Neurocomputing 158 (2015) 210–215

Contents lists available at ScienceDirect

Neurocomputing

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

Brief Papers

Prediction models based on multivariate statistical methods and their applications for predicting railway freight volume

Yandong Yang, Congzhou Yu *

College of Education and P. E., Bohai University, Liaoning 121013, China

ARTICLE INFO

Article history:
Received 30 November 2014
Received in revised form
6 January 2015
Accepted 25 January 2015
Communicated by Yang Tang
Available online 10 February 2015

Keywords:
Prediction modell
Multivariate statistical methods
Partial least squares
Railway freight volume

ABSTRACT

Four prediction models based on the multivariate statistical methods are constructed in this work and they are successfully applied in predicting the Railway Freight Volume (RFV). RFV directly reflects the regional economic states such as production improvement and economic restructuring. Accurately predicting the RFV is of great use in production planning, decision making, labor allocating, etc. In this work, based on the multivariate statistical methods, i.e. ordinary least squares regression (OLSR), principal component regression (PCR), partial least squares regression (PLSR), and modified partial least squares regression (MPLSR), four RFV prediction models are constructed and the detailed comparison is made by implementing them on a practical dataset. From the simulation results, the conclusion can be derived that the MPLSR based prediction model outperforms the other three models.

1. Introduction

With the rapid development of economic globalization and regional economic integration, the logistics industry has entered a new stage of a rational, pragmatic and rapid development [1]. The renascent logistics industry is a service industry and provides great support to other industries. One of the important factors of affecting regional logistics is the economic development in one region. Regional economic development is multi-level, multi-factor and complex system [2,3]. The demand of logistics is just a reflection of the integrated system. It is also an important factor of deciding the development policy of logistics industry [4,5]. However, it is complicated and it cannot exactly describe the quantitative relationship. Fortunately, useful information can be tapped from time series data in the stable situation. A quantitative analysis mathematical model can be established by seeking the variation of logistics. Then the logistics can be predicted. Through the analysis of the logistics, the standard living of a region can be obtained [6,7]. It will bring great help to planning of railway departments, through predicting logistics with current data. Rail freight demand forecast refers to the rail freight market research. After analysis, the scientific methods are used to estimate the future demand for freight and variation tendency. It provides a scientific basis for policy formulation and improving law regulations. In order to plan reasonably or avoid wasting energy, an accurate prediction of railway freight is needed.

In recent years, many scholars have studied railway freight volume forecast in many different methods. In Ref. [8], rough sets theory was used to predict the demand for rail freight. Moreover, it is concluded that the data can be directly used when using rough set. This method proposed in [8] avoids the subjective factors and improves the reliability and validity of the forecast [8]. In Ref. [9], a study of the combination of rail freight forecasting methods is conducted and satisfactory results are shown. The theory of grey system is also introduced in Ref. [10] and it is based on railway freight volume data of Sichuan province. Then, DGM (2, 1) prediction model of Sichuan railway freight volume is built and the predictive values are tested. It is demonstrated that the model is of high precision and it can be used for freight volume forecast conclusion. At last, based on the algorithm, the railway freight volume of Sichuan province from 2010 to 2015 are predicted [10]. In Ref. [11], the method which based on improved BP neural network model is used to forecast railway freight volume. This method is proposed for the shortage of standard BP neural network based on the analysis of railway freight volume forecasting methods. The improved BP algorithm can get better solution, and can also shorten the training time. The method based on GA-BP model can be proposed to resolve the problem of railway freight volume forecasting in Ref. [12]. Reference [12] introduces the gray correlation analysis to determine the correlation of the national rail freight and its main impact factors, then GA-BP neural network prediction model is established according to their correlation factors associated with above standard values. And the feasibility and effectiveness of this method are proved.

By analyzing these study, it is found that the models established by our predecessors exist some errors and the case study is also not perfect. For example, grey model method can predict roughly the change trend of freight volume over a period of time, but the

* Corresponding author.
E-mail addresses: yandong.yang2014@gmail.com (Y. Yang), congzhou.yu2014@gmail.com (C. Yu).

http://dx.doi.org/10.1016/j.neucom.2015.01.046
0925-2312/

uncertainty factors, such as environmental factors and emergency, have a large influence on factors. And BP neural network exists the drawback, such as low convergence speed, falling into local minimum easily and network training sensitively for the initial weights and thresholds. It will produce strong volatility, so the accuracy is not high for the short-term freight volume forecasting. However, in order to forecast the regional logistics demand, choosing the most suitable method is the key to obtain high-precision forecast results. Four algorithms will be proposed to overcome the drawbacks of the aforementioned approaches in this paper. The four algorithms are OLSR, PCR, PLSR and MPLSR [13]. And these algorithms can avoid the disaster of dimensions effectively. At the same time, they can eliminate the influence of the multicollinearity in regression models. They can work on huge amount of data to complete the prediction of railway freight volume. The basic idea of the four algorithms is to represent the high dimensional data set by a smaller set of latent variables. Then they can reduce computational load.

In this paper, the reliability of the four algorithms will be tested according to the data set of rail freight. Jiangxi province is located in the middle of China. The logistics activity plays a key role to link the west, the north and the south part of China. With the domestic logistics heating up, Jiangxi province regard speeding up the development of modern logistics as a strategy option for adjustment of economic efficiency in Jiangxi province. As the eleventh five year plan goes on, GDP and social total consumption have been increasing in Jiangxi province due to the implementation of the Open Strategy. Therefore, it is an important topic to predict the demand of railway logistics so that logistics of Jiangxi province can be planned based on it. In order to predict logistics of Jiangxi province, the multivariate methods will be used. This algorithm can help to predict logistics and build the fitted model. In order to obtain an efficient prediction model, a comparison amongst the models established by ordinary least squares regression, principal component regression [14], partial least squares regression, and a modified partial least square regression is made. And the best model is selected out by the analysis of error of the various forecasting methods. Then the advantage of MPLSR algorithm can be shown by comparison with the other three algorithms. MPLSR has a simpler math description than others. The method is simple in computation, high in reliability and has wide applications. Now MPLSR is used in framework of data-driven approaches [15–17] and system control [18,17], and the best result can be obtained [19]. Here MPLSR is used to forecast railway freight volume in the next five years, and provides suggestions for the future of scientific development.

2. Algorithms introduction

Multivariate statistics is a powerful tool to solve problems of railway logistics forecast, which often includes a large number of data. Multivariate statistics has been extensively applied to many sphere, such as predicting railway logistics by PLS regression [20,21], finding the main factors which influences the freight by PCA [22] and PLS.

For this part, in addition to describing three classical multivariate statistics methods, a modified PLS regression method is also imported, which has a simpler math description compared with standard PLS regression.

2.1. Ordinary least squares regression (OLSR)

Ordinary least squares regression (OLSR) aims to establish the optimal method of mathematical relationships according to the principle of minimum squared error of measure data. OLSR can easily obtain the unknown data. In most cases, OLSR appears in the lower-dimensional case. But, it can also settle the problem generated by multi-dimensional. The following is a brief introduction to the algorithm.

- *Step* 1: Standardize the collected measurable variables and response variables, denoted as $X=[x_1 \cdots x_m] \in R^{n\times m}$, $Y=[y_1 \cdots y_p] \in R^{n\times p}$.
- *Step* 2: Minimize the sum of errors e_i:

$$e_i = y_i - X\beta_i \tag{1}$$

$$\beta_i = \arg\min e_i^T e_i = \arg\min (y_i - X\beta_i)^T (y_i - X\beta_i) \quad i = 1, \ldots, p \tag{2}$$

Because of the invertibility of X^{TX}, the variable β_i can also be calculated by

$$\beta_i = (X^T X)^{-1} X^T y_i \tag{3}$$

- *Step* 3: Establish the regression model:

$$Y = XP + Q \tag{4}$$

in which P is formed by β_i, Q is formed by e_i.

2.2. Principal component regression (PCR)

During multiple linear regression analysis, we often encounter phenomenon of approximate linear relationship between the independent variable. This phenomenon is the so-called collinearity. When collinearity is severe, the regression model established by the least squares method will increase the variance parameters. Making the regression equation becomes difficult. In order to overcome this problem, Hotelling proposed principal component analysis in 1933. The core idea of principal component analysis is to reduce the dimension by regarding a few indicators as the comprehensive index, and try not to change the indicator system to explain the extent of the dependent variable. According to the idea of principal component analysis, W. F. Massy proposed principal component regression in 1965 [14]. Today, PCR has been widely used. And it become a effective method to resolve multicollinearity problem, and the algorithm can be briefly formulated as follows.

- *Step* 1: Standardize the original data. Presented as $X=[x_1 \cdots x_m]^T \in R^{n\times m}$, $Y=[y_1 \cdots y_p]^T \in R^{n\times p}$.
- *Step* 2: Calculate eigenvalues and the corresponding orthonormal eigenvectors of covariance matrix $X^T X$:

$$\frac{1}{n} X^T X = T \Lambda T^T \tag{5}$$

$$\Lambda = \mathrm{diag}(\lambda_1 \cdots \lambda_m), \quad \lambda_1 > \cdots > \lambda_m > 0$$

In which n is the dimension of covariance matrix $X^T X$, Λ is the diagonal matrix, λ_i is the eigenvalues of covariance matrix $X^T X$, T is the so-called orthogonal matrix, and it contains eigenvectors corresponding to eigenvalues.
- *Step* 3: Select k principal components. If the sum of eigenvalues ratio reaches a certain level, such as 85 percent, k is the number of the principal. Then these k components can replace all the independent variables, and calculate the score matrix Z_k.

$$Z_k = X T_{pc} \tag{6}$$

where $T_{pc} = [t_1 \ldots t_k] \in R^{n\times k}$, k is the number of the principal components.
- *Step* 4: Establish prediction model Y:

$$\beta_i = (Z_k^T Z_k)^{-1} Z_k^T y_i, \quad B = [\beta_1 \ldots \beta_p]^T \in R^{m\times p} \tag{7}$$

$$Y = Z_k B + E = XP + E, \quad P = T_{pc} B \tag{8}$$