Air passenger forecasting by using a hybrid seasonal decomposition and least squares support vector regression approach

Gang Xie[a,*] , Shouyang Wang[a], Kin Keung Lai[b]

[a]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[b]Department of Management Sciences, City University of Hong Kong, Hong Kong

Abstract

In this study, a hybrid approach based on seasonal decomposition (SD) and least squares support vector regression (LSSVR) model is proposed for air passenger forecasting. In the formulation of the proposed hybrid approach, the air passenger time series are first decomposed into three components: trend-cycle component, seasonal factor and irregular component. Then the LSSVR model is used to predict the components independently and these prediction results of the components are combined as an aggregated output. Empirical analysis shows that the proposed hybrid approach is better than other benchmark models, indicating that it is a promising tool to predict complex time series with high volatility and irregularity.

Keywords: Hybrid approach; Seasonal decomposition; Least squares support vector regression; Air passenger forecasting

*corresponding author.Tel. +86-10-62610229, Fax: +86-10-62541823.
E-mail address: gxie@amss.ac.cn (G. Xie)

1. Introduction

As incomes and populations have increased and the structure of industry has changed, air transportation has grown considerably around world. The gradual freeing of trade across the globe has added to this growth (Alekseev and Seixas, 2009). With this overall expansion of demand, the patterns of traffic have also changed and become more complex. For example, there is competition between high-speed railroad service and air transport (Park and Ha, 2006). Air passenger forecasting provides a key input into decisions of daily operation management and infrastructure planning of airports and air navigation services, and for aircraft ordering and design (Scarpel, 2013). To meet these new conditions, airlines and airports require enhanced forecasting tools.

Several methods have been used for air passenger forecasting, and including second-degree polynomial (Profillidis, 2000), autoregressive integrated moving average (ARIMA) model and seasonal autoregressive integrated moving average (SARIMA) model (Samagaio and Wolters, 2010), logit model (Dupuis et al., 2012), and gravity model (Grosche et al., 2007). In particular, Alekseev and Seixas (2009) developed a hybrid approach based on decomposition and back-propagation neural network (BPNN) for air transport passenger analysis. The results showed that forecasting performance was improved when data preprocessing of decomposition was fully adopted.

However, BPNN often suffers local minima and over-fitting, and it is sensitive to parameter selection (Xie et al., 2013). Support vector machine (SVM) has been

proved to possess excellent capability for classification and prediction, by minimizing an upper bound of the generalization error (Vapnik, 1995). SVM can be applied to classification and regression, i.e. support vector classification (SVC) and support vector regression (SVR). Since it adopts the structural risk minimization (SRM) principle, SVR can alleviate the over-fitting and local minima issues and its solution is more stable and globally optimum (Xie et al., 2013).

Moreover, in order to reduce the computational complexity of SVM, Suykens and Vandewalle (1999) proposed least squares support vector regression (LSSVR) model, which solves a system of equations instead of a quadratic programming (QP) problem and leads to significantly improved speed of calculations. Due to these advantages of LSSVR model, we employ it as the prediction model for air passenger forecasting. In addition, since using only the univariate time series can reduce the data dimensionality thus improve generalization and forecasting performance (Wang et al., 2011), we will focus on univariate time series models in this study.

To the best of our knowledge, the application of LSSVR for air passenger forecasting has not been studied in the literature. In this study, LSSVR model is integrated with seasonal decomposition (SD) to form a hybrid approach for air passenger forecasting. Empirical analysis is implemented to compare the proposed hybrid approach with other benchmark methods in terms of measurement criteria on the forecasting performance. Finally, some related issues are discussed and conclusions are drawn.

The remainder of the paper is organized as follows. The hybrid approach

SD-LSSVR for air passenger forecasting is proposed in Section 2. Section 3 illustrates the problem by using empirical analysis with experiments. Then, some related issues are discussed in Section 4. Section 5 draws conclusions and suggests some directions for future investigations.

2. The hybrid approach based on SD and LSSVR model

In this section, the overall formulation process of the SD based LSSVR hybrid approach is presented. First, LSSVR model and SD technique are briefly introduced. Then the hybrid approach SD-LSSVR is formulated and corresponding steps involved in its implementation are described in details.

2.1 LSSVR model

In a least squares support vector regression (LSSVR) model, the regression problem can be transformed into an optimization problem, as follows.

$$\text{Min} \quad (w^T w)/2 + (\gamma \sum_{i=1}^{l} e_i^2)/2 \tag{3}$$

$$\text{s.t.} \quad y_i = w^T \varphi(x_i) + b + e_i, \ (i = 1, 2, \ldots, \ l)$$

where e_i is the error variable and γ is the penalty parameter. γ is used to control the minimization of estimation error and the function smoothness.

In order to solve the optimization problem, the Lagrange function is developed as

$$L(w, b, e, \alpha) = (w^T w)/2 + (\gamma \sum_{i=1}^{l} e_i^2)/2 - \sum_{i=1}^{l} \alpha_i [w^T \varphi(x_i) + b + e_i - y_i] \tag{4}$$

where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_l)$ is the Lagrange multiplier. Differentiating L with respect to variables w, b, e and α, we obtain

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{l} \alpha_i \varphi(x_i), \quad (5)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{l} \alpha_i = 0 \quad , \quad (6)$$

$$\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i, \quad (7)$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow w^T \varphi(x_i) + b + e_i - y_i = 0. \quad (8)$$

After solving the above functions, we can obtain the solution of the problem in the following form:

$$f(x) = \sum_{i=1}^{l} w_i K(x, x_i) + b \quad (9)$$

where $K(\cdot)$ is the kernel function. Here, the usual Gaussian RBF $K(x, x_i) = \exp[\|x - x_i\| / (2\sigma^2)]$ with a width of σ is employed.

2.2 Seasonal decomposition

In order to capture seasonal characteristics of observations in different years, we use the most popular seasonal decomposition (SD) method X-12-ARIMA, which is the Census Bureau's latest seasonal adjustment programme (Findley et al., 1998). X-12-ARIMA method decomposes time series y_t into three components, i.e. trend-cycle component tc_t, seasonal factor sf_t and irregular component ir_t, which can be combined into the original data in additive and multiplicative forms, as follows:

$$y_t = tc_t + sf_t + ir_t, \quad (3)$$

$$y_t = tc_t \times sf_t \times ir_t. \quad (4)$$

Comparing the two forms of SD, the multiplicative decomposition is a more

suitable choice for most seasonal time series. The main reasons for the priority are summarized into two aspects: On one hand, the seasonal factor of the multiplicative form is a relative value of the original series; On the other hand, most seasonal time series with positive values has the characteristic that the scale of seasonal oscillations increases in the level of original time series (U.S. Census Bureau, 2011). As a consequence, the multiplicative form is employed for SD via X-12-ARIMA program in this study.

2.3 The hybrid SD-LSSVR approach

Generally speaking, there are three main steps involved in the proposed hybrid approach, i.e. decomposition, single forecast and aggregation. After the three components are predicted by LSSVR as $\hat{tc}_t$, $\hat{sf}_t$ and $\hat{ir}_t$ respectively, they are aggregated as an output $\hat{y}_t$ as follows

$$\hat{y}_t = \hat{tc}_t \times \hat{sf}_t \times \hat{ir}_t . \quad (5)$$

The overall process of the SD-LSSVR approach can be described in **Fig. 1** as the following three main steps:

(1) The original time series y_t (t=1, 2, …, T) is decomposed into seasonal factor (SF) sf_t, trend cycle (TC) tc_t and irregular component (IR) ir_t via multiplicative SD.

(2) For decomposed components sf_t, tc_t and ir_t, the LSSVR is used as a forecasting tool to fit the decomposed components, and to make the corresponding prediction for each one as $\hat{tc}_t$, $\hat{sf}_t$ and $\hat{ir}_t$.

(3) Prediction results of SF, TC and IR are multiplied as an aggregated output $\hat{y}_t$.

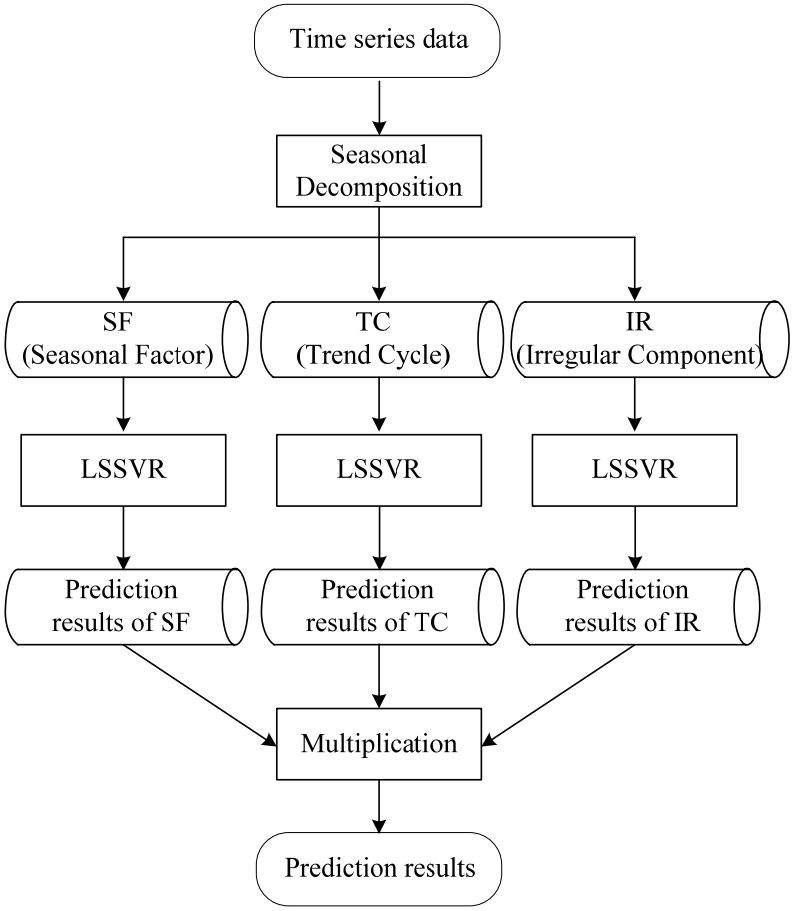

Fig. 1 The overall process of the SD-LSSVR approach

In order to verify the effectiveness of the proposed hybrid approach, the time series of air passengers at Hong Kong International Airport (HKIA) are used as a testing target, which is illustrated in the next section.

3. Empirical analysis

3.1 Data description and experiment design

Hong Kong International Airport (HKIA), located in South China, is one of the biggest airports around the world. In 2012, the number of passengers at HKIA was 55.66 million, with a growth rate of 5.5%. The time series data in this study are

obtained from CEIC Database (http://www.ceicdata.com/). The sample data are monthly data of passengers at HKIA, covering the period from January 1999 to February 2013, with a total of 170 observations, as shown in **Fig. 2**.

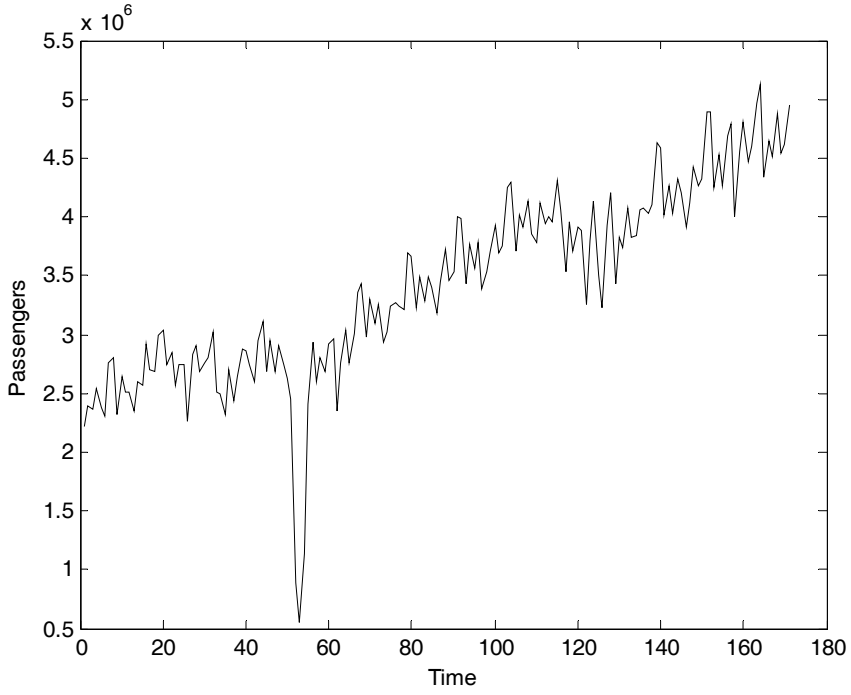

Fig. 2 Monthly air passengers at HKIA

In experiments, training dataset is used to determine the unknown parameters of the pre-defined models. Testing dataset is used to evaluate the forecasting performance. For each out-of-sample observation, its previous data are used as training samples to set the forecasting model for making one-step-ahead forecasting, where the lag period which is determined by analyzing the autocorrelation and partial correlation of the time series.

For comparison of forecasting performance of multiple different models, we choose *RMSE*, *MAE* and *MAPE* as the criteria of measuring level prediction accuracy, as

follows:

$$RMSE = \sqrt{[\sum_{t=1}^{n}(y_t - \hat{y}_t)^2]/n}\,, \tag{12}$$

$$MAE = \sum_{t=1}^{n}|y_t - \hat{y}_t|/n\,, \tag{13}$$

$$MAPE = (100\sum_{t=1}^{n}|1 - \hat{y}_t / y_t|)/n \tag{14}$$

where n is sample size, y_t is real value of the observation and $\hat{y}_t$ is the corresponding forecast in the t th month.

Apart from the level prediction accuracy, directional prediction accuracy is another important criterion for forecasting models. The performance to predict movement direction can be measured by a directional statistic as follows:

$$D_{stat} = \sum_{t=1}^{n} d_t / n \tag{15}$$

where $d_t = \begin{cases} 1, & (y_t - y_{t-1})(\hat{y}_t - y_{t-1}) \geq 0 \\ 0, & otherwise \end{cases}$.

Note that *RMSE*, *MAE* and *MAPE* are measures of the deviation between real and predicted values. Therefore, the forecasting performance is better when the values of these measures are smaller. In addition, D_{stat} provides the correctness of the predicted direction and can also be utilized to evaluate the prediction accuracy. The higher D_{stat} value is, the better forecasting performance is.

Also, Back-Propagation Neural Networks (BPNN) is used as a benchmark model for air passenger forecasting. In addition, we present another hybrid empirical mode decomposition (EMD) and LSSVR (EMD-LSSVR) approach for comparison, and its overall process is shown in **Fig. 3**. In the formulation of EMD-LSSVR approach, the

air passenger time series are first decomposed into several intrinsic mode function (IMF) components and one residual component. Then the LSSVR model is used to predict the components independently and these prediction results of the components are combined as an aggregated output.

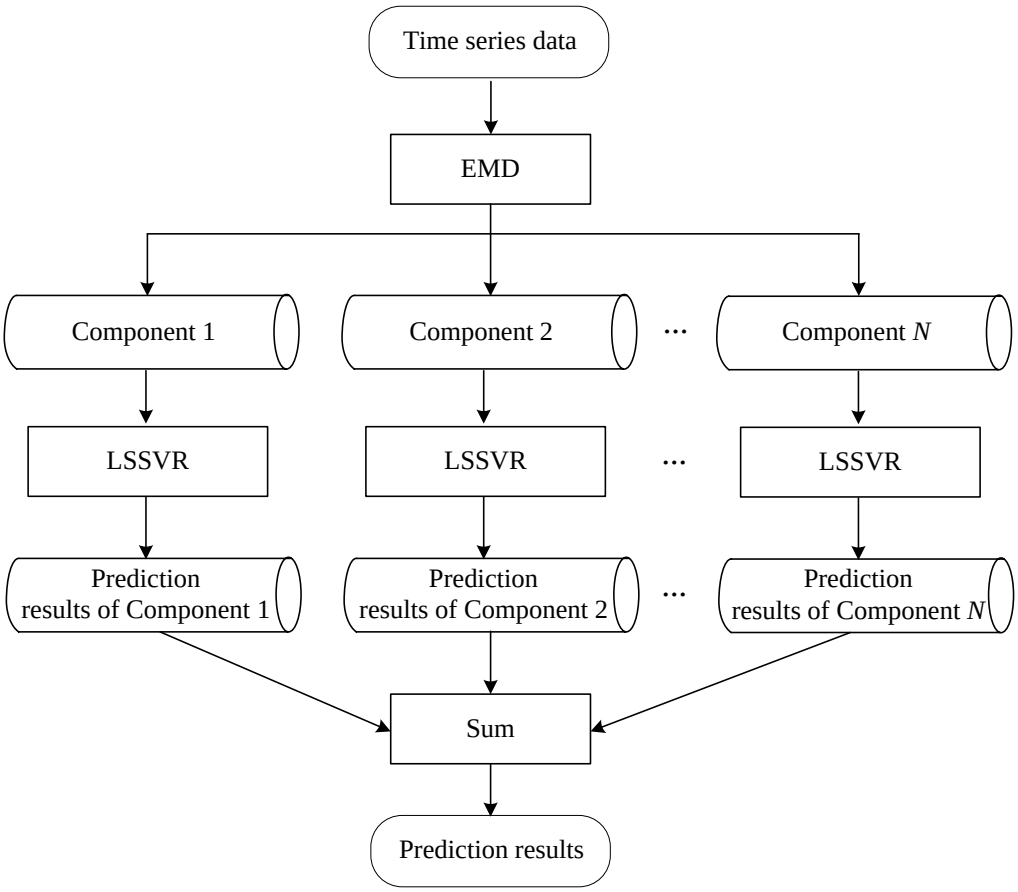

Fig. 3 The overall process of the EMD-LSSVR approach

In this study, ARIMA is implemented in the EVIEWS software package, which is produced by Quantitative Micro Software Corporation. Three single models, BPNN, LSSVR and EMD, are implemented via MATLAB software package.

3.2 Experiment results and analysis

Using 70% monthly data of the time series at HKIA as training dataset (119 observations and the period covering from January 1999 to November 2008), we

apply LSSVR model with Gaussian Kernel for making the one-step-ahead forecasting.

In the LSSVR model, the values of γ and σ^2 parameters are first determined via 10-fold cross-validation grid search method in the range of [0.01, 10000], and then adjusted using the trial-and-error approach to produce the smallest error in the training set (Tay and Cao, 2001).

Apart from single LSSVR, other two single forecasting models ARIMA and BPNN are used for comparison purpose. In the ARIMA(p,d,q) model, the best model for each training sample is determined through the minimization of Schwarz Criterion (SC). The BPNN in this study used 3 input neurons (i.e., p=3), 10 hidden nodes (i.e., q=10) and one output neuron. The BPNN models are iteratively run 10,000 times to train the model using the training subset.

After multiplicative SD is implemented via X-12-ARIMA program for the time series of air passenger at HKIA, we obtain trend cycles (TC), seasonal factors (SF) and irregular components (IR). Then, LSSVR model is used for fitting and forecasting of decomposed components. The real data and forecasts of out-of-sample datasets at HKIA are shown in **Fig. 4**.

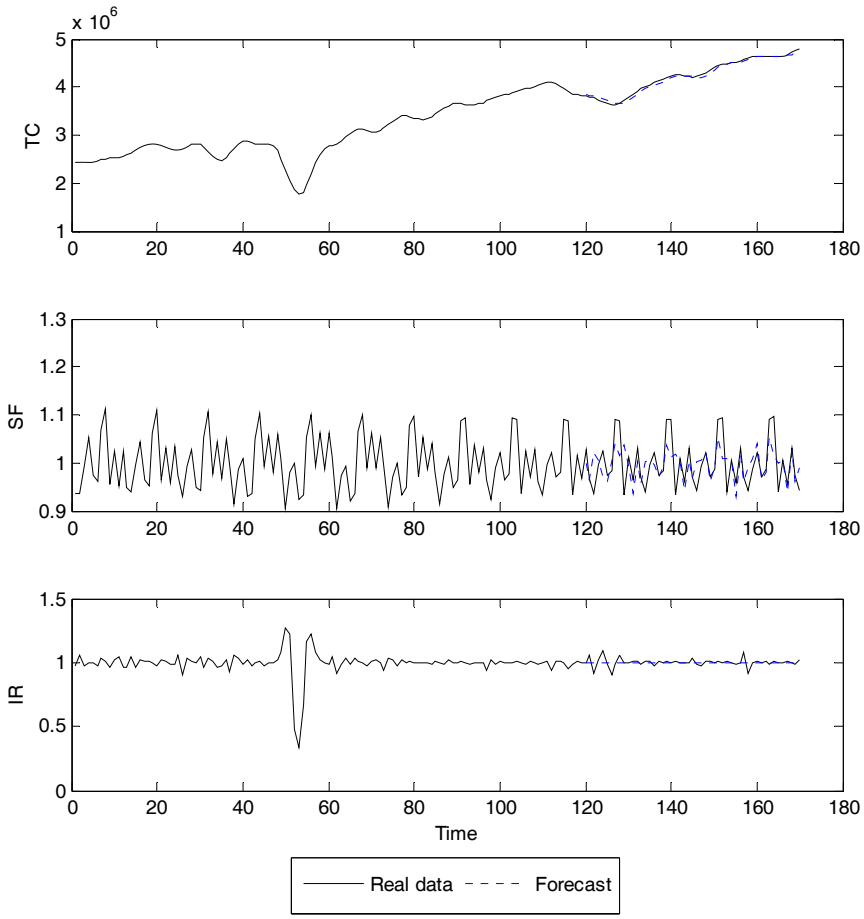

Fig. 4 Comparison of the real data and forecasts in SD-LSSVR approach

In hybrid approach EMD-LSSVR, the air passenger time series at HKIA are first decomposed into 6 intrinsic mode function (IMF) components and one residual component. Then the LSSVR model is used to predict the components. The comparison of forecast values of testing dataset by LSSVR model and real values of the component series are shown in **Fig. 5** as follows.

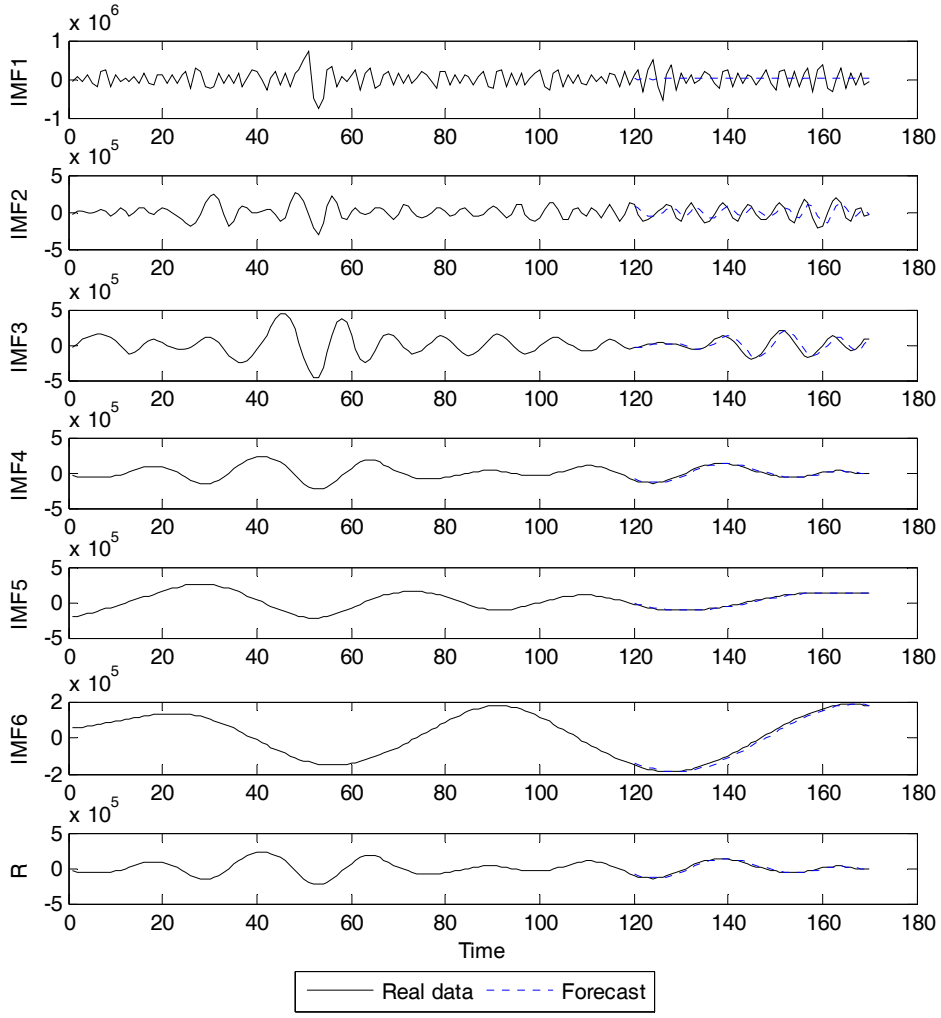

Fig. 5 Comparison of the real data and forecasts in EMD-LSSVR approach

Furthermore, for the robustness evaluation of different methods, we use different ratios of training dataset to sample sizes and three relative ratios of 70%, 80%, and 90% are considered. Then, the forecasting performance in both level accuracy and directional accuracy is listed as in **Table 1**.

Table 1. Robustness evaluation of approaches by different training and testing sample sizes.

Relative ratio(%)	Models	Testing data			
		RMSE	*MAE*	*MAPE*	D_{stat}
70	ARIMA	117.9	109.7	25.1	0.412
	BPNN	55.6	44.8	10.4	0.549
	LSSVR	45.9	35.4	8.3	0.431
	EMD-LSSVR	23.8	18.9	4.6	0.843
	SD-LSSVR	24.8	18.5	4.5	0.843
80	ARIMA	118.9	114.8	25.4	0.412
	BPNN	57.2	48.2	10.4	0.441
	LSSVR	49.8	37.7	8.3	0.382
	EMD-LSSVR	22.3	17.6	3.9	0.882
	SD-LSSVR	22.0	16.4	3.7	0.912
90	ARIMA	110.6	107.5	23.1	0.353
	BPNN	68.3	63.1	13.4	0.353
	LSSVR	55.3	41.8	9.1	0.294
	EMD-LSSVR	23.4	18.5	4.1	0.824
	SD-LSSVR	21.7	14.4	3.2	0.882

After illustrating the proposed approaches by experiments, we make further analysis of some related issues in the following section.

4. Discussion

This section presents an in-depth discussion on the forecasting performance of SD-LSSVR approach. The forecasting performance of SD-LSSVR approach is analyzed firstly on the basis of the experimental results presented in Section 3. Then, deep insights are given for air passenger forecasting.

4.1 Performance comparison and analysis

Using the experiment design and methodologies mentioned above, the forecasting experiments for air passenger at HKIA is performed and accordingly the forecasting performances are evaluated by the four main measure criteria.

In Table 1, the ARIMA model performs the poorest because it is a class of typical linear model and it cannot capture the nonlinear patterns and seasonal characteristic existing in the data series. Also, the forecasting performance of single BPNN and LSSVR models is not good. The reason may be that the data of air passenger at HKIA are complex time series with high volatility and irregularity.

Obviously, the forecasting performance of hybrid approaches EMD-LSSVR and SD-LSSVR is much better than single models, including ARIMA, BPNN and LSSVR models. However, SD-LSSVR is better than EMD-LSSVR, except for RMSE when relative ratio of training dataset to testing dataset is 70%. The reason may be that the seasonality within the time series of air passenger is not described by EMD-LSSVR.

4.2 Deep insights in air passenger forecasting

As ARIMA can’t capture nonlinear characteristics but linear component of time series, its forecasting performance is inferior to other approaches in Table 1. Therefore, nonlinearity should be captured for better forecasting performance. When the data of air passenger are complex time series with high volatility and irregularity, single AI models, i.e. LSSVR and BPNN, are not excellent in air passenger forecasting. On the whole, the proposed hybrid approaches, i.e., SD-LSSVR and EMD-LSSVR, outperform other single approaches. This suggests that decomposition is efficient can effectively improve performance in the case of air passenger forecasting. Also, the results imply that the proposed hybrid approaches can be applied to other complex time series forecasting problems with seasonality.

5. Conclusions and Future Work

In this study, on the basis of seasonal decomposition (SD) and least squares support vector regression (LSSVR) model, we proposed a hybrid SD-LSSVR approach for air passenger forecasting. Based on time series of air passenger at Hong Kong International Airport, empirical analysis is used to illustrate the proposed hybrid approach and compare it with other benchmark methods. Finally, some related issues were discussed and conclusions were drawn.

The contribution of this study is that LSSVR model is firstly used for air passenger forecasting. A hybrid approach is developed for the comparison with benchmark methods. The investigation suggests that decomposition is an effective way to air passenger forecasting. It is important to describe the seasonal characteristic and nonlinear nature of air passenger series for better forecasting performance.

It is expected that future research would benefit from concentrating on other methods for air passenger forecasting, using data from a wider sample of international airports.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 70871107), China Postdoctoral Science Foundation (Grant No. 20060400103) and The Royal Academy of Engineering for research exchanges with China and India scheme.

References

Abdel-Aal, R.E., Al-Garni, A.Z., 1997. Forecasting monthly electric energy consumption in eastern Saudi Arabia using univariate time-series analysis. Energy 22 (11), 1059-69.

Alekseev, K.P.G., Seixas, J.M., 2009. A multivariate neural forecasting modeling for air transport - Preprocessed by decomposition: A Brazilian application. Journal of Air Transport Management 15, 212-216.

Cline, R.C., Ruhl, T.A., Gosling, G.D., Gillen, D.W., 1998. Air transportation demand forecasts in emerging market economies: a case study of the Kyrgyz Republic in the former Soviet Union. Journal of Air Transport Management 4, 11-23.

Dupuis, C., Gamache, M., Pagé, J.-F., 2012. Logical analysis of data for estimating passenger show rates at Air Canada. Journal of Air Transport Management 18, 78-81.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B.C., 1998. New capabilities and methods of the X-12-ARIMA seasonal adjustment program. Journal of Business and Economic Statistics 16, 127-176.

Grosche, T., Rothlauf, F., Heinzl, A., 2007. Gravity models for airline passenger volume estimation. Journal of Air Transport Management 13, 175-183.

Huth, W., Eriksen, S.E., 1987. Airline traffic forecasting using deterministic and stochastic time series decomposition. Logistics and Transportation Review 23, 401-409.

Lai, S., Lu, W., 2005. Impact analysis of September 11 on air travel demand in the USA. Journal of Air Transport Management 11, 455-458.

Park, Y., Ha, H.-K., 2006. Analysis of the impact of high-speed railroad service on air transport demand. Transportation Research Part E 42, 95-104.

Profillidis, V., 2000. Econometric and fuzzy models for the forecast of demand in the airport of Rhodes. Journal of Air Transport Management 6, 95-100.

Profillidis, V.A., 2012. An ex-post assessment of a passenger demand forecast of an airport. Journal of Air Transport Management 25, 47-49.

Rengaraju, V.R., Arasan, V.T., 1992. Modeling for air travel demand. Journal of Transportation Engineering 118, 371-380.

Samagaio, A., Wolters, M., 2010. Comparative analysis of government forecasts for the Lisbon Airport. Journal of Air Transport Management 16, 213-217.

Scarpel, R.A., 2013. Forecasting air passengers at São Paulo International Airport using a mixture of local experts model. Journal of Air Transport Management 26, 35-39.

Sellner, R., Nagl, P., 2010. Air accessibility and growth - The economic effects of a capacity expansion at Vienna International Airport. Journal of Air Transport Management 16, 325-329.

Shmueli, D., 1998. Applications of neural networks in transport planning. Progress in Planning 50, 141-204.

Smyth, A., Christodoulou, G., Dennis, N., AL-Azzawi, M., Campbell, J., 2012. Is air transport a necessity for social inclusion and economic development? Journal of Air Transport Management 22, 53-59.

Suykens, J.A.K., Vandewalle, J., 1999. Least squares support vector machine

classifiers. Neural Processing Letters 9, 293-300.

Tay, F.E.H., Cao, L., 2001. Application of support vector machines in financial time series forecasting. Omega 29, 309-317.

U.S. Census Bureau, 2011. X-12 ARIMA reference manual version 0.3.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Xie, G., Wang, S., Zhao, Y., Lai, K.K., 2013. Hybrid approaches based on LSSVR model for container throughput forecasting: A comparative study. Applied Soft Computing 13 (5), 2232-2241.