Study on the subway transfer recognition during rush hour based on big data

Shushen Yao

School of Civil Engineering and Transportation,
South China University of Technology,
Guangzhou, 510641, China
and
School of Information,
Guangdong Communication Polytechnic,
Guangzhou, 510815, China
Email: 51421326@qq.com

Xiaoxiong Weng*

School of Information,
Guangdong Communication Polytechnic,
Guangzhou, 510815, China
Email: xiaoxiong.weng@hotmail.com
*Corresponding author

Abstract: With the development of the subway network, multipath coexistence becomes very common in big cities. It's followed that the tickets clearing problem is highly concerned by co-investors, which relies on accurate transfer paths identification. Different from the commonly used Logit models for subway transfer recognition problem, we adopted the adaptive Gauss cloud transformation (A-GCT) model, which transformed the distribution of passengers' trip time into multiple concepts of different granularity and evaluated the maturity of the concept by the of parameter named confusion degree (CD). The case in this paper shows that, the A-GCT model has higher accuracy in dealing with uncertain problem such as subway transfer recognition.

Keywords: Gaussian cloud transformation; GTC; subway transfer recognition; big data.

Reference to this paper should be made as follows: Yao, S. and Weng, X. (2020) 'Study on the subway transfer recognition during rush hour based on big data', *Int. J. Information and Communication Technology*, Vol. 16, No. 1, pp.43–52.

Biographical notes: Shushen Yao received his BS and MS degrees from the South China University of Technology (SCUT) in 2002 and 2005, respectively and he is currently a PhD student in the School of Civil Engineering and Transportation, South China University of Technology (SCUT). He also works at the Guangdong Communication Polytechnic as second unit. He has authored and co-authored more than ten peer-reviewed papers in academic journals and conferences. His major research interests include data mining, traffic signal control and public transportation.

Xiaoxiong Weng is currently a Professor in the School of Civil Engineering and Transportation, South China University of Technology (SCUT). Her major research interests include data mining, traffic information system and public transportation.

1 Introduction

1.1 Study background

Subway construction is an effective measure to deal with traffic congestion in big cities. In order to ensure convenient travel, the subway network generally adopts the seamless transfer mechanism, which only records the entry and exit station information and does not record the transfer station information. In fact, in the subway network, for the same starting point and end point, different passengers may choose different paths and there is uncertainty for each trip of the same passenger. The accuracy of subway transfer recognition and the fairness of passenger ticket clearance have become a historical problem in subway operation and management, as well as the core issue concerned by co-investors.

For a long time, the shortest-path-model is widely used in subway passenger flow distribution and ticket clearing problem, but the model usually has a large gap with the actual situation and cannot satisfy the co-investors (Wang, 2015). To solve the above problems, scholars introduced Logit model or its improved model and calculated the passenger flow distribution by building travel impedance function. However, Logit model requires manual investigation data to calibrate the model parameters, which has great uncertainty (Smith, 1979; Yan, 2015; Lin et al., 2012). In recent years, some scholars applied big data technology and machine learning algorithm to subway passenger flow distribution. But these studies are generally based on macro data and research shows that individual-level travel characteristics are often masked by macro-level regularity. In the real environment, the data at the individual level is not easy to obtain due to privacy and the analysis at the individual level is much more complicated. Therefore, these studies are just getting started and face many challenges (Morency et al., 2007; Zhu et al., 2014; Shi et al., 2015; Sun et al., 2015; Zhu et al., 2017; Weng and Wang, 2018).

The subway transfer recognition is essentially an uncertainty problem. Cloud model is a new method to study the uncertainty problem in the field of artificial intelligence (Li and Du, 2014; Li and Liu, 2004; Martin et al., 2006). In this paper, big data processing technology is adopted to collect full sample data of individual travel time, an adaptive Gaussian cloud transformation (A-GCT) method has been proposed to solve the problem of subway transfer recognition. The A-GCT method takes into account the individual travel difference and contains the index of concept confusion degree (CD), which can objectively evaluate the travel time overlap of different transfer paths.

The structure of this paper is organised as follows:

1 We introduce data format and its source in Session 2.

2 We explains how to construct the A-GCT model based on the spectrum function of travel time in Session 3.

3 In Session 4, we studied a specific case.

1.2 Contributions of this paper

Based on the individual travel time data, this paper provides a new idea for solving the problem of passenger ticket clearing in subway, by using the A-GCT method to identify the transfer route. The main contributions of this paper are as follows:

1 Different from the traditional Logit model, this paper construct the travel time spectrum function according to full sample data, to describe the degrees of certainty of travel time.

2 The A-GCT method is applied to subway transfer recognition and its validity is verified by comparing with Bayesian information criterion (BIC) method.

3 The accuracy of the Cloud model is further improved through Section-superposition method.

2 Data specification

The data adopted in this paper is from Guangzhou, China. For the case following, special travel data is selected from the rush hours (between 7:00 to 9:00 and 17:00 to 20:00) of five consecutive working days (March 7th to March 11th, 2016). The original data format is shown in Table 1.

Table 1 Data format

Type	*ID*	*Entry-line*	*Entry-station*
Monthly ticket	**********1905	3	Xiajiao
Entry-time	*Exit-line*	*Exit-station*	*Exit-time*
2016-03-07 18:07:27	3	Shipaiqiao	2016-03-07 18:27:59

3 Transfer recognition based on A-GCT method

3.1 Cloud of travel time

The data of travel time from entry-station to exit-station of all passengers via different routes during the rush hours is collected. The travel time spectrum is formed as Table 2, in which, the travel time spectrum function (Ft) is calculated by using the probability function.

It is assumed that the travel time from entry-station to exit-station constitutes a universal set U, which is denoted by precise numbers, in the problem of subway transfer recognition. The travel time C be the qualitative concept related to U. If there is travel time $x(x \in U)$ for a trip, which is a random realisation of the qualitative concept C and the certainty degree of x for C (i.e., $\mu(x) \in [0, 1]$) is a random value with stabilisation tendency. Then the distribution of x on U is defined as a Cloud.

Table 2 Travel time spectrum

Entry-station	*Exit-station*	*Time*	*Travel time*	*Spectral function of travel time*
Station A	Station B	i	T1	Ft(i, 1)
			T2	Ft(i, 2)
			……	……
			Tn	Ft(i, n)

Factors influencing specific travel time x include transfer route, departure time, passenger flow density, walking time between the gateway and platform and departure interval, etc. The travel time spectrum function describes the probability of all kinds of travel time from Entry-station to Exit-station at time i. So, we could use Ft in Table 2 as the μ(x) for the Cloud model of travel time.

3.2 Transfer recognition

The problem of subway transfer recognition is essentially a clustering and evaluation problem of multiple concepts. In view of the universality of Gaussian cloud in the concept representation (Li and Liu, 2004), an A-GCT model of travel time is constructed in this paper.

1 The concept CD threshold β is set to 0.50 or 0.64 to ensure that the formed concept is relatively maturely, as listed in Table 3.

2 Let the number of peaks m of the frequency distribution of the data sample set be the initial value of the concept number.

3 Use the GCT algorithm to cluster the data set into m Gaussian clouds.

4 Judge CD of each concept, if $CD_k > \beta$, $k = 1, \ldots, m$', then the concept of the number '$m = m - 1$', jump to step 3; otherwise, output m concepts which $CD_k < \beta$.

Table 3 GT division results measured by confusion degree

Confusion degree(CD)	*The concept explanation*
(0.6354, 1)	Confusion
(0.5004, 0.6354)	Not clear
(0.2, 0.5004)	A little clear
(0, 0.2)	Clear

We will get m concepts derived from A-GCT method, usually, it should correspond to the actual M transfer path, that is, the condition of '$m = M$' is satisfied. If '$m > M$', it means that there are some trips that do not conform to the conventional (minimum cost, minimum transfer, etc.), such as some passengers would miss stop or have a rest in the station, etc. This moment, we could eliminate these concepts whose ratio parameters were significantly smaller than others. If '$m = M$', it indicates that some transfer paths overlap too much in travel time and cannot be clearly divided by Cloud model.

3.3 Model improvement

When the A-GCT is used to identify different transfer paths, it is generally necessary to iterate over and over again to reduce the number of concepts m. In order to further optimise the calculation process and improve the accuracy of transfer recognition, the following improvement ideas are proposed:

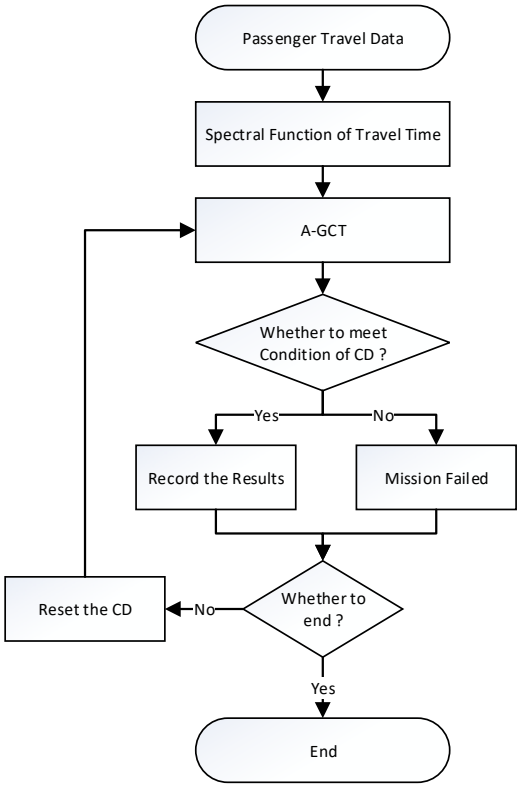
1 Take the transfer station as the demarcation point, divide the travel paths from entry-station to exit-station into several sections.

2 Collect the travel time data of each segment and calculate its expectation (Ex) using Cloud model.

3 The travel time Ex of the corresponding segment is superimposed to obtain the full travel Ex of each transfer path from entry-station to exit-station.

4 For transfer path j, set expectation from A-GCT as Ex_j and expectation from Section Superposition method as Ex_j'. If $\left|Ex_{j1}' - Ex_{j2}'\right| < \left|Ex_{j1} - Ex_{jx}\right|$, it means the $CD_{j1-j2}^{Actual} > CD_{j1-j2}^{A=GCT}$, the reliability of the A-GCT method will be reduced; on the contrary, if $\left|Ex_{j1}' - Ex_{j2}'\right| < \left|Ex_{j1} - Ex_{jx}\right|$, it means the $CD_{j1-j2}^{Actual} > CD_{j1-j2}^{A=GCT}$, the reliability of the A-GCT method will be improved.

Figure 1 Data flow chart



3.4 Programming and execution

In this paper, programs are developed based on the win10 system and Qt 5.9.6 platform. We import full sample data of passengers' travel time from entry-station to exit-station and gets the transfer recognition results and its CD index, as shown in Figure 1.

4 Case study

Select the travel data of five consecutive working days from Jiangnanxi station (entry-station) to Zhujiang New Town station (exit-station) in Guangzhou, a total of 1866 valid samples were obtained during the rush hours (7:00-9:00 and 17:00-20:00). In the subway network, there are two typical transfer paths connecting the above two stations, each of which requires two transfers. Table 4 lists the total distance and estimated travel time of the two transfer paths and Table 5 is the travel time spectrum based on sample data.

Table 4 Transfer path from Jiangnanxi station to Zhujiang New Town station

Path					*Distance (km)*	*Estimated travel time(min)*
Path no.	*Entry-station*	*Transfer-station 1*	*Transfer-station 2*	*Exit-station*		
j1	Jiangnanxi	Changgang	Kecun	Zhujiang New Town	8.7	27
j2	Jiangnanxi	Gongyuanqian	Tiyu Xilu	Zhujiang New Town	10.9	36

Table 5 Travel time spectrum from Jiangnanxi station to Zhujiang New Town station

Entry-station	*Exit-station*	*Time*	*Travel time (Min)*	*Samples*	*Spectral function of travel time*
Jiangnanxi	Zhujiang New Town	Rush Hours (7:00-9:00 and 17:00-20:00)	22	18	0.01
			23	91	0.05
			25	168	0.09
			27	224	0.12
			29	127	0.07
			31	273	0.15
			32	284	0.15
			34	223	0.12
			36	152	0.08
			38	55	0.03
			40	109	0.06
			41	66	0.04
			43	41	0.02
			45	11	0.01
			47	13	0.01

Table 5 Travel time spectrum from Jiangnanxi station to Zhujiang New Town station (continued)

Entry-station	*Exit-station*	*Time*	*Travel time (Min)*	*Samples*	*Spectral function of travel time*
			49	5	0.00
			50	3	0.00
			52	2	0.00
			54	1	0.00
			Sum	1,866	1.00

4.1 BIC method

In this section, we use the Bayesian information criterion (BIC) method to identify the case. The BIC method divides passengers into n categories based on travel time and calculates the proportion of each type of travel. The concept of low proportion is not representative and may be abnormal data. The final identification results based on BIC method are shown in Table 6.

Table 6 Identification results based on BIC method

Path (j)	*Travel time*	
	Ex (min)	*Ratio(%)*
j1	30.02	42
j2	34.20	43
j3	24.62	13

The BIC method runs efficiently, but it does not evaluate the overlapping of two adjacent categories. So it cannot guarantee the independence of all types of travel. For example, we are not sure whether path j1 and path j2 in Table 6 can merge a large class with an Ex of 32 minutes and a Ratio of 85%. The rationality of BIC method may be questioned.

4.2 A-GCT method

We will use the A-GCT method in this section. The travel time clustering failure when we set the concept confusion degree threshold $\beta = 0.50$; and the travel time clustering into two transfer paths when we set the concept confusion degree threshold $\beta = 0.70$, as shown in Figure 2 and Table 7. Although there are two transfer paths in reality, the two transfer paths' travel time obtained by the A-GCT method have a difference of 5 minutes, less than the estimated 9 minutes. At the same time, the two transfer paths identified by the A-GCT method are seriously overlapping, with the confusion degree reaching 0.68, the result of concept division is at the critical point of Confusion and Not clear.

Figure 2 Recognition results based on cloud model (see online version for colours)

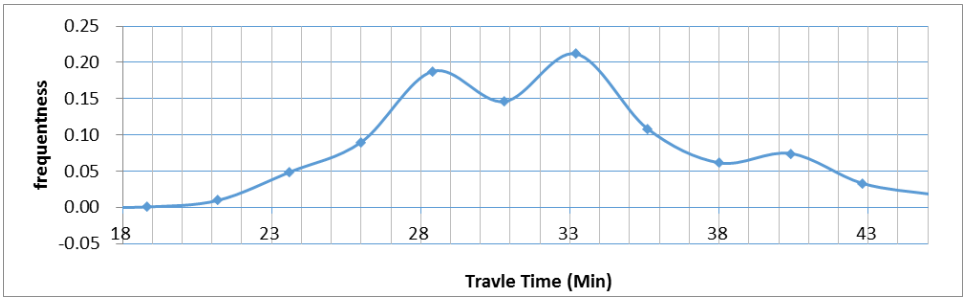


Table 7 Recognition results based on cloud model

Path	*Travel time*		
	Ex(min)	*CD*	*Ratio (%)*
j1	28.61	0.68	52
j2	33.95	0.68	48

It can be seen that the cloud model has also achieved good results in multi-path recognition. More importantly, compared with the BIC method, the A-GCT method can also evaluate the reliability of transfer recognition through CD values.

4.3 Section-superposition method

Taking the transfer station as the demarcation point, the two travel paths in this case are divided into several sections, as shown in Table 8. The accumulated travel time were 32.16 minutes (Path j1) and 33.51 minutes (Path j2) and the difference between them was reduced compared with that of unsegment, it means $\left|Ex'_{j1} - Ex'_{j2}\right| < \left|Ex_{j1} - Ex_{jx}\right|$, and the reliability of the A-GCT method will be reduced, which further verifies the conclusion that the effect of the division of the two transfer paths in this case is not clear enough.

Table 8 The statistical result of section superposition

Path	*Section of the path*			*Travel time*	
	No.	*Entry-station*	*Exit-station*	*Samples*	*Ex (min)*
j1	1	Jiangnanxi	Changgang	1,089	5.45
	2	Changgang	Kecun	961	12.75
	3	Kecun	Zhujiang New Town	3,137	13.96
			Sum	5,187	32.16
j2	1	Jiangnanxi	Gongyuanqian	3,323	9.95
	2	Gongyuanqian	Tiyu Xilu	1,506	16.54
	3	Tiyu Xilu	Zhujiang New Town	2,280	7.02
			Sum	7,109	33.51

5 Conclusions

In this paper, the A-GCT model is constructed based on full sample data of individual travel time, which providing an effective and reliable solution for subway transfer recognition problem. The results and conclusion of this paper are as follows:

1 Commuters who travel in rush hours tend to choose a path with a relatively short travel time, but do not exclude other paths with a small increase in travel time. The actual difference of travel time for each paths will be reduced, which will increase the difficulty of recognition. The reason is that more passengers tend to choose the transfer path with relatively short travel time, which causes these paths to be relatively crowded in the rush hours and the actual time consumption will increase. At the same time, fewer passengers chose the longer route, resulting in less time actually spent on those paths than estimated.

2 The Cloud model not only realises the transfer recognition, but also evaluates the recognition effect evaluates through the CD values. So it is more intuitive and reliable than the BIC method.

3 The improvement using section superposition method can improve the accuracy of the transfer recognition evaluation.

References

Li, D.Y. and Du, Y. (2014) *Artificial Intelligence with Uncertainty*, 2nd ed., National Defense Industry Press, Beijing.

Li, D.Y. and Liu, C.Y. (2004) 'Study on the universality of the normal cloud model', *Engineering Science*, Vol. 6, No. 8, pp.28–33.

Lin, Z., Jiang, M.Q., Liu, J.F. and Si, B.F. (2012) 'Improved logit model and method for urban rail transit network assignment', *Journal of Transportation Systems Engineering and Information Technology*, Vol. 12, No. 6, pp.145–151.

Martin, A., Laanaya, H. and Bos, A.A. (2006) 'Evaluation for uncertain image classification and segmentation', *Pattern Recongnition*, No. 39, pp.1987–1995.

Morency, C., Trépanier, M. and Agard, B. (2007) 'Measuring transit use variability with smart-card data', *Transport. Policy*, Vol. 14, No. 3, pp.193–203.

Shi, J.G., Zhou, F., Zhu, W. and Xu, R.H. (2015) 'Estimation method of passenger route choice proportion in urban rail transit based on AFC data', *Journal of Southeast University*, Natural Science Edition, Vol. 45, No. 1, pp.184–188.

Smith, M.J. (1979) 'The existence uniqueness and stability of traffic equilibria', *Transportation Research Part B*, Vol. 13, No. 4, pp.295–304.

Sun, L., Lu, Y., Jin, J.G. et al. (2015) 'An integrated Bayesian approach for passenger flow assignment in metro networks', *Transport. Res. Part C: Emerg. Technol.*, Vol. 52, pp.116–131.

Wang, M. (2015) *Research on Passenger Flow Distribution of Urban Rail Transit Based on Travel Time Reliability*, Beijing Jiaotong University.

Weng, X.X. and Wang, Z.P. (2018) 'Passenger flow distribution model for urban rail transit based on BP neural networks', *Journal of Chongqing Jiaotong University*, Narural Science, Vol. 37, No. 1, pp.1–12.

Yan, Y. (2015) *Traffic Assignment Problem of Urban Railway Transit Based on Seamless Transfer*, Southwest Jiaotong University.

Zhu, W., Hu, H. and Huang, Z. (2014) 'Calibrating rail transit assignment models with genetic algorithm and automated fare collection data', *Computer, Aided Civ. Inf.*, Vol. 29, pp.518–530.

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H.M. (2017) 'A probabilistic passenger-to-train assignment model based on automated data', *Transport. Res. Part B: Meth.*, Vol. 104, pp.522–542.