m=1}^{M} E_{mt}},$$

where $\alpha \in (0,1)$. In the third variation, *Bates III*, $S_m^2 = \sum_{i=1}^{t-1} w^t (e_{mt})^2$ is estimated, which for $w > 1$ assigns greater weight to more recent errors than ones further in the past. The weight of each forecast becomes

$$w_{mt} = S_m^2 / \sum_{m=1}^{M} S_m^2.$$

The first three methods utilise the variance of forecast errors. *Bates IV* utilises the weighted covariance $C = \sum_1^{t-1} w^t e_{it} e_{jt}$, with $i = 1, 2, 3, \ldots, M$ and $j = 1, 2, 3, \ldots, M$. The weight given to forecast m at time t is $w_{mt} = S_m^2 - C / \sum_{m=1}^{M} S_m^2 - mC$.

The final method, *Bates V*, utilises an exponentially smoothed weighting based on the absolute error of the forecast

$$w_{mt} = \alpha w_{mt-1} + (1-\alpha)\frac{|e_{mt-1}|}{\sum_{m=1}^{M} |e_{mt-1}|}, \tag{7}$$

where α is a smoothing constant between one and zero. The reader is asked to refer to this paper for full details of each method.

4. Empirical evaluation

4.1. Dataset

To empirically evaluate the effect of forecast combination we use a set of 229 products from a major UK fast moving consumer goods manufacturer. The manufacturer specialises in the production of household and personal care products. For each product there are 173 weekly sales observations. The historical data are separated into an in-sample estimation set of 104 weekly observations allowing a reasonable estimate of any seasonal effects when present in the data, and the remaining 69 observations are used as a test set, where out-of-sample forecasts will be evaluated. About 21% of the time series are identified as trended and none as seasonal. Trend was identified using the nonparametric Cox-Stuart test on a centre moving average estimate of each series. Fig. 1 provides representative examples of the time series in the dataset.

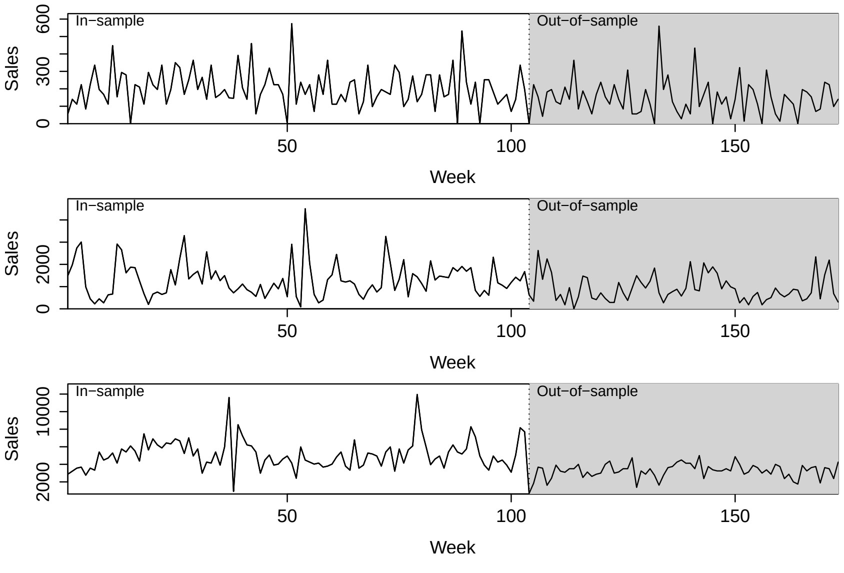

Figure 1: Example time series.

4.2. Forecasting Methods

To conduct the empirical evaluation a number of forecasting methods are used to produce base forecasts. These are subsequently combined using the methods in section 3.

Naïve

The random walk forecast, hereby referred to as *Naïve*, is a fundamental benchmark that requires no parameter identification, and should be outperformed by more complex methods to warrant the additional complexity. Given the most recent actual observation y_t the forecast for h steps ahead is calculated as $\hat{y}_{t+h} = y_t$.

Exponential Smoothing (ETS)

Exponential smoothing methods model the various structural components of a time series: level, trend and season, which may interact with each other in an additive or multiplicative way. Hyndman et al. (2002) proposed a state space formulation, which provides a statistical framework for estimating model parameters, choosing between alternative forms and constructing prediction intervals. The reader is referred to Hyndman et al. (2008) for a detailed description of the model, which we refer to as *ETS*. Here the appropriate model is chosen using the Akaike's Information Criterion (AIC) (Hyndman et al., 2002).

Autoregressive models (AR)

Autoregressive models, named here *AR*, attempt to capture the time series dynamics in a regression framework, thus having a different information base than *ETS*. This is useful so as to provide a variety of forecasts for the combinations. The first p lags of the time series are used as explanatory variables. Seasonal lags are permitted, which are of order P. The resulting forecast is:

$$\hat{y}_{t+h} = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{P} \beta_j y_{t-(js)}, \tag{8}$$

where s is the seasonal length of the time series and α_i and β_j the autoregressive coefficients. In case of nonstantionary time series differencing may be used first. To identify the order of p and P, as well as the need for differencing we follow the methodology proposed by Hyndman and Khandakar (2008). This uses the KPSS and OCSB tests to decide the order of first- and seasonal-differencing respectively and the autoregression order is identified using a stepwise procedure based on AIC.

Autoregressive Integrated Moving Average model (ARIMA)

Autoregressive Integrated Moving Average models extend *AR* to include moving average components. These models, although statistically elegant, are regarded as hard to specify and have not been widely applied in practice. Here we will be using the methodology by Hyndman and Khandakar (2008) to specify the models. In addition to the steps considered in specifying the *AR* models, the order of the moving average is also identified through a stepwise process. We refer to them as *ARIMA*.

Theta method

The Theta method, referred to as *Theta* in our results, was initially proposed by Assimakopoulos and Nikolopoulos (2000) as a decomposition method, but latter shown by Hyndman and Billah (2003) to be in its most basic form, equivalent to a single exponential smoothing with drift. *Theta* can capture seasonality by first de-seasonalising the time series. It has been shown to perform very well in multiple empirical evaluations and specifically in the M3 competition, one of the most well known forecasting competitions, where it ranked overall best (Makridakis and Hibon, 2000).

Multiple Aggregation Prediction Algorithm (MAPA)

This method employs multiple temporal aggregated series of the original time series to achieve better estimation of the various time series components. It was proposed by Kourentzes et al. (2014b) who argued that since by temporally aggregating a time series different structural components are attenuated or strengthened; a series should be modelled across multiple aggregation levels. This way a more holistic identification and estimation of the time series components can be achieved. The combination of the various estimates from the different aggregation levels is done by time series components, which makes *ETS* a natural model to use at each level. The authors showed that this approach resulted in substantial performance improvements over conventional modelling, while at the same time increasing the robustness of the forecasts to model misspecification. Petropoulos and Kourentzes (2014) found similar findings for slow moving items. Here we used *MAPA* with mean combination of the components at the different temporal aggregation levels, as described in detail by Kourentzes et al. (2014b).

All forecasts were constructed using the R statistical package (R Core Team, 2012). *ETS*, *AR* and *ARIMA* are built using the forecast package (Hyndman, 2014). *Theta* is build using the TStools package (Kourentzes and Svetunkov, 2014) and *MAPA* using the MAPA package (Kourentzes and Petropoulos, 2014).

4.3. Experimental setup

For each time series all forecasting methods are fitted using the first 104 observations of the series, and the in-sample errors calculated as the difference between the historical and the fitted values, providing a distribution of in-sample errors. Based on the fitted values and errors, the various combination weights are calculated for the different methods as outlined in section 3. This allows calculating the in-sample fit of the combined forecasts. Subsequently, a rolling origin evaluation is performed on the remaining out-of-sample observations. The manufacturer of our case study is interested in both short and medium term forecasts, therefore we consider the following forecast horizons: t+1, t+3 and t+5 for the out-of-sample period. We also track t+1 in-sample forecast errors. Forecasts from the individual base methods and the combinations of their forecasts are calculated, providing the respective out-of-sample error distributions for each model and time series.

We measure the forecast bias and error using scaled errors (sE) and scaled squared errors (sSE):

$$sE_t = \frac{y_t - \hat{y}_t}{\sum_{i=1}^{n} y_i}, \tag{9}$$

$$sSE_t = \frac{(y_t - \hat{y}_t)^2}{\sum_{i=1}^{n} y_i}, \tag{10}$$

where the denominator is the mean level of the time series and is used to make the errors and squared errors scale independent in order to be able to summarise the results across time series and forecast origins. We do that by calculating the mean and the median of the above metrics, resulting in the scaled mean error (sME) and scaled median error (sMdE) to measure forecast bias and scaled mean squared error (sMSE) and scaled median squared error (sMdSE) to measure the magnitude of forecast errors. For the latter two, we can calculate their square root, resulting in sRMSE and sRMdSE respectively. We adopt these scaled errors instead of percentage errors because our time series contain periods with zero observed demand. Furthermore, we focus on the RMSE because under zero or small bias it approximates the variance of the distributions, the determining factor of the size of safety stock.

5. Results

5.1. Forecasting accuracy and bias

First we present the forecast accuracy and bias results. Tables 1 and 2 summarise the sME/sMdE and the sRMSE/sRMdSE respectively. Values in brackets refer to the median metrics, and the rest to the mean metrics. Each column refers to a specific forecast horizon and the best base and combined forecasts are highlighted in boldface. The best forecast overall in each column is underlined.

In both tables we can observe substantial differences between the sME or sMSE and their median counterparts, providing some evidence that the error distributions deviate from normality. For the base methods *Theta* and *MAPA* perform best in terms of bias and accuracy. This is consistent with the literature (Makridakis and Hibon, 2000; Kourentzes et al., 2014b). Fot the combined forecasts, the best performing ones are the *Median*, *Optimal adaptive RW* and *Regression RW*, depending on the use of mean or median errors. The *Mean* combination performs better than several of the various combination methods (Timmermann, 2006), but is consistently outperformed by the *Median*, as one would expect for distributions of forecasts that may deviate from normality (Kourentzes et al., 2014a).

Comparing base with combined forecasts, the latter are overall better. For all out-of-sample measurements *Median* has the best overall sME and sMSE performance, while *Regression RW* has the best sMdE and sMdSE performance. With regards to the in-sample errors the *Optimal adaptive RW* performs best, however this is not consistent with its out-of-sample performance, which can be attributed the combination weights having overfit. This behaviour is not observed by the relatively simpler *Optimal* and *Optimal adaptive* counterparts.

5.2. Shape and variability of error distributions

Next we evaluate the shape of the error distributions. For RMSE to be appropriate for the calculation of safety stock it is assumed that the errors are normally distributed. Figure 2 shows the percentage of normally distributed in and out-of-sample errors for base and combined forecasts across time series, as tested using the Shapiro-Wilk test for normality.

Table 1: Forecast bias for in- and out-of-sample sets in sME (sMdE)

Method	In-sample	Out t+1	Out t+3	Out t+5
Naïve	0.922 (0.170)	0.777 (0.172)	0.809 (0.177)	0.844 (0.185)
ETS	0.453 (0.090)	0.459 (0.117)	0.474 (0.120)	0.505 (0.124)
AR	0.472 (0.095)	0.500 (0.124)	0.517 (0.127)	0.544 (0.131)
ARIMA	0.459 (0.092)	2.090 (0.124)	0.492 (0.128)	0.519 (0.130)
Theta	0.454 (**0.088**)	0.455 (**0.117**)	0.469 (**0.117**)	0.496 (0.122)
MAPA	**0.446** (0.090)	**0.446** (0.118)	**0.449** (0.119)	**0.472** (**0.120**)
Mean	0.458 (0.091)	0.486 (0.112)	0.456 (0.115)	0.485 (0.117)
Median	0.447 (0.090)	**0.437** (0.112)	**0.448** (0.114)	**0.474** (0.117)
Optimal	0.456 (0.091)	0.500 (0.112)	0.456 (0.115)	0.485 (0.117)
Optimal adaptive	0.439 (0.087)	0.500 (0.112)	0.453 (0.115)	0.482 (0.118)
Optimal adaptive RW	**0.418** (**0.081**)	0.729 (0.119)	0.463 (0.119)	0.488 (0.123)
Regression	0.500 (0.103)	0.508 (0.120)	0.475 (0.123)	0.503 (0.127)
Regression RW	0.473 (0.091)	0.488 (**0.108**)	0.455 (**0.111**)	0.483 (**0.114**)
Outperformance	0.488 (0.092)	0.586 (0.113)	0.483 (0.119)	0.514 (0.122)
Bates I	0.491 (0.101)	0.508 (0.124)	0.527 (0.128)	0.554 (0.132)
Bates II	0.482 (0.095)	0.470 (0.115)	0.480 (0.119)	0.509 (0.122)
Bates III	0.611 (0.146)	0.504 (0.116)	0.478 (0.118)	0.507 (0.121)
Bates IV	0.605 (0.143)	0.491 (0.116)	0.469 (0.119)	0.499 (0.121)
Bates V	31.075 (0.103)	20.729 (0.135)	39.156 (0.140)	48.487 (0.143)

Table 2: Forecast accuracy for in- and out-of-sample sets in sRMSE (sRMdSE)

Method	In-sample	Out t+1	Out t+3	Out t+5
Naïve	0.960 (0.412)	0.882 (0.415)	0.900 (0.421)	0.919 (0.431)
ETS	0.673 (0.300)	0.677 (0.343)	0.688 (0.347)	0.711 (0.352)
AR	0.687 (0.309)	0.707 (0.353)	0.719 (0.356)	0.737 (0.361)
ARIMA	0.677 (0.304)	1.446 (0.352)	0.701 (0.357)	0.721 (0.360)
Theta	0.674 (**0.297**)	0.674 (**0.342**)	0.685 (**0.342**)	0.705 (0.349)
MAPA	**0.668** (0.300)	**0.668** (0.343)	**0.670** (0.344)	**0.687** (**0.346**)
Mean	0.677 (0.302)	0.697 (0.335)	0.675 (0.339)	0.696 (0.342)
Median	0.668 (0.299)	**0.661** (0.334)	**0.669** (0.338)	**0.689** (0.342)
Optimal	0.675 (0.302)	0.707 (0.335)	0.675 (0.340)	0.696 (0.342)
Optimal adaptive	0.663 (0.295)	0.707 (0.335)	0.673 (0.339)	0.694 (0.343)
Optimal adaptive RW	**0.646** (**0.284**)	0.854 (0.345)	0.680 (0.345)	0.699 (0.351)
Regression	0.707 (0.321)	0.713 (0.347)	0.690 (0.351)	0.709 (0.356)
Regression RW	0.688 (0.302)	0.699 (**0.329**)	0.675 (**0.333**)	0.695 (**0.337**)
Outperformance	0.698 (0.303)	0.766 (0.337)	0.695 (0.345)	0.717 (0.349)
Bates I	0.700 (0.318)	0.713 (0.353)	0.726 (0.358)	0.744 (0.363)
Bates II	0.695 (0.309)	0.686 (0.339)	0.692 (0.346)	0.713 (0.349)
Bates III	0.781 (0.382)	0.710 (0.341)	0.692 (0.344)	0.712 (0.348)
Bates IV	0.778 (0.378)	0.700 (0.341)	0.685 (0.345)	0.706 (0.348)
Bates V	5.575 (0.320)	4.553 (0.368)	6.257 (0.374)	6.963 (0.378)

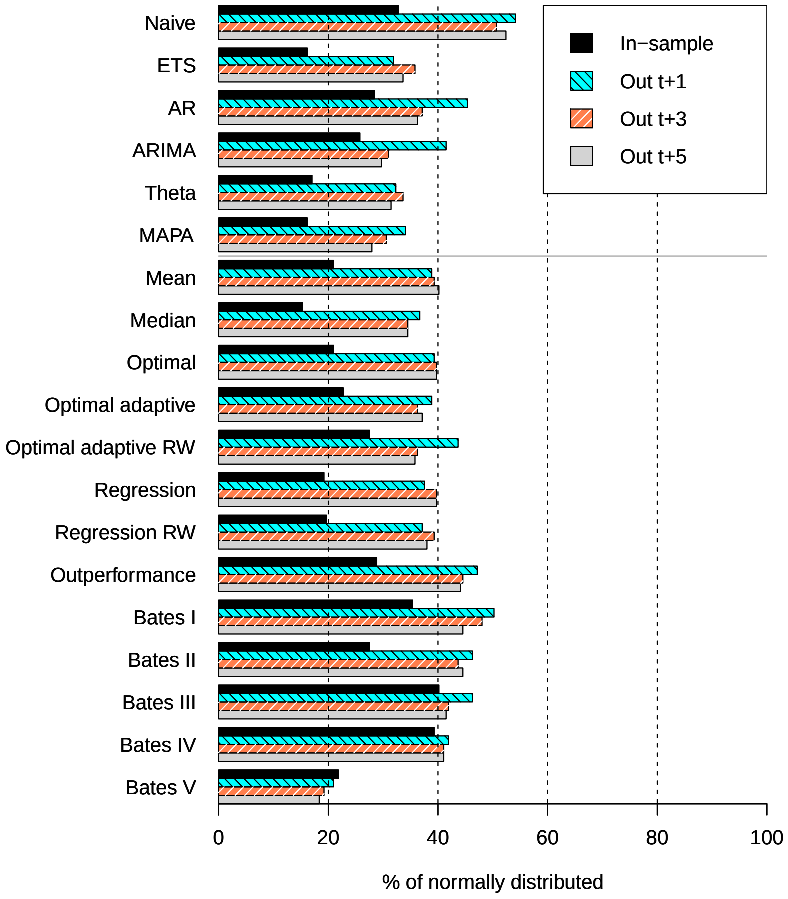

Figure 2: Percentage of normally distributed errors.

The figure provides some interesting insight in the difference between in- and out-of-sample behaviour of the error distributions, with the latter being more normal. Note that the Shapiro-Wilk test tests only deviation of the empirical distribution from the normal and not the forecast performance, which worsens for longer out-of-sample forecast horizons, as indicated by Tables 1 and 2. When comparing base and combination forecast error it is apparent that on average, errors of the combined forecasts exhibit more normal behaviour, with most of the percentages around 40% for the out-of-sample. Crucially, considerably less than 50% of the time series have normally distributed errors, for either base or combined forecasts. This demonstrates that the assumption of normality is violated very frequently. The *Naïve* is an exception to this. Further evidence of the nature of this deviation is provided in Appendix A, where the skewness and kernel density estimations of the error distributions are provided.

Figure 3 provides boxplots of the relative out-of-sample variance for the different forecast horizons over the in-sample variance measured across all time series. The different forecasts are grouped into *Base* and *Combinations* to better highlight the differences between the two groups of forecasts. If the in- and out-of-sample errors would have the same variance then the boxplots should be very close to one, indicated by a horizontal line. This is equivalent to measuring whether the in-sample variance is an appropriate estimation of the out-of-sample variance, as is the standard practice in inventory management. The combined forecasts have smaller relative variance than the base forecasts, supporting the argument that combinations violate less the standard assumptions in comparison to base forecasts.

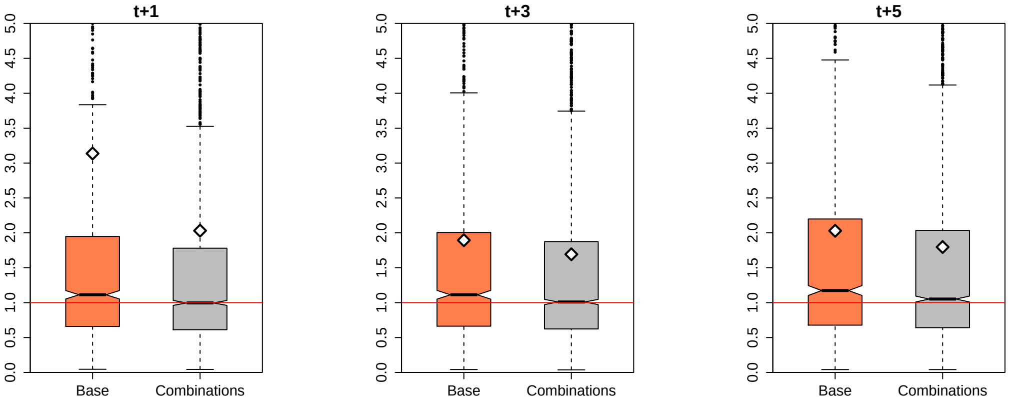

Figure 3: Relative variance of error distribution between in- and out-of-sample for *Base* and *Combination* methods. The mean of the distribution is plotted with a diamond ($\diamond$).

The results so far provide strong evidence that approximating the out-of-sample error variance with the in-sample t+1 RMSE adjusted by $\sqrt{L}$ for longer lead times is inappropriate and overly simplistic, either due to the differences between error distributions shape or variance. Although we do not claim that this is the only way to approximate the out-of-sample variance, its use is widespread and other approaches typically ignore some of the identified distributional differences as well (Chatfield, 2000).

5.3. Error distributions and safety stocks

Up this to point we have seen that combinations on average transform the in- and out-of-sample error distributions closer to normal compared to base forecasts. Furthermore, when the appropriate combination method is used, it is found to have a beneficial impact in terms of forecast bias and accuracy, as well as on the variance of the errors.

We turn our attention to the implication this has on safety stock calculation. We calculate the one-step ahead average safety stock level using the theoretical approach, approximated as $SS = k \cdot sRMSE$, and the empirical in- and out-of-sample distribution for 80% and 95% service levels. Note that the results from the empirical error distributions are bound to be different than the theoretical ones, as the former include any covariance between forecast errors, which are not present in the theoretical formula. The results across series are presented in Table 3. The smallest value in each column is highlighted in boldface.

Table 3: Average in-sample (theoretical and empirical) and out-of-sample t+1 scaled Safety Stock

Method	In-sample Theoretical		In-sample Empirical		Out-of-sample t+1	
	80%	95%	80%	95%	80%	95%
Nave	0.74	1.44	0.95	1.71	0.93	1.65
ETS	0.51	0.99	0.66	1.11	0.71	1.16
AR	0.52	1.02	0.66	1.15	0.73	1.26
ARIMA	0.51	1.00	0.66	1.13	0.72	1.22
Theta	0.51	0.99	0.66	1.11	0.70	1.16
MAPE	0.50	0.98	0.65	1.09	0.70	1.14
Mean	0.51	1.00	0.66	1.13	0.69	1.16
Median	0.51	0.99	0.65	1.10	**0.69**	**1.13**
Optimal	0.51	1.00	0.66	1.13	0.69	1.16
Optimal adaptive	0.50	0.98	0.64	1.10	0.69	1.14
Optimal adaptive RW	**0.49**	**0.95**	**0.63**	**1.07**	0.71	1.17
Regression	0.55	1.07	0.67	1.23	0.70	1.23
Regression RW	0.53	1.03	0.65	1.18	0.68	1.18
Outperformance	0.53	1.03	0.68	1.18	0.71	1.21
Bates I	0.53	1.04	0.69	1.18	0.74	1.27
Bates II	0.53	1.03	0.68	1.17	0.71	1.20
Bates III	0.60	1.17	0.78	1.33	0.70	1.20
Bates IV	0.59	1.16	0.77	1.32	0.70	1.18
Bates V	0.93	1.82	1.11	2.10	0.78	1.70

The average in-sample theoretical estimation of SS is much lower than the average in-sample empirical estimation of SS, and also the out-of-sample t+1 empirical SS. Using in-sample RMSE based on assumption of normality, when distributions are in fact not normal, is on average underestimating safety stock levels compared to both in- and out-of-sample empirical errors. The in-sample empirical SS is much closer to the out-of-sample (t+1) empirical SS, suggesting that using the in-sample empirical distributions of forecast errors might be preferable to the theoretical approach. These findings hold for various service levels between 80% and 95% that were trialled.

To highlight the differences between base and combined forecasts, Figure 4 plots the relative sRMSE of the empirical in- and out-of-sample t+1 errors over the theoretical variance for 80%, 90%, and 95% percentiles of the cumulative distribution, which refer to target service levels. The various base and combination forecasts are grouped. The behaviour of the t+3 and t+5 out-of-sample errors is analogous to the t+1 case and therefore not shown, but obviously any differences are further inflated due to the covariance that is captured in the empirical error distributions and is missing in the theoretical calculation. This covariance has two sources: between errors of forecasts of different steps-ahead and errors forming the cumulative demand over lead time, and therefore the differences increase further for longer lead times.

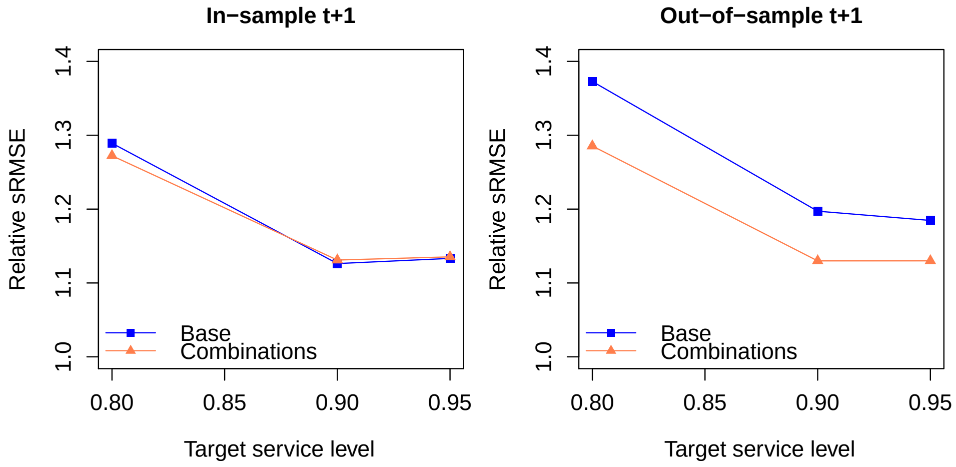

Figure 4: Relative sRMSE of empirical in- and out-of-sample t+1 errors over theoretical.

There is very little difference in the behaviour between the base and the combined forecasts for the case of the in-sample t+1. However, for the out-of-sample case, the combination forecasts result in substantially reduced relative sRMSE, implying lower required safety stocks. These results are based on the empirical distributions and do not assume normality of the errors as the theoretical approach does, which underestimates the observed variance. Furthermore, note that in Figure 4 the behaviour of the combination forecasts between the in- and out-of-sample case is very similar, demonstrating a consistency not seen for the base forecasts. Consequently the safety stocks calculated using the t+1 in-sample empirical error distribution of the combination forecasts are expected to be much more reliable than the ones for the base forecasts.

The variance of combined forecasts is not only lower, but also better behaved in terms of the correlation of the average in- and out-of-sample error variance. Therefore, forecast combinations have a beneficial impact on safety stocks and subsequently on inventory management.

There are important practical implication of these findings. Although the theoretical calculation of the safety stock is convenient, it requires the assumption of normality and ignores any implied covariance. Violating these can result in underestimating the observed error variance. On the other hand, the empirical in-sample t+1 error distribution is a direct alternative for statistical forecasts, as it is a typical output of model fitting. Our analysis shows that while this is a poor approximation of the out-of-sample forecast error behaviour for base forecasts, for the combined forecasts it is accurate.

This is a useful finding for organisations. As long as combination forecasts are used, using the readily available in-sample empirical variance can lead to more reliable safety stock calculations.

6. Conclusions

In the forecasting literature combinations of forecasts are generally considered beneficial and have been found to improve the performance in terms of forecast bias and accuracy. In this paper we focus on evaluating the impact of forecast combinations on the forecast error distribution. This aspect is important for inventory management, and in particular for the calculation of the safety stock.

We find that forecast combinations improve forecast accuracy and bias, in agreement with the literature, but also result in more normally distributed errors. We also provide evidence that the empirical error distribution of the combination forecasts is a good approximation of the out-of-sample one, while this not as evident for the base forecasts. In any case, the empirical distribution provides a better approximation of the out-of-sample behaviour in comparison to the commonly used theoretical approximation, as the latter is based on the strong assumption of normality.

In so doing we identify a practical way for organisations to achieve better approximation of the demand uncertainty when using combined forecasts. Using the empirical distribution of the forecast errors we are able to overcome limitations of the standard theoretical formula for stock calculations which fails to account for deviations from normality and any covariance between forecast errors of the cumulative demand over lead time. The impact of this miscalculation for individual and combination forecasts is explored and the latter is found to be more robust.

Approximating the out-of-sample empirical distribution requires appropriate handling of the data, i.e. the use of a validation sample, and adequate sample size. This may make its use complicated for practice. We proceed to identify an effective approximation of the out-of-sample uncertainty using the empirical in-sample errors that holds for combination forecasts, but not for individual forecasts. This greatly simplifies any calculations, as no special separation of the sample is needed. In fact, this figure is often a result of the forecasting model estimation and can be readily available in organisations.

Translating these results in terms of safety stock we find that combinations behave more consistently between in- and out-of-sample errors and require less safety stock to cover the observed forecast error variance. Thus, the overall conclusion is that when the additional dimensions of this research are considered, forecast combinations improve upon base forecasts, with beneficial implications for inventory management.

Although the focus of the paper was not to compare the different combination methods, we evaluated a wide selection of methods. Simple ones, such as the *Median*, performed at least as good, if not better, than more complex methods. A useful finding is that the widely used *Mean* did not perform well in the presence of irregular data. This is very relevant to inventory management, where special events and promotions are common. This is helpful for practice as it demonstrates that simple and easy to implement combination methods can bring the desired benefits.

References

Aksu, C., Gunter, S. I., 1992. An empirical analysis of the accuracy of sa, ols, erls and nrls combination forecasts. International Journal of Forecasting

8 (1), 27–43.

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.

Assimakopoulos, V., Nikolopoulos, K., 2000. The theta model: a decomposition approach to forecasting. International Journal of Forecasting 16 (4), 521–530.

Axsäter, S., 2006. Inventory Control. Springer, Berlin.

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

Bordley, R. F., 1982. The combination of forecasts: A bayesian approach. Journal of the Operational Research Society 33 (2), 171–174.

Bunn, D. W., 1975. Bayesian approach to linear combination of forecasts. Operational Research Quarterly 26 (2), 325–329.

Chan, C. K., Kingsman, B. G., Wong, H., 1999. The value of combining forecasts in inventory management - a case study in banking. European Journal of Operational Research 117 (2), 199–210.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society. Series A (Statistics in Society) 158 (3), 419–466.

Chatfield, C., 1996. Model uncertainty and forecast accuracy. Journal of Forecasting 15 (7), 495–508.

Chatfield, C., 2000. Time-series forecasting. CRC Press.

Chen, X., Sim, M., Simchi-Levi, D., Sun, P., 2007. Risk aversion in inventory management. Operations Research 55 (5), 828–842.

Clemen, R. T., Winkler, R. L., 1986. Combining economic forecasts. Journal of Business & Economic Statistics 4 (1), 39–46.

Clements, M. P., Hendry, D. F., 2007. An overview of economic forecasting. In: Michael P. Clements, D. F. H. (Ed.), A Companion to Economic Forecasting. Blackwell Publishing Ltd., pp. 1–18.

de Menezes, L. M., Bunn, D. W., 1993. Diagnostic tracking and model specification in combined forecasts of U.K. inflation. Journal of Forecasting 12 (7), 559–572.

de Menezes, L. M., Bunn, D. W., 1998. The persistence of specification problems in the distribution of combined forecast errors. International Journal of Forecasting 14 (3), 415–426.

de Menezes, L. M., Bunn, D. W., Taylor, J. W., 2000. Review of guidelines for the use of combined forecasts. European Journal of Operational Research 120 (1), 190–204.

Dekker, M., van Donselaar, K., Ouwehand, P., 2004. How to use aggregation and combined forecasting to improve seasonal demand forecasts. International Journal of Production Economics 90 (2), 151–167.

Diebold, F. X., Pauly, P., 1987. Structural-change and the combination of forecasts. Journal of Forecasting 6 (1), 21–40.

Diebold, F. X., Pauly, P., 1990. The use of prior information in forecast combination. International Journal of Forecasting 6 (4), 503–508.

Donaldson, R. G., Kamstra, M., 1996. Forecast combining with neural networks. Journal of Forecasting 15 (1), 49–61.

Elliott, G., Timmermann, A., 2004. Optimal forecast combinations under general loss functions and forecast error distributions. Journal of Econometrics 122 (1), 47–79.

Elliott, G., Timmermann, A., 2005. Optimal forecast combination under regime switching. International Economic Review 46 (4), 1081–1102.

Fildes, R., 1989. Evaluation of aggregate and individual forecast method selection rules. Management Science 35 (9), 1056–1065.

Fildes, R., Howell, S., 1979. On selecting a forecasting model. TIMS Studies in Management Science 12, 297–312.

Gerchak, Y., Mossman, D., 1992. On the effect of demand randomness on inventories and costs. Operations Research 40 (4), 804–807.

Granger, C. W. J., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting 3 (2), 197–204.

Guidolin, M., Timmermann, A., 2009. Forecasts of US short-term interest rates: A flexible forecast combination approach. Journal of Econometrics 150 (2), 297–311.

Hyndman, R., 2014. FORECAST package for R v5.5.
URL `http://cran.r-project.org/web/packages/forecast/index.html`

Hyndman, R. J., Billah, B., 2003. Unmasking the theta method. International Journal of Forecasting 19 (2), 287–290.

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

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

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

Jose, V. R. R., Winkler, R. L., 2008. Simple robust averages of forecasts: Some empirical results. International Journal of Forecasting 24 (1), 163–169.

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

Kourentzes, N., Barrow, D. K., Crone, S. F., 2014a. Neural network ensemble operators for time series forecasting. Expert Systems with Applications 41 (9), 4235–4244.

Kourentzes, N., Petropoulos, F., 2014. MAPA package for R v1.7.
URL `http://cran.r-project.org/web/packages/MAPA/index.html`

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

Kourentzes, N., Svetunkov, I., 2014. TStools package for R v1.0.
URL `https://github.com/trnnick/TStools`

Lemke, C., Gabrys, B., 2010. Meta-learning for time series forecasting and forecast combination. Neurocomputing 73 (1012), 2006–2016.

Lütkepohl, H., 2011. Forecasting nonlinear aggregates and aggregates with time-varying weights. Jahrbücher für Nationalökonomie und Statistik, 107–133.

Macdonald, R., Marsh, I. W., 1994. Combining exchange-rate forecasts - what is the optimal consensus measure. Journal of Forecasting 13 (3), 313–332.

Makridakis, S., 1986. The art and science of forecasting an assessment and future directions. International Journal of Forecasting 2 (1), 15–39.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R., 1982. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of Forecasting 1 (2), 111–153.

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

Makridakis, S., Hibon, M., Lusk, E., Belhadjali, M., 1987. Confidence intervals: An empirical investigation of the series in the M-competition. International Journal of Forecasting 3 (3), 489–508.

Makridakis, S., Winkler, R. L., 1989. Sampling distributions of post-sample forecasting errors. Applied Statistics, 331–342.

Min, C. K., Zellner, A., 1993. Bayesian and non-bayesian methods for combining models and forecasts with applications to forecasting international growth-rates. Journal of Econometrics 56 (1-2), 89–118.

Newbold, P., Granger, C. W. J., 1974. Experience with forecasting univariate time series and combination of forecasts. Journal of the Royal Statistical Society Series a-Statistics in Society 137, 131–165.

Pant, P. N., Starbuck, W. H., 1990. Innocents in the forest: Forecasting and research methods. Journal of Management 16 (2), 433–460.

Petropoulos, F., Kourentzes, N., 2014. Forecast combinations for intermittent demand. Journal of Operational Research Society.

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/`

Silver, E. A., Pyke, D. F., Peterson, R., 1998. Inventory management and production planning and scheduling. Vol. 3. Wiley New York.

Stock, J. H., Watson, M. W., 2004. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23 (6), 405–430.

Tian, J., Anderson, H. M., 2014. Forecast combinations under structural break uncertainty. International Journal of Forecasting 30 (1), 161–175.

Timmermann, A., 2006. Forecast combinations. In: Elliott, G., Granger, C., Timmermann, A. (Eds.), Handbook Of Economic Forecasting. Elsevier, Amsterdam.

Appendix A. Deviation of error distributions from normality

Here we provide further evidence on the nature of the deviations from normality that the forecast error distributions exhibit. Figure A.5 plots the distribution of the coefficient of skewness of the in- and out-of-sample scaled errors for all methods across the various time series. The out-of-sample distribution for all horizons are grouped as they have only smalls differences.

The forecast errors of the base forecasts are skewed, as are the errors of the combined forecasts, with the only exception being the *Naïve* method. However the boxplots indicate that some combination methods (for e.g. the *Outperformance* method and *Bates IV* method) are effective at reducing both the in- and out-of-sample skewness, thus reducing the non-normality of the error distributions.

To further understand the shape of the distributions Figures A.6 and A.7 provide the t+1 in and out-of-sample scaled error average empirical distributions, using kernel density estimation. Note that these are smoothed since these are averaged across all time series. Forecast horizons t+3 and t+5 are not provided, as they are very similar to the out-of-sample t+1. These plots

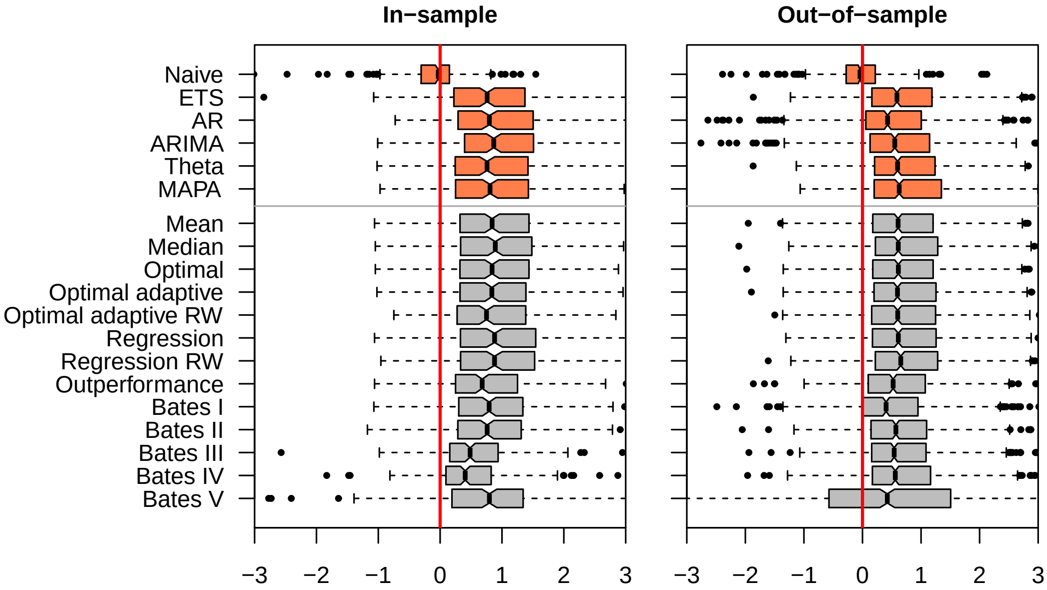

Figure A.5: Boxplots of skewness of error distribution per method. Normal distribution corresponds to 0 skewness.

are illustrative of the shape of the distribution and are not intended to provide a detailed view of the errors of each forecast method. Error distributions exhibit heavy tails and multimodality. The difference in the shape between the in- and out-of-sample errors is particularly clear, explaining the results in Figure 2. Although combinations lead to some improvement in normality, the distributions remain overall non-normal.

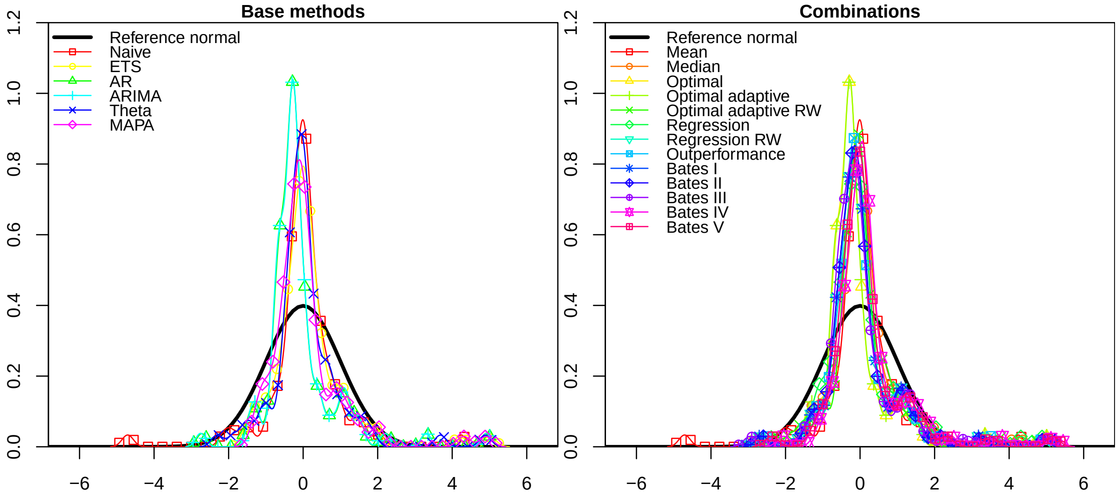

Figure A.6: Densities for in-sample errors.

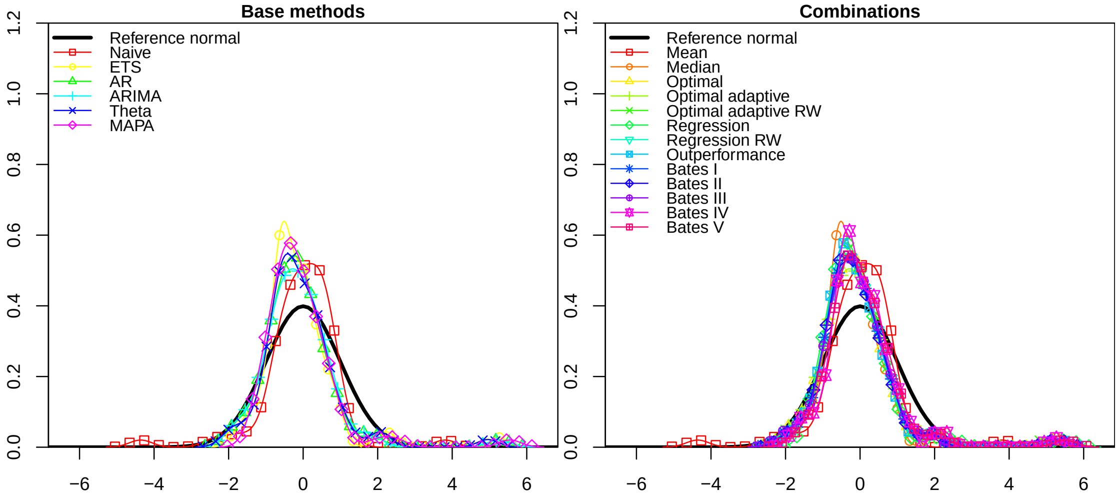

Figure A.7: Densities for out-of-sample t+1 errors.