Transportation Research Part C 39 (2014) 80–93

Contents lists available at ScienceDirect

Transportation Research Part C

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

A self-learning advanced booking model for railway arrival forecasting

Tsung-Hsien Tsai *

Department of Tourism Management, National Quemoy University, 1 University Rd., Jinning Township, Kinmen 892, Taiwan, ROC

ARTICLE INFO

Article history:
Received 15 August 2013
Received in revised form 21 November 2013
Accepted 21 November 2013

Keywords:
Advanced booking model
Booking curve
Case-based reasoning
Arrival forecasting
Revenue management

ABSTRACT

Accurate short-term arrival forecasting is essential information for railway operators to conduct daily operations such as demand management strategies. Conventional time series methods apply historical arrival data which is the accumulation of reservations to project future arrivals. This study aims to utilize reservation data directly and proposes a novel advanced booking model by using the framework of case-based reasoning. The proposed model contains four modules with distinctive functions for similarity evaluation, instance selection, arrival projection, and parameter search. We have the constructed model tested on fourteen daily arrival series and compared its out-of-sample accuracy with that of four traditional benchmarks. The empirical results show that in average the proposed self-learning model may reduce at least 11% of mean square errors (MSE). Moreover, the learning scheme in the model may achieve significant reduction of MSE comparing with performance of other naïve versions.

1. Introduction

Forecasting is no doubt an important skill that railway operators strive to possess current days in which uncertainty and volatility are becoming more and more common circumstances. In order to utilize perishable resources effectively in the dynamic environment, the concept of revenue management (RM) has been proven to be a successful application in the service and transportation industries (Chiang et al., 2007). Kimes (2005) has confirmed the value of RM by showing that the deployment of RM may increase 3–5% revenues in hotel, rental car and airline industries.

The core of RM integrates the functions of forecasting, resource allocation, pricing and overbooking and aim to sell the right products to right customers with right prices at right time (Smith et al., 1992) for attaining maximized revenues. Prediction is responsible for providing accurate inputs in the RM system and may enable operators to have their perishable seats allocated appropriately to avoid either vacancy at the departure day or selling seats to passengers with low willingness-to-pay. Lee (1990) indicates that 10% increase in predictive accuracy can result in 0.5–3% increase in revenue. Zaki (2000) and Weatherford and Belobaba (2002) also support the importance of RM forecasting by providing similar evidences.

Understanding upcoming daily demand helps operators determine the number of seats available for implemented ticket types. In order to do so, railway operators have to frequently track booking information and project future arrivals based on historical data and also external influences. In the literature, traditional seat allocation models desire probability distribution of demand and assumptions such as products are sold in a low-before-high manner or the fare structure (Littlewood, 2005; Belobaba, 1989; Robinson, 1995; Talluri and van Ryzin, 2004). Recently, researchers propose novel concepts using no forecasting and optimization process (van Ryzin and McGill, 2000; Kunnumkal and Topaloglu, 2007) or limited demand infor-

* Tel.: +886 82 313039; fax: +886 82 313086.
E-mail address: thtsai@nqu.edu.tw

0968-090X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2013.11.021

mation (Lan et al., 2008) for determining booking limits. For example, van Ryzin and McGill (2000) propose an adjustment method to determine the optimal protection levels based on RM performance in previous departures. Lan et al. (2008) utilize only estimated upper and lower demand bounds rather than complete arrival information. This new stream of concept provides alternatives for deploying RM applications. However, it is an unanswered question when these novel revenue management procedures can outperform traditional processes or vice versa in what situations. In fact, most RM practices still regard forecasting as a vital step for determining optimal booking limits (Chiang et al., 2007; Bobb and Veral, 2008).

Within the literature of model structure on arrival forecasting, most of them apply historical arrival data for model construction. Comparatively little research has focused on utilizing booking patterns to construct forecasting models. The last two decades have seen the growing importance of building reservation systems in the service and hospitality industries. These reservation systems store clients' requests and record their booking information including when they book, when they use the service, when they cancel, how they pay, so on and so forth. As a result, the aim of this study is to propose a forecasting procedure based on the framework of case-based reasoning for effectively utilizing reservation data to construct arrival forecasting model.

The next section first reviews related works in terms of forecasting tasks and also model structures. The third section introduces the proposed model and the forecasting procedure. The form of booking curves is first briefed and four modules with respective functions are described. Four benchmarks are also stated for the comparing purposes. An empirical study in the third section shows out-of-sample performance of the proposed model and four benchmarks to verify validity. Finally, conclusions are presented and suggestions are rendered for future research.

2. Literature review

2.1. Forecasting issues

In practice, improvement of a forecasting procedure can be classified into several major dimensions including model selection, data editing and identification, instance selection, input selection, novel model structure, and model integration (Tsai et al., 2008). The issue of model selection centers on how to select a suitable model for the task. In the field of short-term traffic forecasting, framework of time series modeling is the most prevailing concept. ARIMA, exponential smoothing, non-parametric regression, artificial neural networks, support vector machine are all potential methods. However, there are no such universal rules for model selection in the literature. Data characteristics, in fact, play a significant role while selecting a suitable methodology for prediction. Some researchers have provided general principles (Armstrong, 2001) and others compare the differences and similarities among different methods (Karlaftis and Vlahogianni, 2011) which may shed some lights on the adoption of methodologies.

Data editing focuses on the treatment of missing data and also the recognition of data characteristics for achieving promising performance. Related works are like Li et al. (2013), Tan et al. (2013), Zhong et al. (2004); the aforementioned studies all propose new models to impute missing data in the traffic flow. On the studies of recognizing traffic data characteristics, Vlahogianni et al. (2006) apply statistical methods for identifying nonlinearity and non-stationary of univariate time series data. Understanding data features is important since it is helpful for model selection, input selection, and the structure of forecasting models.

Selection of instances for constructing models is another research topic. Conventional statistical models apply all available samples for model construction; however, some researchers start concentrating on the selection of most relevant samples such as Morantz et al. (2004), and Zhang and Qi (2005). This stream of research is getting more important since nowadays electronic storage systems are affordable resulting in the collection of abundant (mega) data.

Choosing the most relevant factors to explain the traffic flow and maintain parsimony simultaneously is another vital issue. This topic is important especially for time series forecasting since lag variables can be ample. In addition, spatial–temporal correlations in the traffic flow data are obvious which also creates a situation to include more inputs. For instance, Stathopoulos and Karlaftis (2003) have tested the use of upstream flow information to predict the downstream flows. Yang (2013) also proposes a new input selection method for traffic congestion prediction.

The development of novel models continuously attracts researchers' attention in the literature and potential alternatives have been proposed. Huang and Sadek (2009) propose a spinning network by observing the operation of human memory for forecasting traffic volumes. Fei et al. (2011) utilize Bayesian dynamic linear models to predict on-line short-term travel time with confidence intervals on freeway. Recently, Antoniou et al. (2013) exploit a cluster-based data-driven computational approach to predict local traffic state. All aforementioned models focus on the innovation of model structures and have capabilities to upgrade performance as a result of appropriately explaining the flow or travel time data.

Last but not least, since forecasting is a difficult task to tackle, some studies divide the whole forecasting task into several sub-tasks and apply suitable models for solving sub-tasks and form so-called hybrid models. Related works are like Wang and Shi (2013) who have integrated the advantages of wavelet and support vector machine for traffic speed forecasting. Vlahogianni et al. (2005) use the genetic algorithm for determining the topology of structure while applying neural networks. Dimitriou et al. (2008) incorporate the fuzzy theory in a rule-based forecasting system. Wei and Chen (2012) apply the decomposed theory to obtain several components which are forecasted by respective neural network models.