JOURNAL OF TRANSPORTATION
SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
Volume 12, Issue 4, August 2012
Online English edition of the Chinese language journal

Available online at www.sciencedirect.com
ScienceDirect

Cite this article as: J Transpn Sys Eng & IT, 2012, 12(4), 128−134.

RESEARCH PAPER

Data Mining on Index of Static Load of Freight Cars

LIU Zhili[1,3,*], ZHOU Xilin[2], FANG Qigen[2]

1 MOE Key Laboratory for Urban Transportation Complex Systems Theory & Technology, Beijing Jiaotong University, Beijing 100044, China
2 China Railway Siyuan Survey and Design Group Co., Ltd., Wuhan 430063, China
3 Integrated Transportation Research Center of China, Beijing Jiaotong University, Beijing 100044, China

Abstract: Static load of freight vehicles is one of the important indices to evaluate the efficiency of railway wagon usage, as well an important basis for predicting coefficient of static load in railway line design. Based on sorting out the data structure of static load, clustering algorithm of K-medorids is employed for data mining to find the significant and relatively stable classification characteristics between static load of railway administrations. It is different delivery proportions of heavy goods that are indicated by index factor analysis to cause the static load classification of railway administrations. The data-mining also indicates the differences of static load between different grading railway administrations which cannot be ignored, therefore determining the coefficient of static load in railway design based on divided region is recommended.

Key Words: railway transportation; static load of freight cars; clustering algorithm; index factor analysis

1 Introduction

As an efficiency indicator for railroad freight transportation, Per Truck Static Load (PTSL)[1] is designed as the average weight of goods in one railway freight car when goods are uploaded into rail wagons at the origin station, measured by ton. For PTSL, there are three major statistical meanings. First of all, it is one of the most significant statistical indicators for railroad ministry and administrations to measure the efficiency of wagon utilities. To evaluate whether the wagon is wisely and economically used by employing the index of PTSL can encourage the personnel to increase per truck utilities rationally, which contributes to effective and efficient use of the current rail line transportation capacity. The second, PTSL is also directly relevant to the economic benefits for the goods owner, the higher PTSL the lower freight payment. The last but not the least, PTSL is used to predict PTSL coefficient for future railway line design to convert goods traffic flow into wagon flow. In order to figure out the development tendency of PTSL, we use clustering algorithm to mining the rule of PTSL from historical data.

2 Data structure of PTSL

Since the establishment of railway administrations, the categories of delivery goods vary during different historical periods as well as the PTSLs of different kinds of goods do. The records of PTSL statistical tables in Compilation of the National Railway Statistic include the PTSLs of railway ministry and each railway administration. Each record then includes the average PTSL and category-based PTSL for railway ministry or one railway administration. Our data of PTSL are collected from Compilation of the National Railway Statistics from Year 1985 to 2008. During these twenty-four years, as table 1 shows there are totally 8 times of changes in the data structure of PTSL[2] since the year 1985.

The tremendous change of data structure on categories of goods occurred in 1993 when the numbers of categories of goods increased from 22 to 26 as shown in Table 1. In 2001there were 3 slight name changes in category with Electronics and electricity replaced by Electronics and electrical machinery, Textiles and leather products replaced by Textiles, leather, furs and related products, and Others replaced by Other goods. Compared to the change frequency of goods category, the change of numbers of railway administrations were more frequent during 1985–2008. The relative stable years of both the railway administrations and the goods categories were 1986–1992, 1997–2000, and 2005–2008. For the goods categories in 1986-1992 had

Received date: Mar 12, 2012; Revised date: May 11, 2012; Accepted date: May 18, 2012
*Corresponding author. E-mail: liuzl@bjtu.edu.cn

DOI: 10.1016/S1570-6672(11)60217-1

Table 1 Data structure of PTSL during 1985–2008

No.	Year	No. of RAs	Name list of RAs	No. of categories	Name list of goods categories
1	1985	13	Harbin, Shenyang, Beijing, Hohhot, Zhengzhou, Jinan, Shanghai, Guangzhou, Liuzhou, Chengdu, Lanzhou, Urumqi, Kunming	22	Coal, Coke, Petroleum oil, Iron and steel, Metal ores, Non- metal ores, Mineral materials for construction, Cement, Lumber, Fertilizer and pesticides, Grains and grain products, Cotton, Salt, Household chemicals, Raw chemical materials and their end products, Industrial machinery, Agricultural machinery, Agricultural products, Live & fresh goods or perishable goods, LTL, Phosphate rock, and Others
2	1986–1992	12	RA of Kunming dismissed	22	Ditto
3	1993–1995	12	Ditto	28	Coal, Petroleum oil, Coke, Metal ores, Steel and non-ferrous metals, Non-metallic ores, Phosphate rock, Mineral building materials, Cement, Lumber, Grain, Cotton, Fertilizer and pesticides, Salt, Chemicals, Metal products, Industrial machinery, Electronics and electricity, Agricultural equipment, Fresh goods, Agricultural products, Food products and tobacco, Textiles and leather products, Paper and stationery, Pharmaceuticals, Others, LTL, and Container
4	1996	13	RA of Nanchang established	28	Ditto
5	1997–2000	14	RA of Kunming reestablished	28	Ditto
6	2001–2003	14	RA of Guangzhou renamed as Guangzhou Railway (Group) corp.	28	Electronics and electricity replaced by Electronics and electrical machinery, Textiles and leather products replaced by Textiles, leather, furs and related products, and Others replaced by Other goods
7	2004	15	Qinghai-Tibet Railway Company	28	Ditto
8	2005–2006	18	RA of Taiyuan, Wuhan, Xi'an established	28	Ditto
9	2007–2008	18	RA of Liuzhou renamed as RA of Nanning	28	Ditto

Note: RA is an abbreviation of railway administration.

inconsistency with current statistics of PTSL and the data were obsolete, we only choose the PTSL statistics during two periods of 1997-2000 and 2005-2008 for the further virtual data-mining.

3 Clustering method

The cluster analysis is an important part of data mining technology. It enables finding interesting data distribution model[3] underlying in the data.

Let *E* be the data set of the PTSLs of the periods of 1997-2000 and 2005-2008. *C* is defined as a nonempty subset of *E*, namely $C \subset E$ and $C \neq \cong$. Then the cluster satisfies these two conditions: $C_1 \cup C_2 \cup C_3 \cup \cdots \cup C_K = E$ and $C_i \cap C_j = \cong$ (at any $i \neq j$), the set of $c_1, c_2, c_3, \cdots, c_K$. The first condition ensures each sample in the sample set *E* must belong to a certain subset; the second condition ensures that each sample in the sample set *E* belongs to only one subset.

In this paper, the data matrix is represented by the $n \times p$ matrix, where p is the number of categories of goods and n is the number of railway administrations. Its dissimilarity matrix is a similarity matrix $R(i, j)_{n \times n}$ between any two of the n railway administrations. Although the PTSL is a continuous numerical variable and the measure magnitude of the clustering variables is similar, in order to reduce the measure unit of clustering results, the data should better be normalized before clustering, i.e.:

$$z_{if} = \frac{x_{if} - \overline{xf}}{sf} \tag{1}$$

where,

$$sf = \frac{1}{n} \sum \left| x_{if} - \overline{x}_f \right|$$

is set as the average.

For this type of variable we usually use distance to measure the difference between the two objects. For the distance function, we can apply the Euclidean distance, Manhattan distance and etc. Because there is only continuous numerical type of data sets, Euclidean distance is applied, i.e.:

$$SEUCLID = \sum_{i=1}^{k} (x_i - y_i)^2 \tag{2}$$

where k means in each sample there are k variables; x_i is the value of of the first sample at the *i*th variable; y_i is the value of the second sample at the *i*th variable; SEUCLID represents Euclidean distance squared.

Traditional cluster analysis methods can be divided into five types[4,5].

(1) Based on the segmentation method. This method first