International Journal of Computer Science and Applications,
©Technomathematics Research Foundation
Vol. 13, No. 1, pp. 105 – 115, 2016

# A PARTIALLY COMBINED FRAMEWORK: FORECASTING CONTAINER THROUGHPUT WITH BIG DATA

ANQIANG HUANG

*School of Economy and Management, Beijing Jiaotong University, No.3 Shangyuncun, Haidian District,*
*Beijing 100044, China*
*aqhuang@bjtu.edu.cn*

YAFEI ZHENG

*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, No.55, Haidian District,*
*Beijing 100190, China*
*zhengyafei11@mails.ucas.ac.cn*

ZHENJI ZHANG

*School of Economy and Management, Beijing Jiaotong University, No.3 Shangyuncun, Haidian District,*
*Beijing 100044, China*
*zhjzhang@bjtu.edu.cn*

XIANLIANG SHI

*School of Economy and Management, Beijing Jiaotong University, No.3 Shangyuncun, Haidian District,*
*Beijing 100044, China*
*xlshi@bjtu.edu.cn*

GUOWEI HUA

*School of Economy and Management, Beijing Jiaotong University, No.3 Shangyuncun, Haidian District,*
*Beijing 100044, China*
*gwhua@bjtu.edu.cn*

This study proposes a partially-combined forecasting framework for container throughput based on big data composed of structured historical data and unstructured data. Under the proposed framework, the structured data (the original time series) is firstly decomposed into linear and nonlinear components. Seasonal auto-regression integrated moving average model (SARIMA) is adopted to capture and forecast the linear component, and a combined model composed of least squares support vector regression (LSSVR) and artificial neural network (GP), is applied to modeling the nonlinear component. Next, unstructured data is analyzed by an expert system. With the synthesized expert judgment, the forecasts of linear and nonlinear components are integrated into a final forecast. For the illustration and verification purpose, an empirical study is conducted with the data of Qingdao Port. The results show that the model under the proposed framework significantly outperforms its competitive rivals.

*Keywords*: Container throughput forecast; Partially-combined forecasting model; Big data; Intelligent model.

## 1. Introduction

Information sharing has been recognized as one of the most important approaches to promoting supply chain collaboration and thus leading to ultimate cost savings. However, there is a big gap between the ideal integrated supply chains and the reality, as argued by [Gunasekaran (2004)]. [Premkumar (2000)] listed 6 critical issues concerning successful supply chain collaboration, one of which was reluctance to share information. [Davis and Spekman (2004)] claimed that a number of companies have not yet made any effort to address the 6 issues to enable effective extended supply chain collaboration.

There are several reasons accounting for reluctance to share information among the players along the supply chain. First and the most direct, the players are afraid that the information will be used unfairly to the partners advantage. Second, power regimes and sub-regimes existing in supply chains will impede supply chain optimization [Watson (2004)], considering that sharing information and integrating systems may lead to major upheavals in the power structure [Premkumar (2000)]. Third, [Thonemann (2002)] mathematically demonstrated that although players along supply chains may gain certain benefits from advance information sharing, it actually tends to increase the bullwhip effect. Finally, noise (accurate information) will lead to terrible decision making, a typical example is the telecom industry demand chain where some partners were double forecasting and ration gaming [Heikkilä (2002)]. It is notable that, current supply chains in reality are not collaborative [Davis and Spekman (2004)], owing to the above reasons.

Given the above considerations, it is very difficult for players to acquire valuable information from others in the real world. Fortunately, demand forecast generated from historical data can provide some kind of valuable information. For example, [Aviv (2001)] and [Cachon and Lariviere (2001)] suggested demand forecast sharing to address problems arising from the bullwhip effect. [Yue and Liu (2006)] analyzed the value of demand forecast sharing in a direct channel supply chain. Therefore, the problem of forecasting distorted demand is of significant importance to businesses [Carbonneau *et al.* (2008)].

## 2. Objectives of Research

Container shipping is one of the most important transport approaches for international cargo trade. According to statistics, about 90 percent of international cargo trade (by weight) is carried by sea, most of which is packed in containers. Therefore, investigations on container supply chain have been absorbed much attention. As in other types of supply chain, demand forecast in container supply chains, *i.e.*, container throughput forecast, is also an important tool of information sharing and of much significance for many aspects. For examples, container throughput forecast can provide the sound background of transport capacity allocation, pricing strategy and port enterprise operation.

There are two flaws in previous studies on container throughput forecasting. First, the real data often include much noise and are frequently influenced by irregular events, *e.g.*, financial crisis, earthquakes, abrupt changes in authoritative policies. Consequently, under high volatile circumstance, many traditional forecasting models, which are dependent on solely historical structured data, have been criticized for their poor forecasting performance. The big data theory offers a great opportunity to solve this problem by taking more various types of data into consideration, expert knowledge and text information from Internet, to name but two.

Second, in the existing academic papers, traditional models, such as time series models [Bonham *et al.* (2009)][Chu (2008)], multivariate regression models [Guizzardi and Mazzocchi (2004)][Teyssier (2012)], fuzzy models [Chou *et al.* (2011)] and intelligent models [Heng *et al.* (2009)][Kotegawa *et al.* (2010)], are frequently employed. Under smooth economic environments, the above models can generate satisfactory forecasting performance. Nevertheless, under complex environment, no single model can offer precise forecasts, thus combined models are put into use, *e.g.*, [Totamane *et al.* (2014)] argued that combining different single models could result in higher forecasting accuracy. It is notable that combing single models in an unsuitable way cannot significantly improve forecasting performance, even worse, it will lead to larger errors [Yu *et al.* (2007)]. [Yu *et al.* (2007)] further mathematically proved that only those models with very weak correlation are suitable for constructing a combined model. However, forecasting models in reality are frequently impacted by the same events of strong influence, therefore their forecasts tend to take on very similar patterns and have significant correlation. In this circumstance, traditional combined models tend to fail.

This paper, taking advantage of big data, aims to propose a new container throughput forecasting framework, under which the forecasting model is expected to effectively solve the aforementioned problem.

The remainder of this paper is organized as follows. First, the framework of forecasting with big data using partially combination strategy is described in details. Then, operational steps of the new forecasting model are explained. Finally, an empirical study is conducted to validate the proposed model.

## 3. The Proposed Forecasting Framework

To solve the above mentioned problems, two principles of the newly proposed framework are advanced. First, big data, including both structured and unstructured data, should be used. Especially, expert knowledge and text information acquired by applying text mining technique, should be incorporated into forecasting models. Second, since high correlation between models in reality is resulted from the impact of the same significant event, decomposition methods can be employed to remove the impact and thus decrease the correlation between the models, then the new models with weak correlation can be combined for higher forecasting accuracy.

The principal idea of the proposed framework can explained as follows. Given an original time series $Y$, econometrical models are employed to predict linear component in $Y$, and select the best forecast $L$ by AIC and BIC criteria; The error $R = Y - L$ comprises the impacts from all kinds of special events and random terms, then the impacts from special events denoted by $IR$ can be captured by nonlinear models, and $R = Y - L$ represents the impacts of random term. If the nonlinear models are good enough to describe the impact of special events, $R-IR$ from different nonlinear models are samples drawn from the same population and thus they are independent identically distributed, therefore there are only weak correlation between them; Use suitable method to combine forecasts of the nonlinear models to generate the combined forecast denoted by $NL$; Integrate the linear component forecast $L$ and the nonlinear component forecast $NL$ to generate the final forecast with big data including expert knowledge and text information. Fig. 1 vividly presents the proposed framework and the notation in the figure is the same as the above.

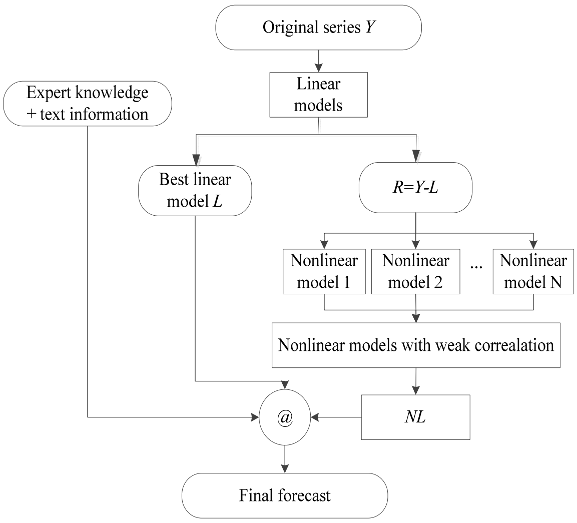

Fig. 1. The partially combined forecasting framework with big data.

As presented in Figure 1, two types of models, *i.e.*, linear and nonlinear, are used in the framework. This paper takes ARIMA as the linear model, LSSVM and GP as the nonlinear models. Under the new framework, the combined model is constructed by the following steps:

(1) Given a time series $Y$, use multiple ARIMA models to predict $Y$ and select the best forecast denoted by $L_f$ Lf by AIC and BIC criteria. $L_f$ is the linear component forecast in $Y$.

(2) Compute the forecasting error by $E=Y-L_f$. Use the LSSVM model and GP model to predict $E$ and the results are denoted by $E_{f1}$ and $E_{f2}$, respectively. Integrate $E_{f1}$ and $E_{f2}$ together and thus obtain $E_f$ that is the nonlinear component forecast in $Y$.

(3) Under the TEI@I methodology [Tian *et al*. (2009)], $L_f$ and $E_f$ together with expert knowledge and text information are integrated into the final forecast.

## 4. Empirical Study

### 4.1. *Data description and evaluation criteria*

This paper selects monthly container throughput of a seaport in China from Jan. 2004 to Sep. 2014 (See Fig. 2), which can be downloaded from the CEIC database. The motivation of selecting these data lies in the fact that in the period from Jan. 2004 to Sep. 2014, all kinds of special events and irregular events happened frequently, *e.g.*, the snowstorms in the south of China, the great earthquake in Sichuan province of China, the earthquake and tsunami in Japan, the subprime mortgage crisis in America, the European debt crisis, etc. This complex circumstance will frustrate the traditional models and provide a good opportunity to validate the newly proposed model.

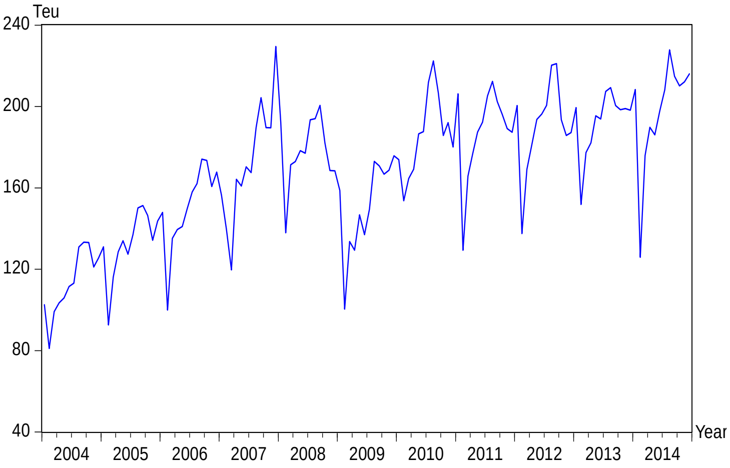

Fig. 2. Container throughput of a seaport of China.

Data from Jan. 2004 to Dec. 2013 are used as the training data set for parameter estimation and the remainder as the test data set for model evaluation. Three criteria are selected to evaluate forecasting performance of different models, *i.e.*, *RMSE* ;MAPE and TPE, and their mathematical expression can be written as:

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}\left(y_i - \widehat{y}_i\right)^2} \tag{1}$$

$$MAPE = \frac{\sum_{i=1}^{T}\left|y_i - \widehat{y}_i / y_i\right|}{T} \tag{2}$$

$$TPE = \left|1 - \frac{\sum_{i=1}^{T}\widehat{y}_i}{\sum_{i=1}^{T}y_i}\right| \tag{3}$$

### 4.2. *Linear component forecasting model*

Fig. 2 presents strong seasonality and outliers in the original series, therefore, X12 program is firstly employed to decompose the original series into seasonal component (SF), trend cycle (TC) component and irregular component (IR). SF represents changes of the time series in different seasons, TC represents the trend of the series in the long run (see Fig. 3). From Fig. 3, container throughput sees an abrupt decline and then a rebound in the period from Apr. 2008 to May. 2010, therefore, a quadratic equation is used to simulate this change.

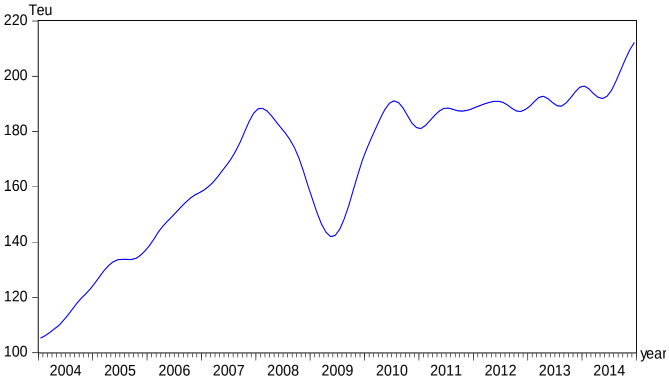

Fig. 3. The TC component of the original series.

Fig. 4 compares seasonality in different moths, which suggests the impacts of February, August, September and October are significantly stronger than in other months, therefore 4 dummy variables including Feb; Aug; Sept and Oct are constructed to capture the impact in the corresponding months.

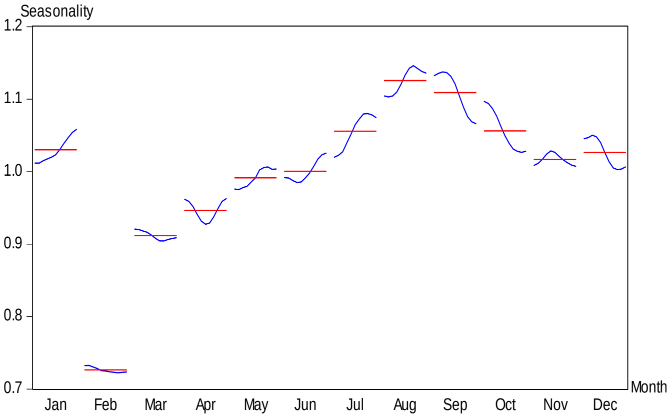

Fig. 4. Seasonality in different months.

Based on the above analysis, a linear regression model is constructed as follows:

$$y_t = a_0 + \alpha_1 Feb_t + \alpha_2 Aug_t + \alpha_3 Sep_t + \alpha_4 Oct_t + \alpha_5 t + \beta_1 ti_t^{\ 2} + \beta_2 ti_t + \beta_3 I_t + e_t \quad (4)$$

where $ti_t$ and $I_t$ are defined as:

$$t_t = \begin{cases} t, & 56 \le t \le 70 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

and

$$I_t = \begin{cases} 1, & 56 \le t \le 70 \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

Augmented Dickey-Fuller statistic is used for the unit root test on the error et. The *p*-value smaller than $10^{-4}$ means et is not a unit root process. From the auto-correlation and partial-correlation chart (Fig. 5), $e_t$ follows the stochastic process of

$$e_t = b_0 + b_1 e_{t-1} + b_2 e_{t-2} + \eta_t\,. \tag{7}$$

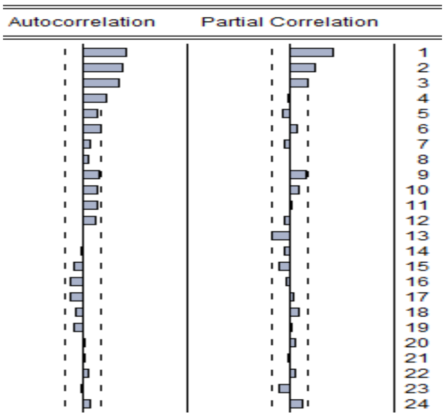

Fig. 5. Auto-correlation and partial-correlation of { $e_t$ }.

Substituting Eq. (7) into Eq. (4) generates Eq. (8)

$$\begin{aligned} y_t = {} & \alpha_0 + \alpha_1 Feb_t + \alpha_2 Aug_t + \alpha_3 Sep_t + \alpha_4 Oct_t + \alpha_5 t + \beta_1 ti_t^{\,2} \\ & + \beta_2 ti_t + \beta_3 I_t + b_1 ar(1) + b_2 ar(2) + \eta_t \end{aligned} \tag{8}$$

where $\alpha_0 = a_0 + b_0$ . Eq. (8) is the forecasting model of the linear component. The detailed estimation results are listed in Table 1.

Table 1. Parameter estimation of the linear component forecasting model.

| **variable** | **parameter** | **Std.** | ***t*-value** | ***p*-value** |
|---|---|---|---|---|
| c | 279.870 | 8.479 | 33.008 | 0.000 |
| *Jan* | -39.407 | 4.612 | -8.544 | 0.000 |
| *Feb* | -81.774 | 4.493 | -18.201 | 0.000 |
| *Oct* | 15.866 | 4.746 | 3.343 | 0.001 |
| *Nov* | 17.824 | 4.666 | 3.820 | 0.000 |
| $ti^2$ | 1.048 | 0.326 | 3.214 | 0.002 |
| $ti$ | -133.085 | 41.061 | -3.241 | 0.002 |
| *I* | 4166.515 | 1284.117 | 3.245 | 0.002 |
| $t$ | 0.676 | 0.120 | 5.648 | 0.000 |
| $ar(1)$ | 0.443 | 0.093 | 4.772 | 0.000 |
| $ar(2)$ | 0.217 | 0.092 | 2.342 | 0.021 |

Table 1. (Continued)

| | | | |
|---|---|---|---|
| $R^2$ | 0.880 | *Adjusted* $R^2$ | 0.868 |
| $F$ | 76.365 | $p$ value | 0.0000 |
| DW | 1.972 | | |

### 4.3. *Nonlinear component forecasting model*

In light of Table 1, the fitted values $L_1$ and the forecasted values $L_2$ of the linear component can be computed. Then the forecasting error $E_1$ can be obtained, which is the nonlinear component of the original series. Carry out the BDS test on $E_1$ and the results presented in Table 2 significantly suggest existence of nonlinear serial correlation, which means that nonlinear forecasting model should be utilized. This section takes advantage of two nonlinear models, *i.e.*, GP and LSSVM, to respectively predict $E_1$, and then integrate the results of two models into the integrated one.

Denote the fitted value of $E_1$ from GP and LSSVM by $E_{lssvm}^1$ and $E_{gp}^1$, depicted by Fig. 6. The fitting error of the two model can be computed by $R_{lssvm}^1$ and $R_{gp}^1$. The correlation between $R_{lssvm}^1$ and $R_{gp}^1$ is very weak, only 0.086, which implies that the GP model and the LSSVM model are suitable for integration. The ANN model is used to integrate the above two nonlinear models to generate the primary forecasted values. Then, to obtain the final forecasted values, expert knowledge and text knowledge is utilized to adjust the primary forecast values by adopting expert system technique. As presented in Fig. 7, the final forecasted values can satisfactorily fit the real ones.

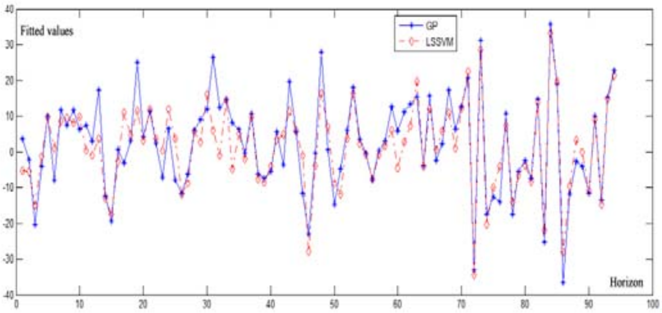

Fig. 6. Fitted values of GP and LSSVM

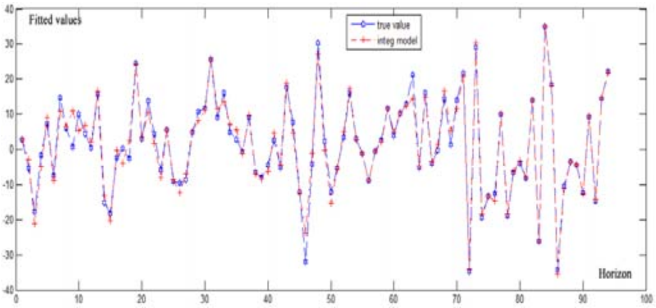

Fig. 7. Final integrated values VS. real data

### 4.4. *Forecasting performance comparison*

Taking advantage of the above-mentioned linear model (Eq. 8), the two nonlinear models (GP and LSSVM) and the integrated model (LSSVM_GP) , we can obtain linear component forecast and different nonlinear component forecasts from different nonlinear models. The final forecasted value of the original data will be computed as the sum of the linear component forecast and the best nonlinear.

In order to demonstrate advantage of the proposed framework over traditional models, this section uses LSSVM, GP, ANN to directly forecast the original data and then constructs a fully combined model composed of the 3 models. Forecasting performances of 4 models, including LSSVM, GP, ANN, LSSVM_GP_ANN and the partially combined model Par_integ, are compared. Fig. 8 vividly depicts the forecasted value of the 4 models, and the performance comparison results are listed in Table 3.

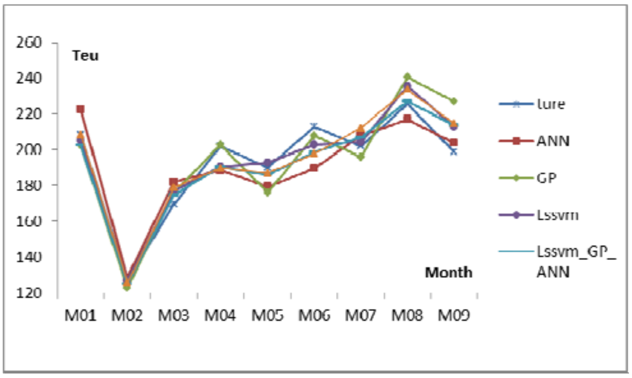

Fig. 8 Forecasted values of the 4 models

Table 3. Performance comparison of the 4 models

| Models | | *RMSE* | *MAPE* | *TPE* |
|---|---|---|---|---|
| Single models | LSSVM | 36.134 | 0.071 | 0.033 |
| | GP | 22.162 | 0.082 | 0.055 |
| | ANN | 23.211 | 0.085 | 0.075 |
| Fully combined model | LSSVM_GP_ANN | 20.693 | 0.051 | 0.045 |
| Partially combined model | Par_Integ | 10.172 | 0.030 | 0.018 |

In light of Table 3, LSSVM in the single models performs best on *TPE* but worst on *RMSE* ; GP is the best single model on both *RMSE* and *MAPE* ; the partially combined model (Par_Integ) significantly outperforms the others on all criteria; It is notable that although performance of the fully combined model (LSSVM_GP_ANN) is slightly better than the single models, it is significantly worse than the partially combined model (Par Integ). The main reason is that, there exist strong correlations between the LSSVM, GP and ANN models (see Table 4) owing to the same events of high importance, to name but two, the subprime mortgage crisis in America, the European debt crisis, thus direct combination cannot bring significant improvement of forecasting performance. By contrast, the partially combined model decomposed the linear component and just combined the nonlinear models with weak correlation (the correlation coefficient is only 0.072 ), consequently can effectively improve forecasting performance.

Table 4. Correlation between 3 single models

| | LSSVM | GP | ANN |
|---|---|---|---|
| LSSVM | 1 | | |
| GP | 0.400 | 1 | |
| ANN | 0.532 | 0.768 | 1 |

## 5. Conclusion

This paper proposes a partially combined forecasting framework with big data. The main difference between the model under this framework and others can be described as follows:

(1) Traditional models construct multiple single forecasting models and then directly combine these models into an unified one. However, when facing high volatility, single models frequently tend be impacted by the same events of strong influence such as earthquake and financial crisis, which will lead to high correlation between the single models. Under this circumstance, direct combination cannot generate significant improvement of forecasting performance.

(2) Under the proposed framework, the original series is firstly decomposed into the linear component representing the long-run trend and the nonlinear component representing the impact of irregular events and stochastic terms. Then various nonlinear models are used to capture the impacts of irregular events and only those models with weak correlation are kept for combination. In order to overcome the weakness of traditional models heavily dependent on historical data, big data including expert knowledge and text information is exploited by expert system.

The empirical study of forecasting container throughput of a seaport in China was carried out to validate the proposed forecasting framework, and the results showed significant superiority of the model under the proposed framework over others.

## Acknowledgments

This research is supported by the Fundamental Research Funds for the Central Universities (2015JBM125), the EC-China project Use of ICT to enable the integration of CSCs (B14F00010), the NSFC project (71390334, 71132008) and Shandong Independent Innovation and Achievement Transformation Special Fund of China (2014ZZCX03302).

## References

Gunasekaran, A. (2004). *Supply chain management: Theory and applications*. European Journal of Operational Research, **159**(2): 265–268.

Premkumar, G.P. (2000). Inter-organization systems and supply chain management: An information processing perspective. Information Systems Management, **17** (3): 56–69.

Davis, E.W., and Spekman, R. (2004). *Extended Enterprise*. Prentice-Hall, Upper Saddle River, NJ.

Watson, G. (2001). *Sub-regimes of power and integrated supply chain management*. Journal of Supply Chain Management, **37**(2): 36–41.

Thonemann, U. W. (2002). *Improving supply chain Performance by sharing advance demand Information*. European Journal of Operational Research, **142**(1): 81–107.

Heikkilä, J. (2002). From supply to demand chain management: Efficiency and customer satisfaction. Journal of Operations Management, **20** (6): 747–767.

Aviv, Y. (2001). The effect of collaborative forecasting on supply chain performance. Management Science, **47** (10): 1326–1343.

Cachon, G.P., and Lariviere, M.A. (2001). *Contracting to assure supply: How to share demand forecasts in a supply chain*. Management Science, **47** (5): 629–646.

Yue, X., and Liu, J. (2006). *Demand forecast sharing in a dual-channel supply channel*. European Journal of Operational Research, **174**: 646–667.

Carbonneau, R., Laframboise, K., and Vahidov, R. (2008). *Application of machine learning techniques for supply chain demand forecasting*. European Journal of Operational Research, **184**(3): 1140–1154.

Bonham, C., Gangnes, B., and Zhou, T. (2009). *Modeling tourism: A fully identified VECM approach*. International Journal of Forecasting, **25**(3): 531–549.

Chu, F.L. (2008). *Analyzing and forecasting tourism demand with ARAR algorithm*. Tourism Management, **29**(6): 1185–1196.

Guizzardi, A., and Mazzocchi, M. (2010). *Tourism demand for Italy and the business cycle.* Tourism Management, **31**(3): 367–377.

Teyssier, N. (2012). *How the consumer confidence index could increase air travel demandforecast accuracy*? Cranfield University, Cranfield.

Chou, T.Y., Liang, G.S., and Han, T.C. (2011). *Application of fuzzy regression on air cargo volume forecast*. Quality & Quantity, **45**(6): 1539–1550.

Heng, H.j., Zheng, B.Z., and Li, Y.j. (2009). *Study of SVM-based air-cargo demand forecast model.* Proceedings of the 2009International Conference on Computational Intelligence and Security (CIS'09), Beijing.

Kotegawa, T., DeLaurentis, D. A., and Sengstacken, A. (2010). *Development of network restructuring models for improved air traffic forecasts*. Transportation Research Part C: Emerging Technologies, 18(6): 937–949.

Totamane, R., Dasgupta, A., and Rao, S. (2014). *Air Cargo Demand Modeling and Prediction*. IEEE Systems Journal, 8(1): 52–62.

Yu, L. A., Wang, S. Y., and Lai, K. K. (2007). *Foreign Exchange Rate Forecasting with Artificial Neural Networks.* Springer, New York.

Tian, X., Lu, X., and Deng, X. (2009). *A TEI@I-Based Integrated Framework for Port Logistics Forecasting*. Proceedings of the International Conference on Business Intelligence and Financial Engineering(BIFE 09), Beijing.