Artif Intell Rev
DOI 10.1007/s10462-012-9361-z

Performance evaluation of weights selection schemes for linear combination of multiple forecasts

Ratnadip Adhikari · R. K. Agrawal

Abstract Time series modeling and forecasting are essential in many domains of science and engineering. Extensive works in literature suggest that combining outputs of different forecasting methods substantially increases the overall accuracies as well as reduces the risk of model selection. The most popular method of forecasts combination is the weighted averaging of the constituent forecasts. The effectiveness of this method solely depends on appropriate selection of the combining weights. In this paper, we comprehensively evaluate a wide variety of benchmark weights selection techniques for linear combination of multiple forecasts in terms of their prediction accuracies. Nine real-world time series from different domains and five individual forecasting methods are used in our empirical work. A robust scheme is also suggested for fairly ranking the combination methods on the basis of their forecasting performances. Our study precisely demonstrates the relative strengths and weaknesses of various benchmark linear combination techniques which evidently can be of much practical importance.

Keywords Time series · Combining forecasts · Linear combination · Weighted average

1 Introduction

Improvement of time series forecasting accuracy has continuously attracted attentions of researchers during the last two decades and as a result various important forecasting methods have been developed in literature. However, the accuracy of a method is very much problem specific and no general conclusion can be made in this regard (Clemen 1989; Gooijer and Hyndman 2006). Also, one cannot exactly identify the best forecasting model for a time series

R. Adhikari (✉) · R. K. Agrawal
School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi 110067, India
e-mail: adhikari.ratan@gmail.com

R. K. Agrawal
e-mail: rkajnu@gmail.com

Published online: 19 October 2012

but conceptually different models may play a complementary role in the approximation of the actual data generating process (Terui and Van Dijk 2002). In view of these facts, it is wise to use a number of different forecasting methods and then aggregate their outputs through a suitable combination technique (Winkler and Makridakis 1983a; Clemen 1989; Terui and Van Dijk 2002; Gooijer and Hyndman 2006). The practice of combining forecasts benefits from the strengths of the constituent models and effectively reduces the errors arising from faulty assumptions, bias or mistakes in the data (Armstrong 2001). An extensive body of literature has shown that combined forecasts can remarkably improve the overall accuracies and often significantly outperform the individual methods.

The most intuitive and popular way of forecasts aggregation is to linearly combine the constituent forecasts. Various important methods have been proposed in literature for selecting the combining weights. The easiest among them is the simple average in which all forecasts are weighted equally. Research evidences show that the naïve simple average provides reasonably improved accuracies in many situations (Gooijer and Hyndman 2006; Jose and Winkler 2008; Andrawis et al. 2011). But, the sensitiveness of this method to the extreme errors often deteriorates the overall forecasting precisions. Other variations and alternatives of the simple average have also been investigated in literature. Some important among them include the trimmed mean, the median (i.e. the ultimate trimming), the Winsorized mean, etc. (Armstrong 2001; Stock and Watson 2004; Jose and Winkler 2008). Although these statistical methods are quite easy to implement and interpret but they overlook the past information regarding the precisions of the participated forecasts as well as the relative dependence among the forecasts. Assigning the weights to the individual models on the basis of their previous forecasting performances is both a rational as well as an advantageous approach (Clemen 1989; Aksu and Gunter 1992; Armstrong 2001; Zou and Yang 2004; Gooijer and Hyndman 2006). As such, numerous linear combination schemes have been developed in literature in which the combining weights are judgmentally selected according to the relative precisions of the contributing forecasts. It is important that while combining multiple forecasts, a lot of consideration is to be given for maintaining the trade-off between the accuracy enhancements and associated computational cost. An ensemble mechanism, requiring a lot of computational time is not practically suitable for large scale datasets. Furthermore, it has been observed that simple combination methods often provide notably better forecasting results than more complicated and sophisticated techniques (Clemen 1989; Jose and Winkler 2008).

The last two decades have witnessed a surge of research works on time series modeling and forecasting. During this period, many important algorithms have been developed for individual forecasting as well as for combining forecasts. At present, the selection of the most promising combination scheme is itself quite challenging (Lemke and Gabrys 2010). With the increase in the number of available forecasts combination techniques, the studies related to their comparative assessment are also growing continuously. Some remarkable early contributions were made by Newbold and Granger (1974), Makridakis et al. (1982), Winkler and Makridakis (1983a). In their influential work, Menezes et al. (2000) reviewed the performances of various forecasts combination methods and provided important practical guidelines for combining forecasts. Makridakis and Hibon (2000) conducted the renowned *M-3 competition* which presents an extensive analysis of a wide variety of time series forecasting methods and combination techniques. Recently, Lemke and Gabrys (2010) investigated different meta-learning approaches to identify the best individual model as well as the most adequate combination scheme for a given forecasting situation. These works are inevitably helpful in assessing the relative efficiencies of different combination methods. However, most of these studies are performed on time series which show regular patterns (e.g. stationary or seasonal time series). Also, most of them analyze the general aspects of model combinations and as such there is a lack of research evidences which minutely explores performances of different weighting schemes for linear combinations of multiple forecasts.

In this paper, we present a comprehensive comparison of ten different weighted linear ensemble techniques for nine practical time series. Our aim here is not to perform an exhaustive analysis with all possible combination methods but to study the effects of a number of benchmark weight assignment mechanisms for linear combination of multiple forecasts. It is evident that the error reduction rate of an ensemble increases with the increase in the number of individual methods. But, it was empirically observed that a degree of saturation in the accuracy is reached after about four or five methods and the variability of accuracy among different combinations diminished as the number of component models are further increased (Winkler and Makridakis 1983a,b; Armstrong 2001). In view of these findings, five individual forecasting methods are used to create all combination models in this paper. We also provide a rational as well as robust technique to rank the different combination methods on the basis of their forecasting performances.

The rest of the paper is organized as follows. In Sect. 2, we describe the linear ensemble framework and various weight assignment schemes. Section 3 presents a relative assessment of the different combination techniques which are used in this paper. Section 4 briefly describes the five individual forecasting methods. The empirical results are reported in Sect. 5 and finally the paper is concluded in Sect. 6.

2 Linear combination of multiple forecasts

A combination approach attempts at improving the overall forecasting accuracy while at the same time decreasing the model selection risk (Winkler and Makridakis 1983b; Lemke and Gabrys 2010). By combining multiple methods we try to reasonably increase the forecasting precision but at the expense of increased computational complexity. Thus, one should possess rigorous insights into the properties of all component models and should carefully utilize the available knowledge in order to develop an effective forecasts combination mechanism (Kuncheva 2004).

In a linear combination method, the combined forecast is calculated through a linear function of the contributing individual forecasts. Let, $\mathbf{Y} = [y_1, y_2, \ldots, y_N]^{\mathrm{T}} \in \mathbb{R}^N$ be the actual time series dataset whose forecasts are obtained through n different models. Then each forecaster can be expressed as a function $f_i : \mathbb{R}^N \to \mathbb{R}^N$ $(i = 1, 2, \ldots, n)$, where it should be noted that f_i transforms the actual dataset $\mathbf{Y}$ to its corresponding forecast and it may not be an explicit function. A linear combination of these n forecasts of $\mathbf{Y}$can be expressed as:

$$\hat{\mathbf{Y}} = F\left(f_1(\mathbf{Y}), f_2(\mathbf{Y}), \ldots, f_n(\mathbf{Y}), \mathbf{W}\right) \tag{1}$$

where $F : \mathbb{R}^N \to \mathbb{R}^N$ is the linear combination function and $\mathbf{W} = [w_1, w_2, \ldots, w_n]^{\mathrm{T}} \in \mathbb{R}^n$ is the weight vector. The function F is linear in all f_i $(i = 1, 2, \ldots, n)$ and $\mathbf{W}$.

Taking $\hat{\mathbf{Y}}^{(i)} = f_i(\mathbf{Y}) = \left[\hat{y}_1^{(i)}, \hat{y}_2^{(i)}, \ldots, \hat{y}_N^{(i)}\right]^{\mathrm{T}}$ and $\hat{\mathbf{Y}} = \left[\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\right]^{\mathrm{T}}$, the linear combination scheme (1) can be conveniently expressed as:

$$\left.\begin{array}{l} \hat{y}_k = w_1\hat{y}_k^{(1)} + w_2\hat{y}_k^{(2)} + \cdots + w_n\hat{y}_k^{(n)} = \sum_{i=1}^{n} w_i\hat{y}_k^{(i)} \\ \forall k = 1, 2, \ldots, N. \end{array}\right\} \tag{2}$$

The weights are often restricted to be unbiased (i.e. they add up to unity) and nonnegative. A schematic depiction of the linear ensemble mechanism is presented in Fig. 1.

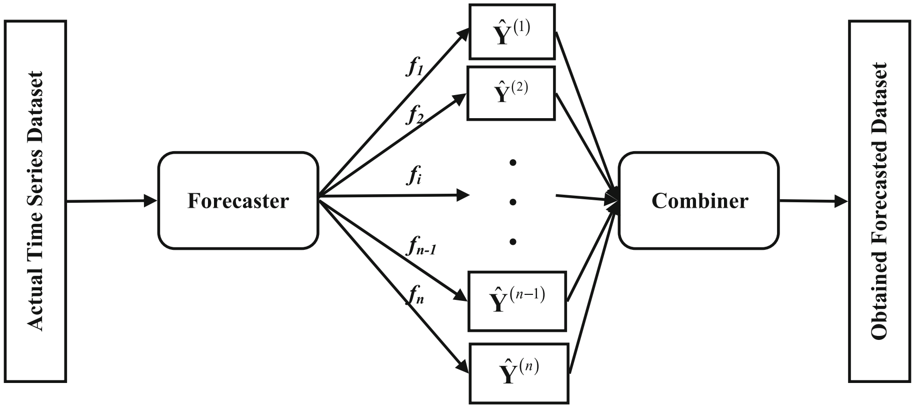

Fig. 1 A linear combination of multiple forecasts

Starting with the initial classic works of Bates and Granger (1969), Newbold and Granger (1974), Winkler and Makridakis (1983a) various important contributions have been made in the area of forecasts combination. Here we discuss about a number of benchmark weights assignment techniques for linear combination of multiple forecasts.

2.1 The basic statistical methods

The easiest method of combining is the simple average in which all component forecasts are weighted equally. Numerous research evidences have shown that the naïve simple average often remarkably improves overall forecasting accuracies (Winkler and Makridakis 1983a; Jose and Winkler 2008; Lemke and Gabrys 2010). However, it is well-known that simple averages themselves are very sensitive to extreme values and so other forms of statistical averaging are also suggested in literature. Two popular alternatives are the trimmed mean and the median (Jose and Winkler 2008; Lemke and Gabrys 2010). Assuming that the sequence $\left\{\hat{y}_1^{(i)}, \hat{y}_2^{(i)}, \ldots, \hat{y}_N^{(i)}\right\}$ is sorted in ascending order, the trimmed mean is mathematically given by:

$$t_k(\alpha) = \begin{cases} \frac{1}{n-2\alpha} \sum_{i=\alpha+1}^{n-\alpha} \hat{y}_k^{(i)} \\ \forall 0 \leq \alpha < n/2; \quad k = 1, 2, \ldots N \end{cases} \tag{3}$$

A trimmed mean averages the individual forecasts by excluding the worst performing $\beta\%$ ($\beta = 100(2\alpha/n)$) of the models. The simple average and median themselves are the two extremes of trimming corresponding to $\beta = 0$ and $\beta = \lfloor n/2 \rfloor$, respectively. A 10–40 % trimming is generally recommended (Jose and Winkler 2008; Lemke and Gabrys 2010).

2.2 The error-based methods

In such a method, the time series to be forecasted is divided into two complementary subsets, viz. the training and the validation sets. The component models are fitted on the training set and their obtained forecast errors on the validation set are recorded. The combining weight to each individual forecast is then taken to be inversely proportional to the forecast error of the corresponding model (Armstrong 2001), i.e.

$$w_i = e_i^{-1} \Big/ \sum_{i=1}^{n} e_i^{-1} \qquad (4)$$
$$\forall i = 1, 2, \ldots, n.$$

Here e_i denotes the error obtained by the ith forecasting model. The scheme (4) assures that the model with more error is assigned the less weight to it and vice versa.

Among the various measures for evaluating the forecast errors, three performance measures are used in this paper. These are: the Mean Absolute Error (MAE), the Mean Squared Error (MSE) and the Symmetric Mean Absolute Percentage Error (SMAPE), which are defined below:

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|y_t - \hat{y}_t\right|, \quad \text{MSE} = \frac{1}{N}\sum_{t=1}^{N}\left(y_t - \hat{y}_t\right)^2,$$
$$\text{SMAPE} = \frac{1}{N}\sum_{t=1}^{N}\frac{\left|y_t - \hat{y}_t\right|}{\left(y_t + \hat{y}_t\right)/2} \times 100.$$

Here, y_t and $\hat{y}_t$ are respectively the actual and forecasted values and N is the number of forecasted observations.

In this paper, we refer the three error-based methods as EB-I, EB-II and EB-III which respectively corresponds to the estimation of combining weights through validation MAE, MSE and SMAPE values.

2.3 The least square regression (LSR) method

This method determines the weights of a linear combination by minimizing the Sum of Squared Error (SSE), calculated from the target and forecasted observations. The usual formulation of the linear combination scheme (2) can be written more conveniently in matrix form as:

$$\hat{\mathbf{Y}} = \mathbf{UW} \qquad (5)$$

where, $\mathbf{U} = \begin{bmatrix} \hat{y}_1^{(1)} & \hat{y}_1^{(2)} & \ldots & \hat{y}_1^{(n)} \\ \hat{y}_2^{(1)} & \hat{y}_2^{(2)} & \ldots & \hat{y}_2^{(n)} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_N^{(1)} & \hat{y}_N^{(2)} & \ldots & \hat{y}_N^{(n)} \end{bmatrix}, \quad \mathbf{w} = [w_1, w_2, \ldots, w_n]^{\mathrm{T}}.$

Then the forecast SSE is given by:

$$\begin{aligned} \text{SSE} &= \sum_{k=1}^{N}\left(y_k - \hat{y}_k\right)^2 \\ &= (\mathbf{Y} - \mathbf{UW})^{\mathrm{T}}(\mathbf{Y} - \mathbf{UW}) \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - 2\mathbf{W}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{Y} + \mathbf{W}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{UW} \end{aligned} \qquad (6)$$

Minimizing the SSE with respect to $\mathbf{W}$, the desired weight vector is obtained as:

$$\mathbf{W} = \mathbf{U}^{+}\mathbf{Y} \qquad (7)$$

Here, $\mathbf{U}^{+} = \left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1}\mathbf{U}^{\mathrm{T}}$ is the pseudo-inverse of $\mathbf{U}$ which is defined if $\mathbf{U}^{\mathrm{T}}\mathbf{U}$ is non-singular (Frietas and Rodrigues 2006).

It should be noted that the forecast SSE is unknown in advance and so in practical applications it is estimated from the available data through various methods. A straightforward approach used in this paper is to determine a sample SSE through in-sample training and validation sets and then use it in (6) for minimization (Lemke and Gabrys 2010).

2.4 The differential weighting methods

Similar to the LSR method, the weights of a linear combination of forecasts can be found by minimizing the variance of the combined forecast error (Newbold and Granger 1974; Winkler and Makridakis 1983a; Chan et al. 2004). However, this approach requires knowledge of the covariance matrix of forecast errors which is unknown in practice. As a remedial measure, five differential weighting schemes were suggested by Newbold and Granger (1974). Later, Winkler and Makridakis (1983a) empirically compared these five classic methods and found that two of them notably outperformed the others. These two methods, as defined below are used in the present paper.

Differential weighting scheme (DWS)-I:

$$w_i = \left(\sum_{s=t-v}^{t-1} \left(e_s^{(i)} \right)^2 \right)^{-1} \Big/ \sum_{j=1}^{n} \left(\sum_{s=t-v}^{t-1} \left(e_s^{(j)} \right)^2 \right)^{-1} \qquad (8)$$
$$\forall i = 1, 2, \ldots, n.$$

Differential weighting scheme (DWS)-II:

$$w_{i,t} = \beta w_{i,t-1} + (1-\beta) \left[\left(\sum_{s=t-v}^{t-1} \left(e_s^{(i)} \right)^2 \right)^{-1} \Big/ \sum_{j=1}^{n} \left(\sum_{s=t-v}^{t-1} \left(e_s^{(j)} \right)^2 \right)^{-1} \right] \qquad (9)$$
$$\forall i = 1, 2, \ldots, n.$$

Here, n is the number of forecasting methods, t is the time period of forecast, $w_{i,t-1}$ is the weight assigned to the ith method based on the data preceding period $t-1$, v and β $(0 < \beta < 1)$ are two constant parameters and $e_t^{(i)}$ is the percentage forecast error at time t which is defined as:

$$e_t^{(i)} = \frac{y_t - \hat{y}_t^{(i)}}{y_t} \qquad (10)$$
$$\forall i = 1, 2, \ldots, n.$$

In both these schemes, the combining weights are based on the reciprocal of the forecast SSE values. The weights in DWS-I are inversely proportional to the SSE values, whereas the weights in DWS-II are obtained through exponential smoothing of those in DWS-I (Winkler and Makridakis 1983a).

The choice of the parameters v and β are crucial for effective performances of the differential weighting methods (8) and (9), respectively. The smaller values of v restrict the estimation to the most recent observations, whereas smaller values of β assure that more weight is given to recent observations. Following the works and recommendations of Winkler and Makridakis (1983a), we use $v = 12$ and $\beta = 0.7$ in this paper.

2.5 The outperformance method

The outperformance method, proposed by Bunn (1975), adopts a Bayesian framework of subjective probabilities to assign the weights in a linear combination of forecasts. The weight of a component forecasting model is determined on the basis of the number of times it outperformed others in the past (Bunn 1975; Lemke and Gabrys 2010). For simplicity, we consider the problem of linearly combining two forecasting models F_1 and F_2 as follows:

$$\begin{aligned} \hat{y}_k &= w\hat{y}_k^{(1)} + (1-w)\,\hat{y}_k^{(2)} \\ \forall k &= 1, 2, \ldots, N. \end{aligned} \tag{11}$$

Suppose that the performances of the two models are measured in terms of some absolute error and are recorded for M times. Then it is logical to assume that the model with a smaller absolute error outperformed the other. If the number of times F_1 outperformed F_2 is a fraction k of M, then it can be considered that k follows a beta distribution $\mathrm{B}\,(k\,|a_1, a_2)$ $(0 \leq k \leq 1; a_1, a_2 > 0)$ (Bunn 1975). Now for each forecast realization i, we define the Bernoulli variable:

$$\left.\begin{aligned} \delta_i &= 1, \quad \text{if } F_1 \text{has outperformed } F_2 \\ &= 0, \quad \text{otherwise} \end{aligned}\right\} \tag{12}$$

Then after j forecast realizations, the distribution for k will be $\mathrm{B}\left(k\,\middle|\,a_1 + \sum_{i=1}^{j}\delta_i, a_2 + j - \sum_{i=1}^{j}\delta_i\right)$ and accordingly the optimal combination weight for F_1 is given by:

$$w = \bar{k} = \left(a_1 + \sum_{i=1}^{j}\delta_i\right)/(a_1 + a_2 + j) \tag{13}$$

The extension of the above framework for n models can be obtained by assuming that the outperformance fractions k_i $(i = 1, 2, \ldots, n)$ of the component models follow the n-parameters Dirichlet distribution, which is the multivariate analogue of the beta distribution (Bunn 1975). In our implementation of the outperformance method, we use the Dirichlet distribution with all parameters as unity. This will provide equal weights initially to all the component models.

3 Relative assessment of forecasts combination methods

In the previous section, we have described five major classes of linear forecasts combination techniques. The selection of a suitable combination scheme for a particular forecasting problem is itself a nontrivial task and should be done through careful analysis. Here, we present a comparison of the strengths and weaknesses of the discussed linear combination methods.

- **The statistical methods:** The statistical combination methods are easy to implement as well as computationally economical as they do not require any estimation of weights or other related parameters (Gooijer and Hyndman 2006; Jose and Winkler 2008). However, the major drawback with these methods is that they ignore the past forecasting performances of the component models and also the relative dependence among the forecasts. As a result, the combined forecasts obtained through these methods are inefficient when the forecast errors are significantly correlated (Gooijer and Hyndman 2006).

- **The error-based methods:** These methods apparently overcome the limitations of the statistical combination schemes by assigning weights to the different forecasting models on the basis of their past track record (Winkler and Makridakis 1983a). The combined forecast precisions through these methods precisely depend on the absolute error measure used for evaluating the forecast errors of the individual models.
- **The LSR method:** Unlike other techniques, this method provides a robust mathematical approach for estimating the combining weights from past observations as well as past forecasts of the contributing models (Gooijer and Hyndman 2006; Lemke and Gabrys 2010). The selection of proper training and validation sets are important or else there may be instability in the estimated weights.
- **The Differential Weighting Methods:** Like the error-based and LSR schemes, the weight estimations through differential weighting methods also solely depend on the past forecast errors of the component models. However, these methods determine the combining weights adaptively from the training data rather than optimizing them. This is an efficient approach, especially for nonstationary and seasonal time series (Winkler and Makridakis 1983a).
- **The Outperformance Method:** This method follows an alternative epistemic approach of Bayesian probabilities in order to assign subjective weights to the component models and it is found to be an effective forecasts combination technique (Bunn 1975; Lemke and Gabrys 2010). The appropriate number of forecasts realization and selection of parameters of the associated probability distribution have utmost importance for the success of this method.

An ideal ensemble scheme is expected to provide reasonably increased accuracy at the expense of moderate computational works. A linear combination of forecasts should perform at least as good as the simple average otherwise there is no point in using it. Thus, one way to relatively assess the precision of any linear combination scheme is to compare its obtained forecasting accuracy to that of the simple average method.

4 Individual forecasting methods

Numerous empirical studies suggest that reasonable enhancement of the combined forecasting accuracy is achieved if individual forecasts are made through independent models (Armstrong 2001) Further, it was also observed that best accuracies are often obtained by combining about four to five component models (Winkler and Makridakis 1983a,b; Armstrong 2001). According to these guidelines, we use the following five component methods for creating all ensembles in this paper:

- *The autoregressive integrated moving average (ARIMA) model.*
- *The support vector machine (SVM) model.*
- *The iterated artificial neural network (ANN) model.*
- *The iterated Elman ANN (EANN) model.*
- *The direct EANN model.*

Next, we briefly describe the aforementioned five forecasting methods.

4.1 The autoregressive integrated moving average (ARIMA) methods

These are the most widely used statistical techniques for time series forecasting and are developed by Box and Jenkins (1970). These models assume that successive observations of a time series are linearly generated from its past values and a random noise process. An ARIMA(p, d, q) model is mathematically given by:

$$\phi(L)(1-L)^d y_t = \theta(L)\varepsilon_t \tag{14}$$

where

$$\phi(L) = 1 - \sum_{i=1}^{p}\phi_i L^i, \quad \theta(L) = 1 + \sum_{j=1}^{q}\theta_j L^j \quad \text{and} \quad Ly_t = y_{t-1}.$$

The parameters p, d, q, respectively represent the number of *autoregressive*, *degree of differencing* and *moving average* terms with y_t being the actual time series and ε_t being a white noise process. A single differencing is often sufficient for practical applications. The appropriate ARIMA model parameters are usually determined through the well-known Box–Jenkins iterative model-building procedure (Box and Jenkins 1970; Zhang 2003). The ARIMA(0, 1, 0), i.e. $y_t - y_{t-1} = \varepsilon_t$ is the popular *Random Walk (RW)* model which is frequently used in forecasting nonstationary and chaotic time series. A generalization of the basic ARIMA model, viz. the *Seasonal ARIMA (SARIMA)* was also suggested by Box and Jenkins for modeling and forecasting seasonal time series. This model performs sequences of ordinary as well as seasonal differencing of the series in order to capture the seasonal and nonseasonal relationships among the successive observations.

4.2 The support vector machines (SVMs)

SVMs, developed by Vapnik (1995) are a class of robust statistical methods based on the Structural Risk Minimization (SRM) principle. The objective of SVM is to find a linear decision rule with good generalization ability. Time series forecasting is a branch of Support Vector Regression (SVR) in which an optimal separating hyperplane is constructed to correctly classify real-valued outputs (Vapnik 1995; Suykens and Vandewalle 1999).

For the training dataset of N points $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ with $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, SVM attempts to approximate the unknown data generation function in a linear form. Using Vapnik's ε-insensitive loss function, the SVM regression is converted to a Quadratic Programming Problem (QPP) to minimize the empirical risk. Solving the QPP, the optimal decision hyperplane is given by:

$$y(\mathbf{x}) = \sum_{i=1}^{N_s}\left(\alpha_i - \alpha_i^*\right) K(\mathbf{x}, \mathbf{x}_i) + b_{\text{opt}} \tag{15}$$

where, N_s is the number of support vectors, α_i and α_i^* ($i = 1, 2, \ldots, N_s$) are the Lagrange multipliers, b_{opt} is the optimal bias and $K(\mathbf{x}, \mathbf{x}_i)$ is the kernel function.

Various SVM kernels exist in literature and a popular one is the Radial Basis Function (RBF) kernel, defined as $K(\mathbf{x}, \mathbf{y}) = \exp(-||\mathbf{x} - \mathbf{y}||^2/2\sigma^2)$ where σ is a tuning parameter. The RBF kernel is used in this paper and SVM parameters are estimated through grid search techniques.

4.3 The artificial neural networks (ANNs)

ANNs belong to the class of computational intelligence models and have gained immense popularity in time series forecasting domain due to their unique nonlinear, nonparametric,

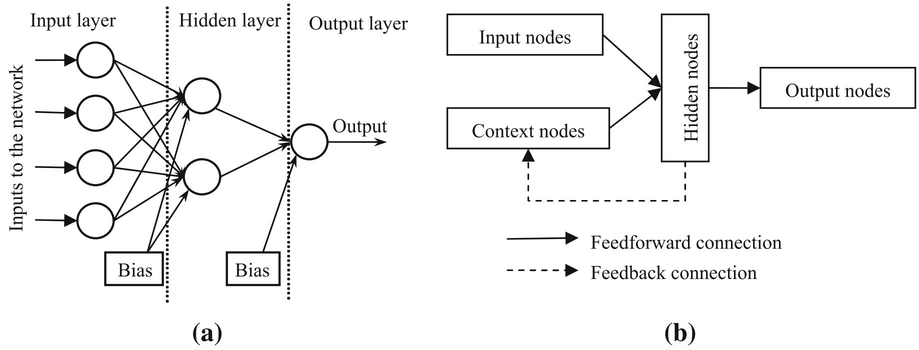

Fig. 2 **a** Structure of a typical MLP **b** structure of an Elman network

data-driven and self-adaptive nature (Zhang et al. 1998; Zhang 2003; Lemke and Gabrys 2010). The most common ANN forecasting model is the Multilayer Perceptron (MLP) with single layers of inputs, hidden and output nodes, as depicted in Fig. 2a. There are two methods of forecasting with ANNs: *iterative* and *direct* (Hamzaçebi et al. 2009; Lemke and Gabrys 2010). In the iterative approach, the single predicted value at each step is used as an input for the next prediction and the process is repeated until needed. On the other hand in the direct approach, the number of output neurons can be equal to the number of observations to be forecasted.

The appropriate model designing is crucial for the success of ANN forecasting. In this paper, we use the well-known *Bayesian Information Criterion (BIC)* (Zhang et al. 1998; Gooijer and Hyndman 2006) for selecting suitable ANN structures and the *Resilient Propagation (RP)* (Reidmiller and Braun 1993) as the network training algorithm.

4.4 The Elman artificial neural networks (EANNs)

An EANN is a recurrent neural network which consists of one extra layer viz. the context layer and two types of connections: *feedforward* and *feedback*. The outputs of the hidden layer in each step are recursively fed back to the context layer. This practice makes the network dynamic so that it can perform non linear time-varying mappings of the associated nodes (Lim and Goh 2005; Lemke and Gabrys 2010). The structure of an EANN model is shown in Fig. 2b.

It has been observed that EANNs require more hidden nodes than the simple feedforward ANNs in order to properly model the temporal relationships, although no general guidelines in this regard exist in literature (Lemke and Gabrys 2010). In this paper, we use about 20–30 hidden nodes and the training algorithm *traingdx* (Demuth et al. 2010) for fitting the EANN models.

5 Empirical works

5.1 Methods and data

Nine time series from different real-world domains are used in our empirical work. These are collected from the well-known Time Series Data Library (TSDL) (Hyndman 2011), a

Table 1 Descriptions of the nine time series datasets

Time series	Description	Type	Total size	Testing size
Canadian lynx (LYNX)	Number of lynx trapped per year in the Mackenzie River district of Northern Canada (1821–1934)	Stationary, nonseasonal	114	14
Sunspots (SNSPOT)	The annual number of observed sunspots (1700–1987)	Stationary, nonseasonal	288	67
USA real GNP (RGNP)	Real GNP of USA in billions of dollars (1890–1974)	Non-staionary, nonseasonal	85	15
Child births (BIRTHS)	Births per 10,000 of 23 year old women in USA (1917–1975)	Non-staionary, nonseasonal	59	10
Airline passengers (AP)	Monthly number of international airline passengers (in thousands) (January 1949–December 1960)	Monthly, seasonal	144	12
USA accidental deaths (USAD)	Monthly number of accidental deaths in USA (1973–1978)	Monthly, seasonal	72	12
Red wine (RW)	Monthly Australian sales of red wine (thousands of liters) (January 1980–July 1995)	Monthly, seasonal	187	19
Quarterly beer production (QBP)	Quarterly USA beer production (millions of barrels) (First quarter of 1975–fourth quarter of 1982)	Quarterly, seasonal	32	8
USA expenditure (UE)	Quarterly USA new plant/equipment expenditures (1964–1976)	Quarterly, seasonal	52	8

publicly available online repository of time series datasets. Each series is divided into three disjoint subsets, viz. *validation*, *training*, and *testing*. The training and validation sets are used for model fitting and parameter estimation, respectively, whereas the testing set is kept for assessing out-of-sample forecasting accuracies of fitted models. Here, we are considering short-term forecasting and so the size of each testing set is kept reasonably small. Table 1 provides descriptions about the time series datasets and Fig. 3 depicts their time plots. The horizontal and vertical axis of each time plot, respectively represents the indices and actual values of the observations.

The experiments in this study are carried out on MATLAB. The neural network toolbox (Demuth et al. 2010) is used for the ANN and EANN models. The network training parameters are either set to their default values or are selected on the basis of the performance of the corresponding model on the validation set. For each time series, the fitted ANN and EANN models are trained for a maximum of 2,000 epochs. The ARIMA models are implemented through the effective MATLAB scripts and programs, developed by Hurd (2012).

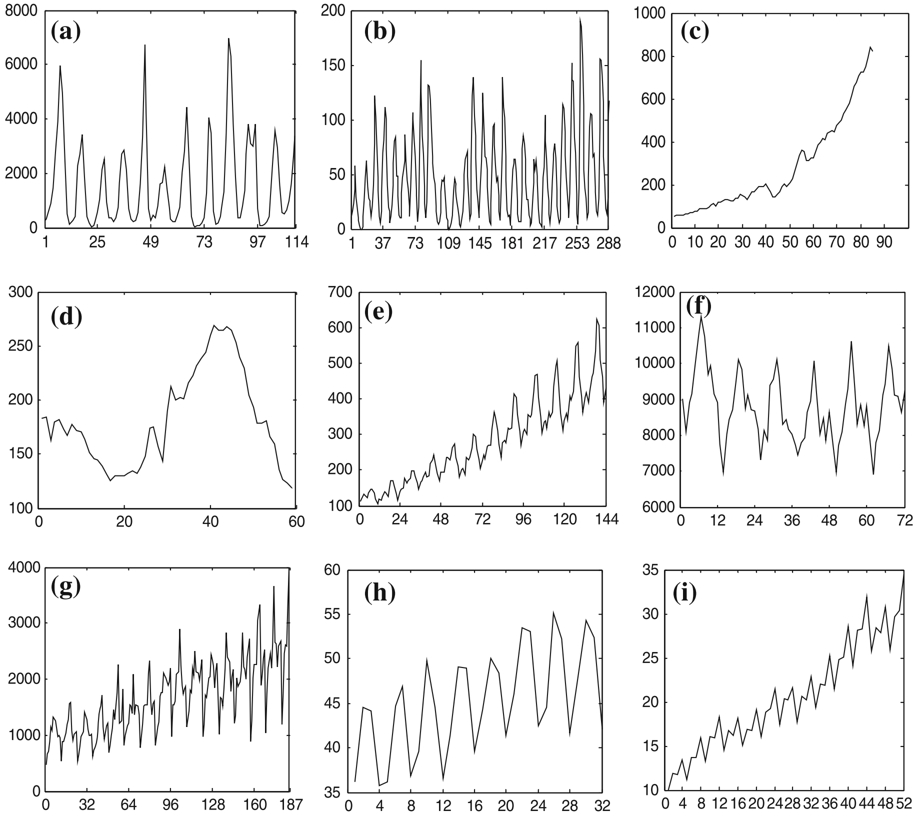

Fig. 3 Time plots of: **a** LYNX, **b** SNSPOT, **c** RGNP, **d** BIRTHS, **e** AP, **f** USAD, **g** RW, **h** QBP, **i** UE

The SVM tuning parameters are determined precisely through the grid search technique as recommended and utilized by Chapelle (2002).

The observations of each dataset are normalized to [0, 1] before fitting the forecasting models. These normalized observations are again transformed back to their original values in the testing phase. Mathematically, the data normalizations are carried out as follows:

$$\begin{aligned} y_i^{(\text{new})} &= \frac{y_i - y^{(\min)}}{y^{(\max)} - y^{(\min)}} \\ \forall i &= 1, 2, \ldots, N \end{aligned} \tag{16}$$

where, $\mathbf{Y} = \left[y_1, y_2, \ldots, y_{N_{\text{train}}}\right]^{\mathrm{T}}$ is the training dataset and $\mathbf{Y}^{(\text{new})} = \left[y_1^{(\text{new})}, y_2^{(\text{new})}, \ldots, y_N^{(\text{new})}\right]^{\mathrm{T}}$ is the normalized dataset, $y^{(\min)}$ and $y^{(\max)}$, respectively are the minimum and maximum values of the training dataset $\mathbf{Y}$.

The MSE and SMAPE are used for evaluating forecasting accuracies of all fitted methods. These are relative error measures and are quite useful for wisely comparing performances of different models. The less are the values of both these error statistics, the better is the forecasting performance of the fitted model.

Table 2 Forecasting results of the individual methods

Models	LYNX	SNSPOT	RGNP	BIRTHS	AP	USAD	RW	QBP	UE	Mean
ARIMA										
MSE	**0.0153**	803.34	1,906.05	98.00	0.0291	**10.07**	18.36	4.13	1.88	315.76
SMAPE	**3.40**	44.44	4.92	4.71	2.48	**2.88**	13.75	3.37	4.01	9.33
SVM										
MSE	0.0527	792.96	1,194.95	**61.87**	**0.0177**	16.20	**12.85**	**1.51**	**1.61**	231.34
SMAPE	6.12	33.38	4.00	**4.53**	**2.35**	3.79	**13.08**	**2.17**	**3.51**	**8.10**
Iterated ANN										
MSE	0.0319	**669.03**	813.38	93.37	0.0378	11.52	27.23	1.78	2.76	179.90
SMAPE	5.25	40.47	2.75	5.48	3.39	3.21	16.94	2.44	4.71	9.40
Iterated EANN										
MSE	0.0365	899.89	**322.17**	75.50	0.0333	14.16	38.99	2.34	2.58	**150.63**
SMAPE	5.48	**30.44**	**2.27**	5.41	3.32	3.91	18.82	2.88	3.94	8.50
Direct EANN										
MSE	0.0189	1,479.64	1,772.28	124.8	0.0967	22.57	19.70	5.31	1.94	380.71
SMAPE	3.80	53.62	5.12	5.66	5.64	5.00	16.08	3.58	3.80	11.37
Mean values										
MSE	0.0311	928.97	1,201.77	90.71	0.0429	14.91	23.43	3.01	2.15	251.67
SMAPE	4.81	40.47	3.81	5.16	3.44	3.76	15.74	2.89	3.99	9.34

5.2 Results and discussions

The LYNX and SNSPOT are both stationary series and exhibit regular patterns. Following previous well-known works (Zhang 2003), the logarithms to the base 10 of the lynx data are used in the present study. The ARIMA(12, 0, 0) (i.e. AR(12)) and ARIMA(9, 0, 0) (i.e. AR(9)) are found to be the most parsimonious ARIMA models for these two datasets, respectively. The $7 \times 5 \times 1$ and $11 \times 9 \times 1$ iterated ANN structures are respectively used for LYNX and SNSPOT series. Also, for fitting EANN models, 25 hidden nodes are used in both cases.

The RGNP and BIRTHS are both nonstationary time series, having quite irregular patterns. In particular, the RGNP series shows an apparent upward trend as can be seen from its time plot. The random walk is the suitable ARIMA model for these types of data (Zhang 2003). After residual analysis on the basis of BIC, the adequate iterated ANN models for these two datasets are determined to be $4 \times 10 \times 1$ and $6 \times 6 \times 1$, respectively. Corresponding direct EANN models have the respective structures $12 \times 25 \times 15$ and $8 \times 25 \times 10$. Each of the remaining five time series exhibits strong seasonal pattern which is clearly visible from the respective time plot. The SARIMA is the most appropriate ARIMA structure for this type of dataset (Box and Jenkins 1970). While fitting ANN and EANN models to these five time series, the numbers of input nodes are chosen to be equal to their respective seasonal periods. Thus, the numbers of input nodes are respectively 4 and 12 for quarterly and monthly series.

The obtained forecasting results of the individual and combination methods are presented in Tables 2 and 3, respectively. In each of these tables, the best results (i.e. least error measures) are presented in bold letters. Additionally, the forecast MSE values for AP, USAD and RW datasets are given in transformed scales (original MSE = MSE $\times 10^4$).

Table 3 Forecasting results of the combination methods

Combination methods	LYNX	SNSPOT	RGNP	BIRTHS	AP	USAD	RW	QBP	UE	Mean
Simple average										
MSE	0.0172	693.11	301.78	45.52	0.0189	7.28	15.85	1.47	**0.91**	118.44
SMAPE	3.75	30.85	2.05	3.62	2.23	2.69	12.71	2.42	**2.28**	6.96
Trimmed mean (40 % trimming)										
MSE	0.0222	743.97	301.22	39.74	0.0196	9.30	15.69	1.46	1.11	123.62
SMAPE	4.21	31.63	1.96	**3.33**	2.33	3.01	12.79	2.34	2.63	7.14
Median										
MSE	0.0232	795.67	296.84	**39.35**	0.0176	11.14	14.73	1.85	1.28	128.99
SMAPE	4.29	31.01	2.00	3.42	2.25	3.28	12.34	2.52	2.67	7.09
EB-I										
MSE	0.0133	650.66	271.06	46.51	0.0154	6.44	12.80	1.40	1.26	110.02
SMAPE	3.37	**30.75**	2.09	3.46	2.01	2.56	11.53	2.34	2.68	**6.75**
EB-II										
MSE	0.0112	**616.85**	302.03	49.36	0.0144	5.38	**12.32**	**1.33**	1.61	109.88
SMAPE	3.10	30.80	2.28	3.69	**1.98**	2.38	**11.38**	**2.29**	3.54	6.83
EB-III										
MSE	0.0129	633.97	273.94	46.63	0.0155	6.53	12.64	1.40	1.25	**108.49**
SMAPE	3.31	31.99	2.10	3.48	2.01	2.58	11.49	2.34	2.68	6.89
LSR										
MSE	0.0254	881.16	**263.48**	81.06	0.0188	6.41	16.91	2.78	1.70	139.28
SMAPE	4.10	113.62	**1.90**	4.31	2.27	2.64	13.86	3.06	3.84	16.62
DWS-I										
MSE	**0.0095**	637.13	316.25	50.87	0.0146	5.82	15.03	1.41	1.59	114.24
SMAPE	**2.83**	37.43	2.33	3.76	1.99	2.46	12.27	2.38	3.50	7.66
DWS-II										
MSE	0.0142	638.82	279.10	46.37	0.0168	6.73	15.59	1.45	0.98	109.90
SMAPE	3.46	32.32	2.05	3.59	2.08	2.61	12.58	2.41	2.29	7.04
Outperformance										
MSE	0.0152	686.79	285.65	45.87	**0.0143**	**5.21**	12.83	1.41	1.19	115.44
SMAPE	3.60	31.65	2.18	3.64	2.16	**2.29**	12.16	2.35	2.72	6.97
Mean values										
MSE	0.0164	697.81	289.14	49.13	0.0166	7.02	14.44	1.60	1.29	117.83
SMAPE	3.60	40.20	2.09	3.63	2.13	2.65	12.31	2.44	2.88	7.99

The following important observations are noticed after careful analysis of Tables 2 and 3:

(a) The least MSE and SMAPE values in Table 2 (which are shown in bold) do not occur uniformly through the rows and columns. This fact signifies that the obtained accuracies notably vary among the individual forecasting methods and also no single method alone could provide best results for all datasets.

(b) The mean MSE and SMAPE values across all combination methods are quite less than those across the individual forecasting methods for each dataset. This justifies that forecasts combinations significantly improve the accuracies together with reasonably decreasing the model selection risk.

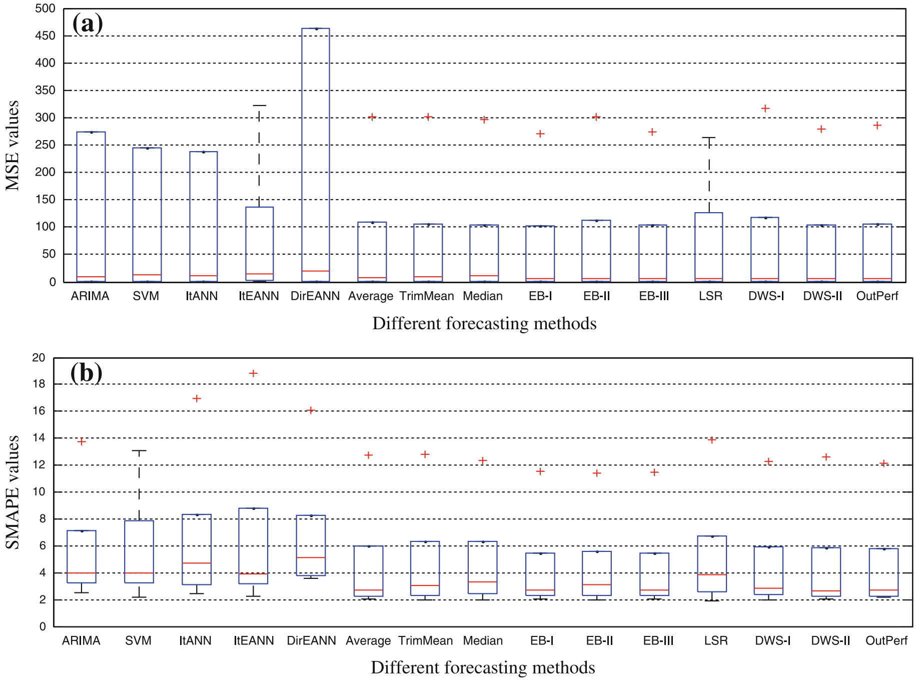

Fig. 4 Box plots of the effects of applying different forecasting methods on: **a** MSE, **b** SMAPE

(c) It can be clearly seen that on an overall basis many combination procedures significantly outperformed all individual methods. Even the mean values of MSE and SMAPE through the combined methods are less than the best MSE and SMAPE values among the individual methods for five and six datasets, respectively.

(d) Although, overall accuracies are clearly improved through combining forecasts but a particular combination scheme could seldom beat the best individual method in terms of both performance measures for all datasets. Thus, it is evident that the selection of a proper linear combination scheme for a time series should be done after a careful analysis.

We show the effects of applying different individual and combination methods on the obtained MSE and SMAPE values across all datasets using boxplots in Fig. 4a and 4b, respectively.

Figure 4a and 4b clearly depict that the values of both the error measures are reasonably reduced through applying different forecasts combination techniques. In Fig. 5a and 5b, we respectively show the number of times a combination scheme outperformed the best individual methods in terms of MSE and SMAPE values.

A look at the bar diagrams in Fig. 5a and 5b reveals some important facts. We can see that the three error-based schemes outperformed the best individual MSE values for nine datasets, whereas DWS-I, DWS-II and the outperformance could achieve this result for eight datasets. On the other hand, two and four combination schemes outperformed the best individual SMAPE values for seven and six datasets, respectively. The statistical methods performed relatively poor in terms of MSE but slightly better in terms of SMAPE. On the basis of both error measures, the five combination methods: EB-I, EB-III, DWS-I, DWS-II and outper-

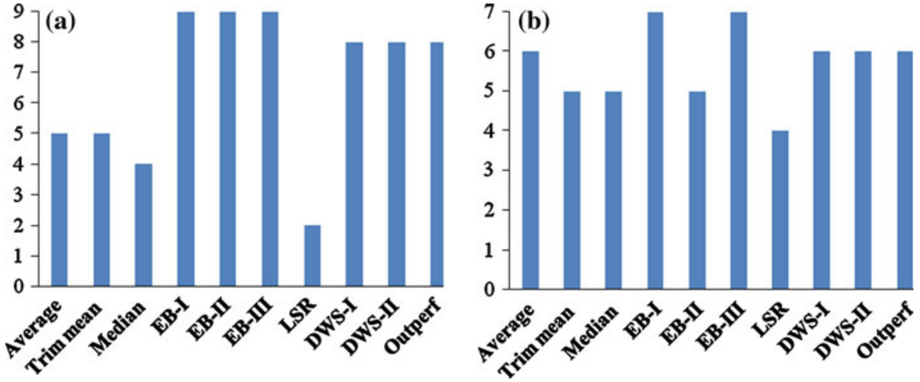

Fig. 5 *Bar diagrams* showing the total number of times a combination method outperformed the best individual forecasting methods in terms of: **a** MSE, **b** SMAPE

formance achieved notably good forecasting results, whereas LSR performed worst among all.

In order to have a fair idea about the relative performances of each combination method, we provide a robust ranking mechanism. At first, for all datasets we record the percentage reduction in the forecasting error of the best individual method by each linear combination scheme. Next, we calculate the average of the recorded nonnegative values for each combination scheme across all datasets. These averages are then sorted from smallest to largest and accordingly the combination methods are assigned consecutive numerical ranks, starting with one. A mathematical description of this ranking technique is presented below.

Let, n_f, n_c, respectively denote the numbers of individual and combination methods and n_d be the number of datasets. Let, e^c_{ij} $(i = 1, 2, \ldots, n_c;\ j = 1, 2, \ldots, n_d)$ be the error measure of the ith combination method for the jth dataset and e^b_j $(j = 1, 2, \ldots, n_d)$ be the best error measure among all individual methods for the jth dataset. We compute the percentage accuracy improvement of a combination method as follows:

$$p^c_{ij} = \begin{cases} \left(\frac{e^c_{ij} - e^b_j}{e^c_{ij}}\right) \times 100, & \text{if } e^c_{ij} \geq e^b_j \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

Then, the average percentage improvement of a combination method is calculated as:

$$p_i = \frac{1}{n_d} \sum_{j=1}^{n_d} p^c_{ij}$$
$$\forall i = 1, 2, \ldots, n_c. \tag{18}$$

Finally, the rank r_i, where $r_i \in \{1, 2, \ldots, n_c\}$ is assigned to the ith combination method, such that:

$$r_i \geq r_j, \text{if } p_i \geq p_j$$
$$\forall i, j = 1, 2, \ldots, n_c. \tag{19}$$

The assigned ranks to the ten forecasts combination methods on the basis of MSE and SMAPE are shown as the bar diagrams in Fig. 6a and 6b, respectively. In both these figures, the rank of each method is represented on the top of its corresponding bar.

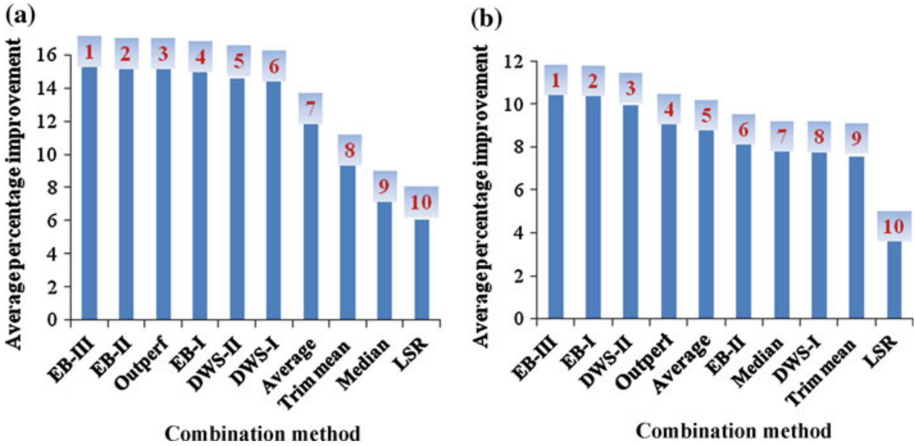

Fig. 6 Bar diagrams of the ranks of the combination methods on the basis of: **a** MSE, **b** SMAPE

Figure 6a and 6b depicts the overall relative position of each of the ten combination methods. We can see that on the basis of MSE the first three ranks are achieved by the schemes EB-III, EB-II and outperformance, whereas on the basis of SMAPE the first three ranks are achieved by the schemes EB-III, EB-I and DWS-II. In terms of both error measures, EB-III topped the rankings and LSR achieved the last rank. The worst position of LSR scheme can be attributed to the instability in optimizing the validation SSE. The outperformance method also performed quite well, attaining the 3rd and 4th ranks in terms of MSE and SMAPE, respectively. Among the three statistical methods, simple average has the best ranks on the basis of both error measures. It seems from Fig. 6a and 6b that through some robust weight selection technique one can achieve reasonably more accuracy than the naïve statistical combination methods.

As EB-III and LSR achieved the first and last positions, respectively through our ranking system for both error measures, so it may be interesting to visualize the diagrams which show the actual observations and the two corresponding forecasts by these two combination methods. These diagrams for all nine datasets are presented in Fig. 7.

6 Conclusions

Time series modeling and forecasting are fundamental to many practical decision-making processes. Improving forecasting accuracy is a challenging yet a crucial task which has been continuously attracting the attentions of researchers during the last two decades. Extensive works in this domain suggest that the forecasting accuracies can be substantially improved through combining forecasts from conceptually different models.

In this paper, we have meticulously investigated various benchmark linear combination methods for aggregating multiple forecasts. Five individual models and ten popular techniques for estimating the combining weights are considered. All combination methods are chosen to be relatively simple and less intricate as it is important to maintain the trade-off between accuracy improvement and computational cost. Empirical analysis is conducted on nine different real-world time series and performances of the fitted models are evaluated through MSE and SMAPE. Obtained results clearly demonstrate that most of the combina-

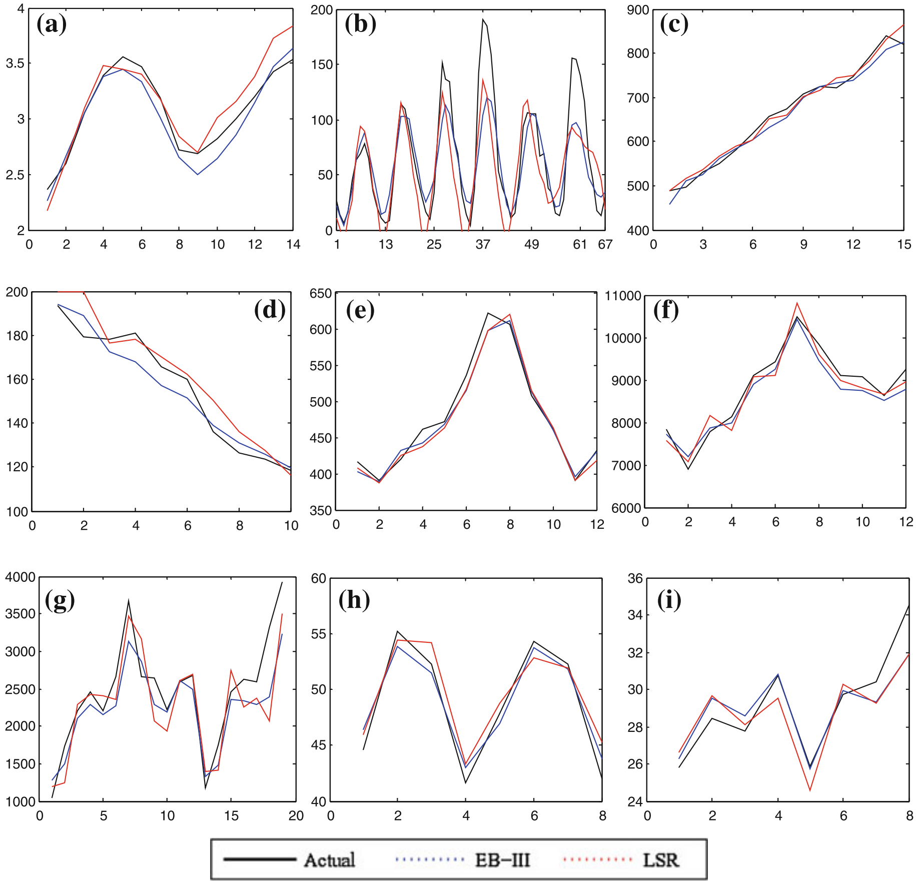

Fig. 7 Diagrams showing the actual and forecasted observations for: **a** LYNX, **b** SNSPOT, **c** RGNP, **d** BIRTHS, **e** AP, **f** USAD, **g** RW, **h** QBP, **i** UE

tion schemes performed significantly better than all individual methods. Moreover, the mean error measures across all combinations are quite less than those across the individual methods for each dataset. This signifies that together with improving accuracies, combining forecasts also reduces the model selection risk to a great extent. In order to compare different linear combinations, we have ranked them on the basis of their average percentage improvements over the best individual methods across all datasets. It is observed that the schemes EB-III, EB-II and outperformance achieved the first three ranks on the basis of MSE, whereas the schemes EB-III, EB-I and DWS-II achieved the first three ranks on the basis of SMAPE. For both error measures, the first and last ranks are attained by EB-III and LSR schemes respectively. The outperformance method also provided quite good combining accuracies and obtained 3rd and 4th ranks on the basis of MSE and SMAPE, respectively. The simple average is found to be superior among the three statistical methods. In a nutshell, our findings indicate that a linear combination of forecasts is a much better approach than relying on a single individual method. Further works in future can be performed with more diverse forecasting models as well as time series datasets.

Acknowledgments The authors are indebted to the reviewers for their constructive suggestions which significantly helped in improving the quality of this paper. In addition, the first author expresses his gratitude to the Council of Scientific and Industrial Research (CSIR), India for the obtained financial support in performing this research work.

References

Aksu C, Gunter S (1992) An empirical analysis of the accuracy of SA, OLS, ERLS and NRLS combination forecasts. Int J Forecast 8(1): 27–43

Andrawis RR, Atiya AF, El-Shishiny H (2011) Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition. Int J Forecast 27(3):672–688

Armstrong JS (2001) Principles of forecasting: a handbook for researchers and practitioners. Kluwer Academic Publishers, Boston, USA

Bates JM, Granger CWJ (1969) Combination of forecasts. Oper Res Q 20(4): 451–468

Box GEP, Jenkins GM (1970) Time series analysis, forecasting and control, 3rd edn. Holden-Day, California

Bunn D (1975) A Bayesian approach to the linear combination of forecasts. Oper Res Q 26(2): 325–329

Chan CK, Kingsman BG, Wong H (2004) Determining when to update the weights in combined forecasts for product demand—an application of the CUSUM technique. Eur J Oper Res 153(3): 757–768

Chapelle O (2002) Support vector machines: introduction principles, adaptive tuning and prior knowledge. Ph.D. Thesis, University of Paris, France

Clemen RT (1989) Combining forecasts: a review and annotated bibliography. J Forecast 5(4): 559–583

De Gooijer JG, Hyndman RJ (2006) 25 years of time series forecasting. J Forecast 22(3): 443–473

De Menezes LM, Bunn DW, Taylor JW (2000) Review of guidelines for the use of combined forecasts. Eur J Oper Res 120(1): 190–204

Demuth H, Beale M, Hagan M (2010) Neural network toolbox user's guide. The MathWorks, Natic

Frietas PSA, Rodrigues AJL (2006) Model combination in neural-based forecasting. Eur J Oper Res 173(3): 801–814

Hamzaçebi C, Akay D, Kutay F (2009) Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting. Expert Syst Appl 36(2): 3839–3844

Hurd HL (2012) A collection of MATLAB programs to do various time series tasks. http://www.stat.unc.edu/faculty/hurd.html. Accessed 12 Feb 2012

Hyndman RJ (2011) Time series data library (TSDL). http://robjhyndman.com/TSDL/

Jose VRR, Winkler RL (2008) Simple robust averages of forecasts: some empirical results. Int J Forecast 24(1): 163–169

Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley, NJ

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

Lim CP, Goh WY (2005) The application of an ensemble of boosted elman networks to time series prediction: a benchmark study. J Comput Intell 3(2): 119–126

Makridakis S, Hibon M (2000) The M3 competition: results, conclusions and implications. Int J Forecast 16(4): 451–476

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.
J Forecast 1(2): 111–153

Newbold P, Granger CWJ (1974) Experience with forecasting univariate time series and the combination of forecasts (with discussion). J R Stat Soc A 137(2): 131–165

Reidmiller M, Braun H (1993) A direct adaptive method for faster backpropagation learning: the rprop algorithm. In: IEEE international conference on neural networks (ICNN), San Francisco, USA, pp 586–591

Stock JH, Watson MW (2004) Combination forecasts of output growth in a seven-country data set. J Forecast 23(6): 405–430

Suykens JAK, Vandewalle J (1999) Least squares support vector machines classifiers. Neural Process Lett 9(3): 293–300

Terui N, Van Dijk HK (2002) Combined forecasts from linear and nonlinear time series models. Int J Forecast 18(3): 421–438

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Winkler RL, Makridakis S (1983a) The combination of forecasts. J R Stat Soc A 146((2): 150–157

Winkler RL, Makridakis S (1983b) Averages of forecasts: some empirical result. Manag Sci 29((9): 987–996

Zhang GP (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50: 159–175

Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: the state of the art. Int J Forecast 14((1): 35–62

Zou H, Yang Y (2004) Combining time series models for forecasting. Int J Forecast 20(1): 69–84