Transportation Research Part E 108 (2017) 160–178

Contents lists available at ScienceDirect

Transportation Research Part E

journal homepage: www.elsevier.com/locate/tre

Data characteristic analysis and model selection for container throughput forecasting within a decomposition-ensemble methodology

Gang Xie *, Ning Zhang, Shouyang Wang

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

A R T I C L E I N F O

Article history:
Received 12 January 2017
Received in revised form 17 August 2017
Accepted 30 August 2017
Available online 1 November 2017

Keywords:
Container throughput
Data characteristic analysis
Model selection
Time series forecasting
Decomposition-ensemble methodology

A B S T R A C T

In this study, a novel decomposition-ensemble methodology is proposed for container throughput forecasting. Firstly, the sample data of container throughput at ports are decomposed into several components. Secondly, the time series of the various components are thoroughly investigated to accurately capture the data characteristics. Then, an individual forecasting model is selected for each component based on the data characteristic analysis (DCA). Finally, the forecasting results are combined as an aggregated output. An empirical analysis is implemented for illustration and verification purposes. Our results suggest that proposed hybrid models can achieve better performance than other methods.

1. Introduction

With the current development of economic globalization, ports have become more and more important in the operation of international trade activities (Notteboom, 2016; Yap and Lam, 2006). A port is characterized by a functional and spatial clustering of cargo transport, storage and transformation processes, all of which are linked to global supply chains (Twrdy and Batista, 2016). The handling of maritime cargo at specialized terminals remains a core function of ports (de Gooijer and Klein, 1989; Steenken et al., 2004; Coto-Millán et al., 2005). As many manufacturing industries have experienced rapid growth, particularly in emerging economies, many coastal cities in those economic centers have invested heavily in ports (Jeevan et al., 2015). The prediction of container throughput at a port helps port managers make not only strategic decisions on developing the port in terms of scale, general layout and district division, but also helps with tactical and operational decisions (mid-term and short term decisions) such as operations planning decisions within the port, the scheduling of port equipment, etc. Accurate predictions of container throughput mean that investments in port capacity and other transportation infrastructure can be made consistent with the needs generated by that traffic (Lam et al., 2004; Levine et al., 2009; Petering, 2009). If these throughput predictions are not accurate enough, policy bias will occur, which could cause huge financial losses. Therefore, developing an effective container throughput forecasting model has become a crucial task (Geng et al., 2015).

Containerization is an important element of the logistics and security innovations that revolutionized freight handling in the 20th century. The pattern characteristics of container throughput time series include cyclicity, seasonality, mutability and randomicity. These traits are determined by the economic structure and market development of the port's hinterland

* Corresponding author. Tel. +86-10-82541368; fax: +86-10-62541823.
E-mail addresses: gxie@amss.ac.cn (G. Xie), qinian1008@163.com (N. Zhang), sywang@amss.ac.cn (S. Wang).

https://doi.org/10.1016/j.tre.2017.08.015
1366-5545/

(Tian et al., 2013). Currently, the length of the available time series is from 10 to 30 years; the time granularity is one month or one year. In practice, there are distinct seasonal characteristics of container throughput at ports (Chou et al., 2003; Peng, 2006). For example, the Chinese New Year has a significant impact on container throughput at all Chinese ports (Liang and Chou, 2003; Chen and Chen, 2010). In addition, the time series of container throughput at ports may fluctuate due to specific events, including, for example, the financial crisis of 2008, large-scale manufacturing transfers, a dock workers' strike, etc. Any or all of these (or similar) events may result in breakpoints within the time series. As a consequence, the intrinsic complexity and volatility of the global economy and trade is what causes container throughput series to appear nonlinear and non-stationary. Container throughput forecasting is also a very complex and dynamic process. In this study, we further analyze the data characteristic of complexity, which covers various nonlinear characteristics, e.g., chaoticity, fractality, irregularity and long-range memorability (Tang et al., 2014; Yu et al., 2015). To achieve better forecasting performance, conducting a data characteristics analysis (DCA) is a useful exercise to undertake before attempting to predict the container throughput at any given port.

Some econometric models have previously been used for container throughput forecasting. Those are the error correction model (ECM) (Fung, 2002; Hui et al., 2004), the multivariate regression model (Chou et al., 2008; Veenstra and Haralambides, 2001), and the seasonal autoregressive integrated moving average (SARIMA) model (Schulze and Prinz, 2009), as well as other models, for example, exponential smoothing (Liu et al., 2010) and the Vector Auto-Regression (VAR) model (Tian et al., 2013). However, econometric models are built on linear assumptions. As such, they cannot capture the nonlinear patterns hidden in the original data, which in turn leads to poor forecasting performance, especially with regard to some time series with nonlinearity. For the description of nonlinear characteristics in the time series of container throughput at ports, artificial intelligence (AI) models have been used. These AI models include a back-propagation neural network (BPNN) model (Liu et al., 2010), a genetic programming (GP) model (Chen and Chen, 2010) and a transfer forecasting model guided by a discrete particle swarm optimization (TF-DPSO) (Xiao et al., 2014) algorithm.

For better forecasting performance, many hybrid approaches have been developed. Combining projection pursuit regression (PPR) with a genetic programming (GP) algorithm, Huang et al. (2015) proposed a hybrid method to forecast the container throughput of Qingdao Port. To eliminate the influence of outliers, Huang extended a local outlier factor (LOF) as a means to detect outliers in the time series. Then, different dummy variables were constructed to capture the effect of outliers, based on domain knowledge. The results of Huang's research show that the proposed method significantly outperforms artificial neural network (ANN), SARIMA, and PPR models. In addition, decomposition has been used to develop hybrid approaches for forecasting container throughput. Peng and Chu (2009) presented the classical decomposition (CD) model, which decomposes the time series of container throughput at ports into four separate components, namely trend, cyclical, seasonal and irregular factors. In the CD approach, the least square method is applied, in order to derive a linear regression equation for the estimate of the trend component. However, the CD model is also built on linear assumptions. As such, the CD model cannot capture the nonlinear patterns hidden in the original container throughput time series. Consequently, Xie et al. (2013) proposed several hybrid approaches based on the least squares support vector regression (LSSVR) model and preprocessed methods, including SARIMA, seasonal decomposition and classical decomposition. The empirical analysis presented in Xie et al. (2013) concluded that the proposed hybrid approaches can achieve better forecasting performance than individual approaches. For good forecasting performance, it is important to describe the seasonal nature and nonlinear characteristics of container throughput series, which can be realized efficiently by employing decomposition and the "divide and conquer" principle (Xie et al., 2013). However, in the above-named studies, DCA has not yet been implemented to the time series of the components before models are developed for the purpose of container throughput forecasting.

Existing studies merely consider certain given data characteristics of time series. For example, Peng and Chu (2009) investigated the seasonality of container throughput at ports, and Xie et al. (2013) took the mutability of a time series into account, while other important data characteristics were ignored in the modeling process in these studies. Therefore, it is necessary to conduct DCA before a container throughput forecasting model is developed. As such, we propose a forecasting procedure within a decomposition-ensemble methodology. To the best of our knowledge, the application of DCA for container throughput forecasting has not yet been studied or published in any existing literature.

In this study, we propose a novel DCA-based decomposition-ensemble methodology. With our methodology, a number of hybrid approaches are developed for container throughput forecasting. Firstly, the sample data of container throughput at ports are decomposed into several components. Secondly, the time series of the components are thoroughly investigated, in order to capture the data characteristics (including stationarity, seasonality, mutability and complexity). Then, a single model is selected for the prediction of each component based on the DCA results. In particular, when the time series of the components have the characteristics of being nonstationary or complexity, artificial intelligence (AI) models are accordingly developed. Finally, the forecasting results (in terms of the decomposed subtasks) are combined as an aggregated output. An empirical analysis is implemented for illustration and verification purposes. Finally, some related issues are discussed, and our conclusions are drawn and presented.

The remainder of this paper is organized as follows: The proposed novel decomposition-ensemble methodology is described in detail in Section 2. Here, the overall process of the proposed methodology, decomposition methods, data characteristic analysis (DCA), and forecasting modeling are thoroughly discussed. Section 3 illustrates the problems related to the forecasting of container throughput by using an empirical analysis with experiments. Then, a number of related issues are discussed in Section 4. Section 5 presents our conclusions and suggests some directions for future investigation.