Evaluation of urban public transportation service

Abstract: Evaluation of urban public transportation service is an important part of the urban public transport system and the improvement of public transport operational efficiency. The reasonable and effective evaluation of urban transportation services is beneficial to the optimization of urban traffic management, traffic accidents and traffic accidents. Based on IC card data and GPS positioning information, the paper takes Matlab, Arcgis, SQL Server software as the working platform to construct the relevant model.

At the problem 1, At least three methods are required to deal with the original data.At first, We found that the original data is primarily GPS positioning information through the analysis. The location accuracy and latitude have very obvious errors, so the error message is mainly in the form of isolated points. In this paper, we first construct the Gauss mixture model, K- cluster analysis and SOM self organizing neural network model. The three methods are to complete the division of the original data time, clean up the extraction of the field and the operation of the bad values. These three methods are mainly used to cluster the data and identify the outliers in order to clean up.

At the problem 2, the problem requires the use of IC card data and GPS positioning information to complete the establishment of the OD matrix and related analysis. According to the characteristics of the original data, the paper first through the server SQL software for the IC card and GPS data table connection, and then to the bus and IC card by car tail number matching, and then in accordance with the GPS record time and IC card consumption time to match, use the Arcgis software combined with GPS positioning data and bus site distribution data to determine the site. Then according to the two site model to determine the station point for the OD matrix construction, and continue the relevant analysis based on the actual situation

At the problem 3, the problem requires to build a mathematical model to complete the bus station, subway station service level of dynamic evaluation, and select the three most unreasonable site. In this paper, described as the problem of 3,at first,we searched indicators that can reflect the level of public transport service. mainly divided into two aspects of the direct response indicators and indirect indicators. The direct reaction index includes two aspects of site index and the ratio of I/O to point flow respectively, to evaluate the site refers to the site on time punctuality of bus, mainly composed of pure punctuality indicators and other site effects of two parts.And indirect evaluation index mainly refers to the location of the bus station, in this we build a site selection model, the main factors include the annual average rainfall, the surface gradient, population density and bus density, obtain the relevant data through the Shenzhen Transportation Bureau or other relevant platform, through the analytic hierarchy process to determine the weight, and weighted overlay.

Key words: Clustering analysis Matching Dynamic evaluation OD matrix

Contents

1.Problems Background

Evaluation of urban public transport services is an important part of the urban public transportation system construction and an important part of the bus operator efficiency. In the many problems caused by China's urbanization process, the traffic is not only the greatest impact factor, but also the most attentional aspect. Because the various aspects of the city's economy and life are closely related with the traffic. The current condition of traffic congestion in some cities is quite serious. At present, In addition to Beijing, Shanghai, Shenzhen and other large cities, traffic jams have been more frequent appeared in a number of other cities in China, and congestion is not only reflected in the intersection peers but also in the large urban area.The road capacity has been saturated. According to the relevant materials, the city's economic loss due to the impact of urban traffic-related problems is up to hundreds of billions each year . For planning data transport business, management and planning department, the traditional bus stations, lines and hubs are only based on statistics collected by the competent authorities and manual inventory.Today the automated collection techniques become more developed, if we can automatically analyze residents travel demand and make use of public transportation system data, consumption data bus card, subway card spending data and taxi location data to evaluate the existing transit planning facilities services (including regular bus stops and subway stations) dynamically.So,we can significantly improve the traditional transit planning, efficiency and quality of design and management.

2.Problem analysis

2.1 Problems restatement

Based on the research of IC card data and GPS information and the relevant geographic factors in the research area, we can construct the relevant OD matrix and analyze them. At the same time, we can make a dynamic evaluation of the existing bus service facilities to achieve the optimization of public transport routes in the study area, and provide reference for the development of the relevant transport sector policy. To this end, we need to solve the following three problems:

1.Use at least three methods to preprocess the data and compare the results of the data processing.

2.OD matrix analysis based on global positioning system and public transportation card consumption.

3.Please establish the mathematical model of the dynamic evaluation of the service of the bus station and the subway station, and rearrange the most unreasonable three bus stops.

2.2 Problem analysis

This problem is mainly related to analysis and evaluation of Shenzhen City Road traffic, the key lies in data mining which based on the data provided by the accessory and the internet, and get the raw data to solve the problem, then according to the subject requirements of the relevant knowledge, reference to the relevant data to establish the corresponding mathematical model, thus solving the relevant problems.

In view of the problem a, this paper adopted three kinds of methods for data preprocessing, the difference of the three methods is the exploration of the wrong data, three models were constructed to deal with the error data, respectively, based on Gauss mixture model, SOM self-organization neural network and K means clustering model.

In view of the problem b, we established a OD matrix estimation model based on the two site analysis model, Respectively, including looking for the get on the bus and get off the bus. In order to make good use of the raw data, we use SQL software to carry out the relevant data connection and query, so as to construct the OD matrix, and its correlation analysis.

In view of the problem c, the service level of the site is directly and indirectly evaluated in this paper. Indirect evaluation is the suitability of the location of the bus or subway station construction, and solving the suitability model by constructing a geographic location; Direct evaluation mainly refers to through the site service directly relevant factors, such as site on time, the I/O streams into consideration, mainly by constructing real-time dynamic evaluation model to solve it.

3. The model assumptions

1.it is assumed that Shenzhen bus IC card data and GPS information represent the entire urban residents travel to a certain extent , the selected data has a strong representation.

2.it is assumed that Shenzhen residents’ travel meets the two-site model, the second residents travel by bus on the train station at the same time should be the first time residents travel by bus to get off site.

3.the noise of the random errors recorded by IC card and GPS data is negligible, the error is mainly gross error derived from instrument recording and data storage brings.

4.It is assumed that at the time of a single line of public transportation OD matrix analysis, bus trips is exclusive, namely don’t consider the impact of other trips to the site.

4 .Symbol Description

Symbol	Meaning
a	The proportion of single-Gaussian
P	probability
ς	Posterior probability
u	average
α	learning rate
d	Arrive station distance
X(t)	Vector value at time t
r	Output node location
JC	Quality metrics

Z	The clustering center
β	Passengers on the train auxiliary parameters
γ	Passengers get off the auxiliary parameters
ν	Get off the number of passengers
PY	On time reliability
TD	The time interval
θ	Time range parameter
C	The ratio of the number of passengers getting on or off
δ	The number of the standard range of on and off

5. The establishment solution of the model and solution

5.1 Technical route

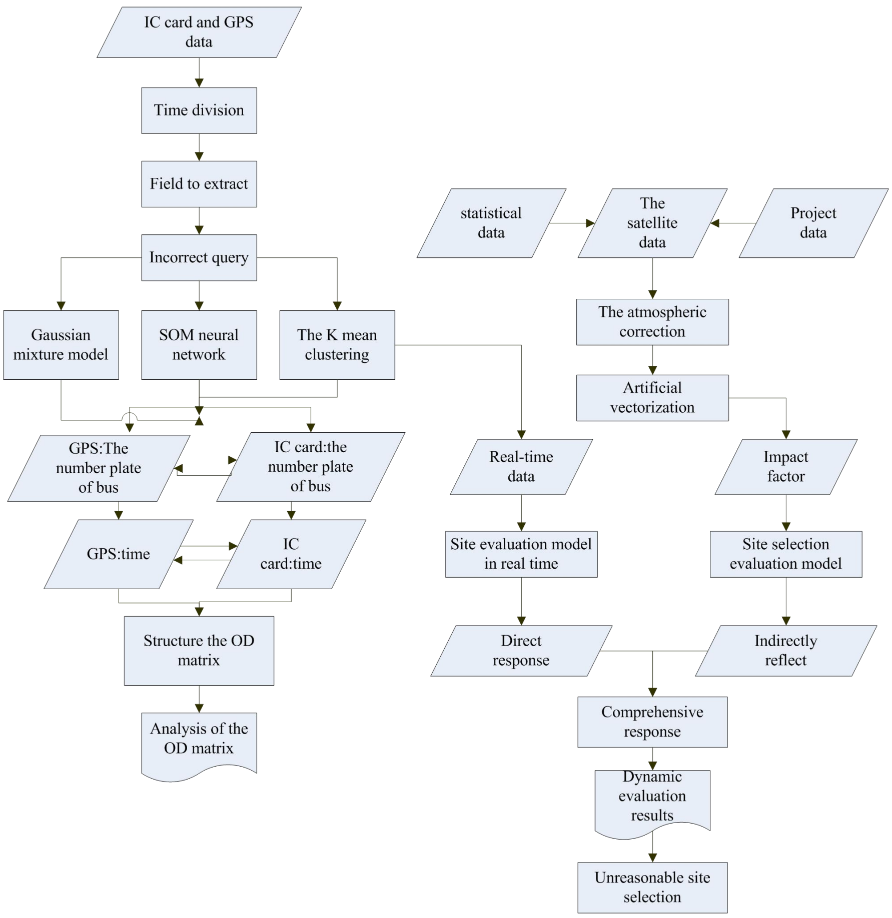


Figure 1 model building technology route

5.2 The first problem to use at least three methods to complete the data preprocessing

During the data acquisition of the IC card data and GPS positioning information, equipment technical conditions and transmission conditions are affected, the collected data has certain defects in quality.Therefore, The preprocessing work based on the city bus preliminary data is indispensable in order to protect the underlying data complete and accurate.

Pretreatment of the data mainly includes three aspects, the selection of the analysis period, the data field selection and wrong data cleaning .The difference of data preprocessing methods selection focused on the third step error data cleaning.The former two-step process don’t have much change . Wherein the specific process is as follows:

5.2.1 Selection of period

City bus operators run trips carried out strictly in accordance with time and closely related to the characteristics of urban residents’trip with some periods of the periodicity.[2] General city bus system regard days, weeks, months as the basic unit of analysis,wherein one day can be separated smaller by the hour periods . In this article we derivate the travel 0D matrix while we regard one day as unit in the second question.

5.2.2 Data field select

There are millions number of Bus IC card and GPS data source in one day while data records often contain abundant data.

The fields of a data which is meaningful to data mining are selected and removed.It helps improving the bus data mining quality and efficiency. the fields of bus IC card data include card card type, credit card time, line number, vehicle number. GPS positioning data mainly select longitude, latitude positioning, recording time, speed, license plate number.

5.2.3 Erroneous data cleaning

Limited by equipment and transmission conditions, the data sources that are collected will usually produce erroneous data, we clean or correct the error data in the data field, thus ensuring the quality of data mining. In this paper, we give three kinds of models of error data cleaning, which are the Gauss mixture model, the SOM self-organizing neural network model, and the K mean clustering model. Model is introduced and the advantages and disadvantages of analysis are as follows.

5.2.3.1 Model one the establishment and solution of Gauss mixture model

(1) The establishment of the model of the Gauss mixture model is the extension of a single Gauss probability density function. There are a number of observation data, data number is n, the distribution in the D dimensional space is not an ellipsoid, so it is not suitable to describe the probability density function of these data points with a single Gauss density function. At this point, we use a flexible scheme, assuming that each point is generated by a single Gauss distribution, and this batch of data were generated by M (explicit) single Gauss model, which one of the data X_i is specific to a single Gauss model is not known, and the proportion a_j of each single Gauss model in the mixed model is unknown. All data points from different distributions are mixed together, which is called the Gauss mixture distribution. Mathematically speaking, we think that the probability distribution of these data can be expressed by a weighted function:

$$p(x_i)=\sum_{j=1}^{M}a_jN_j(X_i;\mu_j,C_j),\sum_{j=1}^{M}a_j=1 \quad (1)$$

among them

$$N_j(x;\mu_j,C_j)=\frac{1}{\sqrt{(2\pi)^n}\left|C_j\right|}exp[-\frac{1}{2}(X-\mu_j)^TC_j^{-1}(x-\mu_j)] \quad (2)$$

SGM represents the j-th of PDF.

Order $\varphi_j=(a_j,\mu_j,C_j)$,GMM have M SGM, Now, we need to estimate all the parameters of GMM by sample set X: $\phi=(\varphi_1,\cdots,\varphi_M)^T$ The probability formula of the sample X is:

$$P(X\mid\phi)=\prod_{i=1}^{N}\sum_{j=1}^{M}a_jN_j(X_i;\mu_j,C_j) \quad (3)$$

(2)Solving model

The first step: the covariance matrix Cj0 is set to the unit matrix, the prior probability of each model is a_{j0}=1/M; the mean u_{j0} is set to random number.

The second step: the estimation step (E-step)

The posterior probability of the a_{j0} is

$$\varsigma_{ij}=\frac{a_jN_j(X_i;\phi)}{\sum_{k=1}^{M}a_kN_k(X_i;\phi)},1\le i\le n,1\le j\le M$$

The third step: the maximization step (M-step)

Update weights:

$$a_j=\frac{\sum_{i=1}^{N}\beta_{ij}}{N} \quad (5)$$

Update mean:

$$\mu_j=\frac{\sum_{i=1}^{N}X_i\beta_{ij}}{\sum_{i=1}^{N}\beta_{ij}} \quad (6)$$

Update variance matrix:

$$C_j=\frac{\sum_{i=1}^{N}\beta_{ij}(X_i-\mu_j^T)(X_i-\mu_j^T)^T}{\sum_{i=1}^{N}\beta_{ij}} \quad (7)$$

The fourth step :convergence condition

The second step and the third step is to repeat the above three values, until the $\left|p(X|\phi)-p(X|\phi')\right|<\varepsilon$, which is the value of the updated parameters, that is, the results of the two iterations are less than a certain extent, the end of the iteration, usually $\varepsilon=10^{-5}$.

The process of the above algorithm is programmed by Matlab software, and the GPS latitude and longitude information is taken as the original information as shown in Figure 2. Algorithm output results are shown in the following figure, for the convenience of viewing, the output of the results were made of the thermal map, as shown in Figure 3 and frequency, as shown in figure 2.

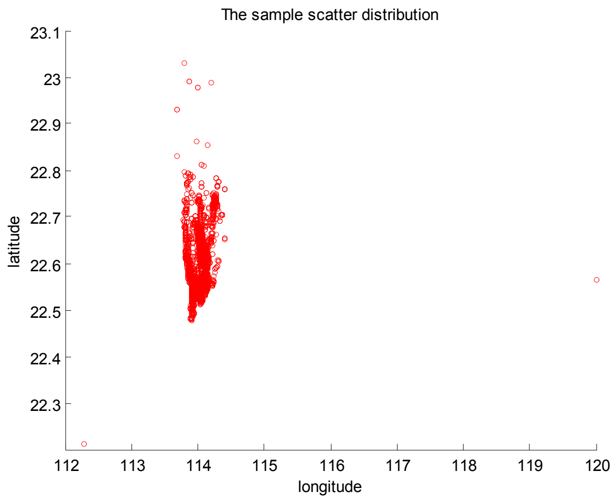


Figure 2 original data distribution

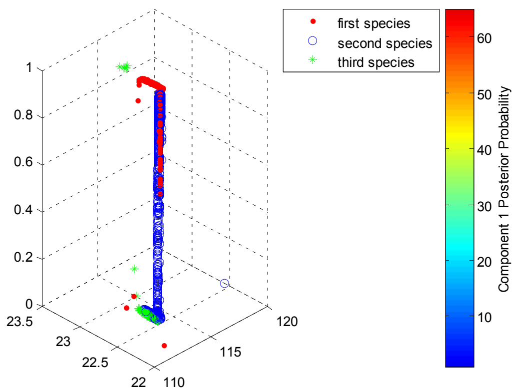


Figure 3 gaussian clustering heat

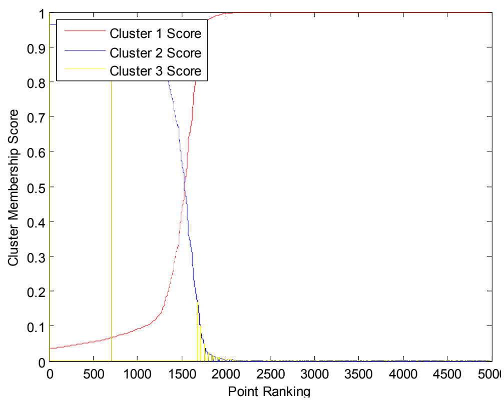


figure4 gaussian clusteringprobability

According to figure 3,4, the noise information is mainly for the third class or the probability of a small number, after the data back to check, you can get some of the wrong data as shown in the table1.

Table1 Part erroneous data

longitude	latitude	license plate
0	0	YBN0287
0	0	YBF1417
0	22.59	YB75502
0	0	YBF0147
0	0	YB76445
0	0	YBU0046
0	0	YBH6447
0	0	YBD1126
0	0	YB92093

5.2.3.2 Establishment and solution of SOM self organization neural network model

There are two layers in SOM self-organizing neural network model and SOM solving network: The input layer and output layer; An input layer (single neuron arrangement) will gather outside information to the output layer each neuron by weight vector; Output layer also called competitive layer neurons, has a one-dimensional arrangement, linear two-dimensional and three-dimensional array of grid array. It has two distinctive features:

A. Topology mapping structure dosen't achieve by movement of neurons reorganizition, but formed constitute a whole by individual neurons in different excited state;

B. The formation of this structure has the characteristics of self-organization.Neuronal tissue topology is its fundamental characteristics;

C. According to the geometric center or features the cluster as flowed also called two-dimensional array.

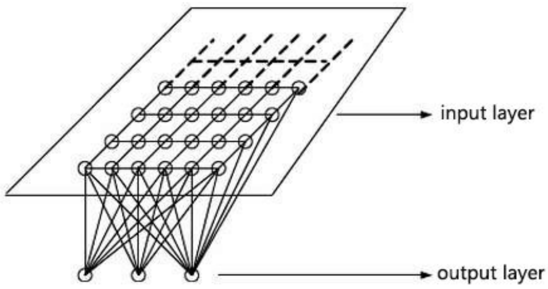


Figure 5 SOM neural distribution

According to Kohonen algorithm,the affecting to Neighboring neurons of winning neurons is from far to near, from excitement to inhibited gradually, namely winning neuron should adjust the weights, the neural element surrounding weight vectors have varying degrees of adjustment. Weight adjustment function include Mexican hat function, top hat function, chef hat function.

(1)The operating principle

Its operation is divided into two stages of training and work. When it training, The samples of training set are input randomly .For a particular input mode, the output level will be a node generating maximum response and win. The Network use a large number of training samples adjust the network weights through self-organization. Finally ,the output layer nodes become specific pattern-sensitive neurons and the corresponding inner star vector weight vectors become the center of each input mode.

(3)Learning (learning algorithm SOM network uses)

A**:** Initialization

The output of each weight vector was assigned small random number and normalized, choosed random values [0,1].The only restriction is mutually diferent, to get w_i(i=1,2,...,m) to establish the initial winning neighborhood N*(0) and the initial value of a learning rate, M is the number of neurons of the output layer

B: Accepting input (sampling)

An input mode was chosen from the training set randomly and normalized to get X_i(i=1,2,...,n), N is the number of neurons in input layer.

C**:** Calculating distance of time

Calculate the input vector at time t the distance to all output node

$$d = \sum (x(t) - w(t))\hat{}2 \qquad (8)$$

X (t) is the input vector value at time **t**

D: Looking the winning node

Calculate the minimum distance **d** as the winning time of node j *.

E: Define the winning neighborhood N * (t)

J * as the center to determine the weight adjustment at time **T** , namely the surrounding areas function of the winning neurons. Using Gauss neighborhood function, generally an initial neighborhood N*(0) is large (about 50% to 80% of the total node).In the training process,the N*(t)shortens with systolic training time.

N*（t）=exp（-|r1-r2|）/2b(t)b(t) （9）

Where r1, r2 are the position of the output node, **b** reflects the range of the neighborhood

F:Adjust weights

Adjust the weights of all nodes in the winning neighborhood N * (t),adjust the weight vector by updating formula.

W（t=1）=W（t）=a（t）N*（t）（x（t）-W（t））　（10）

G: Ending judgment

When learning rate a (t) <= a (min) (after all samples learned)after training; When it is not satisfied,end the condition, then go to Step B to continue.

The aboveal gorithm Coding, specific coding program stored in the annex, the program output shown in Figure 6, it can be seen that a minimum number of points that is classified class and other classes larger distance between classes in Matlab, based on experience seen such as noise, anti-check data may get an error value.

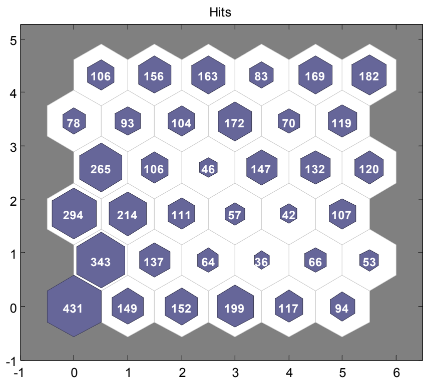


Figure 6 a

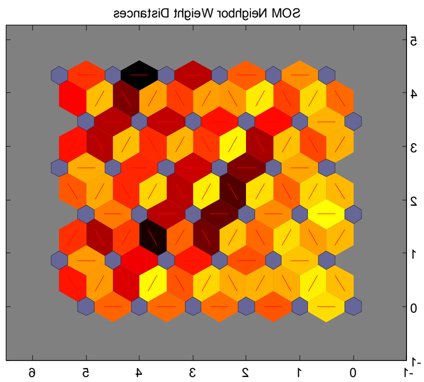


Figure 6b

Figure 6 a:SOM neural network distribution; b:SOM neurons distance distribution

5.2.3.3 Model two—The establishment and solution of K mean clustering model

(1)The process of model establishment

A. Set up the number of data samples that need to be clustered is n, I=1, select K initial clustering center: $Z_j(I)$,j=1,2...k.

B. Calculate the distance from each data sample to the cluster center.
$D(x_i,Z_j(I))$,i=1,2...n, j=1,2...k;
if $D(xi,Z_k(I))=\min\{D(x_i,Z_j(I))\}$, i=1,2...n
Then, the sample belongs to the class.

C. In general, the error square and the target function as a quality measure of the clustering are adopted:

$$JC(I)=\sum_{j=1}^{k}\sum_{k=1}^{nj}\left\|x_k{}^j-Z_j(I)\right\|^2 \qquad (11)$$

D. To determine whether the algorithm reaches the end condition:

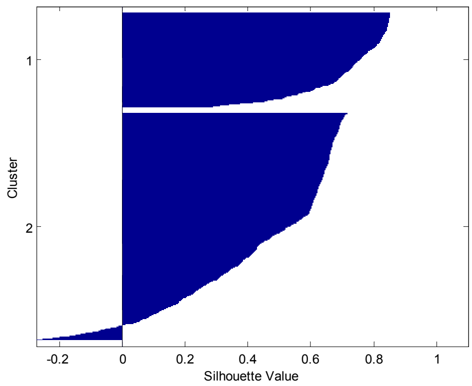
if $\left\|x_k{}^j-Z_j(I)\right\|<\xi$, Then indicates the end of clustering, otherwise, I=I+1, calculate K new clustering centers, and return B, the new clustering center is calculated using the following formula:

$$Z_j(I+1)=\frac{1}{n_j}\sum_{i=1}^{k}X_i{}^j, j=1,2,\cdots,k \qquad (12)$$

The traditional K means clustering algorithm has the characteristics of simple computation and fast clustering speed.

(2) Solving the model

Using Matlab to achieve the above process, it can be obtained as shown in the figure below, figure 7a shows that the clustering effect is divided into 3 categories, figure 7b shows that the clustering effect is divided into 5 categories.



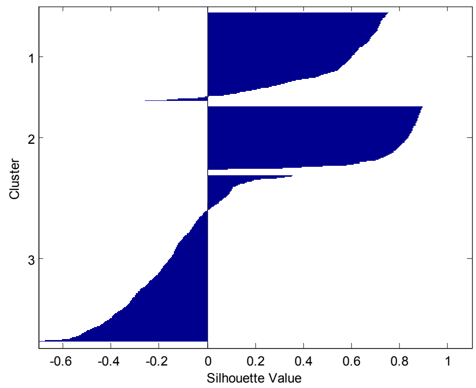


Figure 7a figure 7b

Figure 7a:K - clustering analysis distribution of classification number 3;b:K - clustering analysis distribution of classification number 5

According to figure 7 and figure 8, we can find that the error information is a small group, and the corresponding original data can be found to find the corresponding noise information, and then remove.

5.2.3.4 Comparison of three methods

The above three methods are suitable for the elimination of false information, this is because the error information is generally refer to the noise point information, mainly due to the instrument itself or the use of improper use of the operator, the main feature is the difference between the other data, that is, in the form of isolated points. The three method is a classical algorithm of clustering analysis, so it can be based on the idea of clustering. It is different that the Gauss mixed model is suitable for a large number of data, the effect of pretreatment of the data is best in this paper, because of the noise information of this paper is mainly based on GPS latitude and longitude error information, and the bus has a certain degree of clustering, so it is easy to find the inter class attribute. The disadvantage is that if the distribution of noise information is scattered or distributed in the original data is small, it is not suitable for exploration, and further classification of class attributes is needed. SOM self-organization neural network model is a intelligent neural network algorithm for unsupervised learning. The clustering of data is divided by the Euclidean distance. The distance measurement of the topological neurons can be very obvious. The results show that the large data is limited by the size of the step size, if erroneous data profiling with high accuracy is required to increase the step, this will result in the increased number of classification which is more difficult to find on the erroneous data, we suggest that the proposed step length should be set to about 10 or batch processing. K mean clustering algorithm design principle is based on probability density, which can be more accurate to identify the original data, and with the increase in the number of classification can be more accurate data mining. The disadvantage is that the algorithm is also suitable for classification with strong inter class attributes. The difference is very small or the fusion noise is less effective.

5.3 The second problem building OD matrix through IC card data and

GPS positioning information

5.3.1The building of model four

For this study about OD Matrix Estimation of the Resident bus travel, we need to analyze GPS information and public transportation IC card information to get the information about getting on the station, getting off the station, getting off time. And completing the inhabitants bus study on origin-destination travel[3]. For the station residents get on in the bus travel, we can complete by following steps:

(1)Get their credit card number by the residents bus IC card-based information table, since uniqueness, we may find the corresponding buses from the bus GPS information table;

(2)After finding the corresponding buses'number,extract the time residents get on the bus from the residents travel information table. Isolated the latitude and longitude position corresponding to the bus from the corresponding GPS-based information tables, namely latitude and longitude position where residents getting on the train;

(3)GIS bus electronic map as described in this part 4.1.4.2 of this article, find out the latitude and longitude position where residents commuting on the train, corresponding to the appropriate site,called the bus station the resident bus travel on. the off site and off time can't be extract the information directly from the underlying table, we need to calculate the theoretical study data related to fusion. By projections and analysis, we can see that there is transition of most travelers on the bus when they get off site or get on site, that is the last trip departure is the end of the return of the site, and the last time the end of the trip is a starting point backhaul. In this state, we are constrained by certain conditions, bus travel is estimated to meet the residents of the two-site model proposed travel, bus travel for the study of urban residents will improve great convenient. While in this assumption great errors exist, this article will control it by specific constraints.

So for record passenger bus IC card information of two consecutive day, the second time residents travel by bus on the station at the same time should be the first time residents travel by bus to get off site. Thus, this paper obtained the off site, on site and off time by the following method:

(1)Consists of two information bus IC basic information table recorded continuously, extract passenger information card where the second bus number and boarding time;

(2)Find a resident in accordance with the corresponding method on bus travel car sites, investigate the residents of bus travel to get off the site;

(3)After the corresponding bus travel to get off the site of the residents, look for the bus by the GIS electronic map latitude and longitude location of the site, then by bus with GPS information table on the basis of passenger car time and direction of travel, from GPS-based passenger information table extract take a trip time through the position of the bus, called geting off the bus travel time. From the above analysis step, we can basically get residents bus travel beginning and ending point for complete information, including travel time on the bus, on the train station, get off at the time, off site, the bus and other basic information, shown as follow form.

At the same time, it introduces "a certain amount of bus travel" principle in the course of the research study assisted amended. The so-called "total bus travel must be" based on the principle that the number of passengers on the bus should get off equal the number of get on, the number of passage can get through the IC card information, the total number of bus passengers in obtaining the case we would get at the same time the total number of passengers. At the process of estimating number of passengers get off, since the impact of statistical information and other factors, the number of passengers be off and the number of passengers on the bus are not equal. Therefore, in practical processing, increase certain auxiliary conditions herein, set auxiliary parameters, adjust the number of passengers getting off and make it equal to the number of passengers on the train, so realistic objective case, the findings of this article is more scientific and reasonable application stronger. According to this principle, "the total amount of bus travel must" adjust auxiliary parameters main cause of bus travel on the train passenger count equal to the number of passengers get off the bus travel statistics and, therefore, this paper calibrate the number of passengers on the bus trip car auxiliary parameters and bus travel the number of passengers to get off the auxiliary parameters .as followed:

$$\beta = \frac{A \cdot \eta}{\sum_{i=1}^{m}\sum_{j=1}^{n} \upsilon_{ij}} \qquad (13)$$

$$\gamma = \frac{A \cdot \eta}{\sum_{j=1}^{n}\sum_{k=1}^{m} \nu_{jk}} \qquad (14)$$

Wherein β , γ , respectively, the number of car passengers to get off the auxiliary parameters and number of passengers auxiliary parameters, υ_{ij} represents the first i sites take the first swipe j number of passengers on the bus trip car trips.

A represents the j-line shift in the number of passengers to get off the k-th site, and m represent the total number of sites , n represents arrange class times. According to the analysis, Line number of passengers from the car i-th site, the k-th number of passengers get off the bus station is:

$$\lambda_{ik} = \sum_{j=1}^{n} \lambda_{jik} \qquad (15)$$

A represents the number of bus passengers j-shift car from the i-th on site car, get off the k-th site.From the above analysis, the paper get the basic information and its auxiliary parameter single-line bus passenger OD and comprehensive integration, the availability of the bus lines OD matrix such as:

Table3 OD matrix structure

go off / go on	1	2	…	m	合计
1	0	$\frac{\beta}{\eta}\lambda_{12}$	…	$\frac{\beta}{\eta}\lambda_{1m}$	$\frac{\beta}{\eta}\sum_{i=1}^{m}\lambda_{1i}=\frac{\beta}{\eta}\sum_{j=1}^{n}\alpha_{1j}$
2	$\frac{\beta}{\eta}\lambda_{21}$	0	…	$\frac{\beta}{\eta}\lambda_{2m}$	$\frac{\beta}{\eta}\sum_{i=1}^{m}\lambda_{2i}=\frac{\beta}{\eta}\sum_{j=1}^{n}\alpha_{2j}$
…	…	…	0	…	…
m	$\frac{\beta}{\eta}\lambda_{m1}$	$\frac{\beta}{\eta}\lambda_{m2}$	…	0	$\frac{\beta}{\eta}\sum_{i=1}^{m}\lambda_{mi}=\frac{\beta}{\eta}\sum_{j=1}^{n}\alpha_{mj}$

So far, this article is substantially complete single line bus OD Matrix Estimation study, a single bus lines OD matrix, for between research lines reasonable respective sites passenger distribution, excavation line peak passenger section, provide the basis adjustment for a single line optimization, and as the city's bus OD Matrix basic information.

5.3.2 Solution of model four

For single-line fixed trips, the M322 route of the bus company in eastern Shenzhen is the research route in this paper, the GPS data and IC card data for data matching, due to insufficient amount of data exists, as well as to make the results more reasonable to show, M322 route's 42 sites have been selected, 14 sites were chosen to build, select the results shown in Figure 9, the same construction method, wherein M322 route and site distribution shown in Figure 8.

Figure 8　Shenzhen east bus M322 route and site distribution map

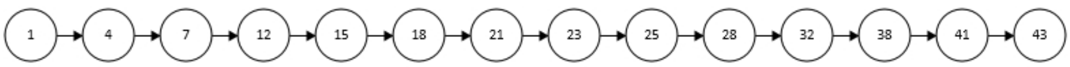


Figure 9　Distribution map of public transit station

Using SQL database language programming, get the GPS data and the IC card data connection, and then according to the time and the car tail number to get the data (part) as shown in table 4, SQL into code stored in the appendix.

Table 4 GPS and IC card matching data

YB76558	YBC7947	YBJ4042	YBL8100
YB87849	YBH0948	YBK6510	YBL8128
YB98892	YBH1025	YBL4847	YBL8289
YBB1167	YBH1072	YBL5457	YBL8290
YBC2906	YBH1090	YBL7487	YBL8323

According to the car tail number and time matching data, combined with the GPS bus positioning data, according to the nearest principle of the bus station, the property division, the results of the model are shown in table 5.

Table5 M322 route OD matrix

D / U	1	2	3	4	5	6	7	8	9	10	11	12	13	14	sum
1	0	11	27	17	74	23	21	19	39	32	56	32	45	67	463
2	12	0	18	27	69	16	24	22	41	48	43	22	39	54	435
3	9	13	0	8	44	14	23	25	36	33	42	24	43	53	367
4	12	9	32	0	35	24	31	19	47	29	48	26	41	56	409
5	7	16	41	9	0	32	13	17	42	45	72	45	64	56	459
6	11	14	28	21	28	0	11	13	37	42	39	38	56	47	385
7	7	14	24	23	46	17	0	16	42	43	47	26	43	51	399
8	8	13	35	26	37	19	23	0	43	54	25	43	56	62	444
9	21	19	26	31	35	17	27	15	0	55	34	46	44	34	404
10	13	23	23	34	37	21	23	32	47	0	43	56	68	74	494
11	15	27	24	30	37	25	29	26	49	58	0	48	56	71	495
12	13	24	31	36	43	25	25	34	42	46	31	0	51	64	465
13	12	26	25	31	47	23	26	35	45	37	26	36	0	62	431
14	11	24	35	33	41	26	42	27	41	32	22	23	31	0	388
sum	151	233	369	326	573	282	318	300	551	554	528	465	637	751	

According to table 5 and figure 8 shows that the OD matrix, the initial site 1 tangkeng station located at the station, Overall, the site behind the line 3,5 and higher number off site, while the number of passengers in the car more on these sites. In these sites should strengthen the relevant management measures, such as increasing the fence or arrange the relevant personnel to carry out supervision, to prevent the safety problems associated with the congestion caused by passengers. In-depth analysis of the OD matrix, some sites are clearly non-matching between the number of people get on the bus and the number of people get off the bus, such as the South Lake Industrial Zone, the number of people who get on the bus are more than the number of people get off the bus. This is mainly due to the larger population density, mainly residential area, more people take the bus, but these areas are not related to consumer spending, such as tourism, shopping, etc. The Universe Park, treasure and farmers market, the liquidity of the regional population is large, strong consumer appeal, so the car number and the number of passengers are more. But some stations such as the Shahu market, the area and the number of passengers on the small area is mainly due to the convenient traffic, while the distance from the residential area near, travel population mainly to residents, through other means of transport means of transport, such as bicycles, cars etc. Therefore, it can reduce the density of the bus station in the area, add the parking position of the bicycle or car, so as to effectively carry out traffic control and prevent the occurrence of traffic jam.

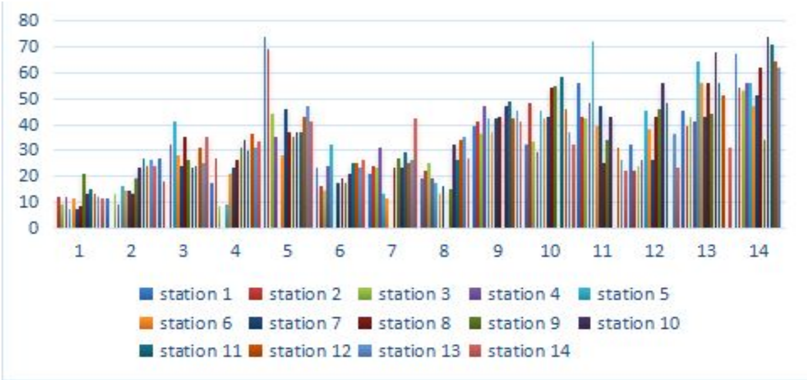


Figure 10 Hop on and off the site number statistical figure

5.4 Dynamic Evaluation of bus station, subway stations service levels

The problem is to establish a mathematical model, to complete the Shenzhen bus station, subway station service level of dynamic evaluation, according to this problem, this paper established the site selection including site suitability that is reasonable, as well as real-time data on the site dynamic evaluation model, the model was constructed as follows.

5.4.1 Establishment of model five

(1) Model building process

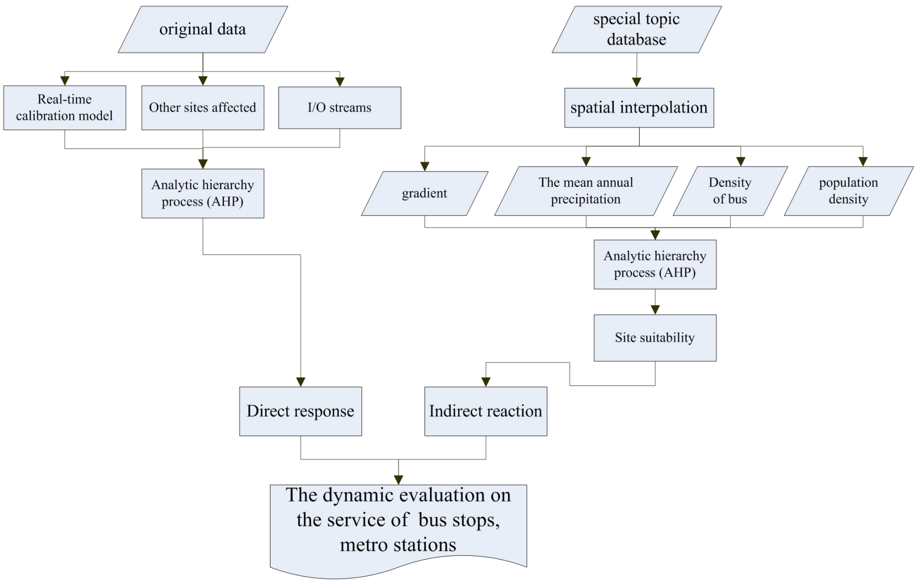


Figure 11 Site service dynamic evaluation technology roadmap

(2)Indirect response: bus station site evaluation model

The location of the bus station to a certain extent determines the existence of the site, a site is not suitable for public transport sites, for example, in the steep gradient or sparsely populated area is bound to have a negative impact on the service[5]. In this paper, we consider the main factors that affect the site selection, use AHP and GIS geographic information system to construct a suitable evaluation model of the location.

According to Wang Lin and Chen Dapeng in the "bus station site selection analysis and fuzzy evaluation" of the bus station for the suitability of the site[2]. In this paper, the suitability of the bus station during the evaluation process to select a slope, precipitation, population density and buses density evaluation, the right to use the analytic hierarchy process re-determination. Wherein each index is shown in table 6 judgment matrix.

Table 6 location suitability index judgment matrix

	Slope information	Precipitation	Density of population	Bus density
Slope information	1	3	1/5	1/7
Precipitation	1/3	1	1/7	1/9
Density of population	5	1/7	1	1/3
Bus density	7	9	3	1

According to the judgment matrix, the weight of each index is obtained by using the analytic hierarchy process, which is shown in table 7:

Table 7 The suitability index weight

Index	weight
Slope information	0.1101
Precipitation	0.0449
Density of population	0.5741
Bus density	0.2709

Maximum eigenvalues: 4.0785, and CI= 0.0262, CR= 0.0294 to meet the requirements.

According to the statistical yearbook of Shenzhen City Bureau of statistics, combined with the satellite images of Shenzhen City, the artificial vector is obtained after the spatial interpolation, which is based on the gradient information of Shenzhen, Shenzhen precipitation, Shenzhen population density, and Shenzhen bus density distribution as shown in figure 12,13,14,15.

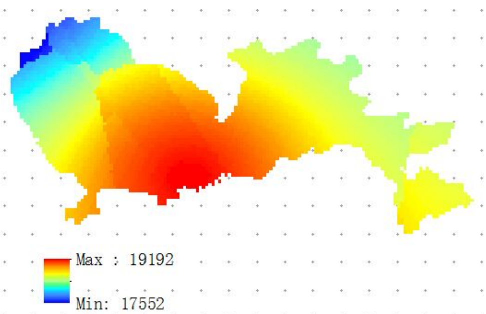


Fig. 12 Annual average precipitation distribution in Shenzhen City

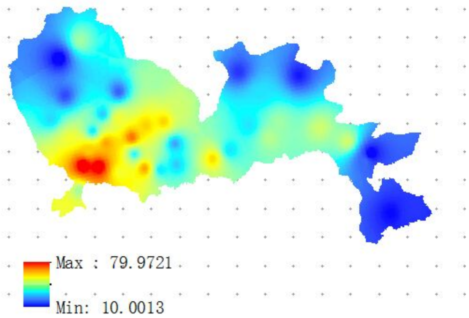
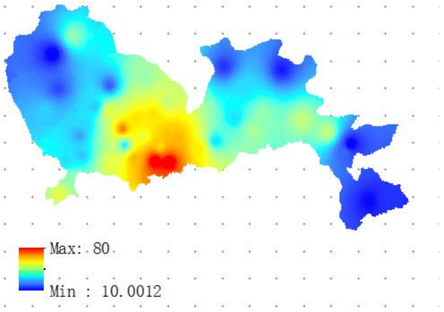


Fig. 13 Population density distribution map of Shenzhen City



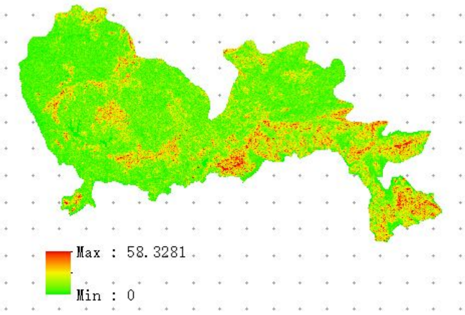


Fig. 14 Density distribution map of Shenzhen City   Fig. 15 Shenzhen ground slope distribution map

Each index data relevant provisions based on the national classification and quantification of the results obtained in accordance with table 8 classification.

Table 8 The suitability rating scale

weighted value	Grade
<5	Very inappropriate
5-15	Inadequate
15-25	More appropriate
>25	Very suitable

We get the Shenzhen bus station site suitability evaluation grade distribution map, as shown in figure 16.

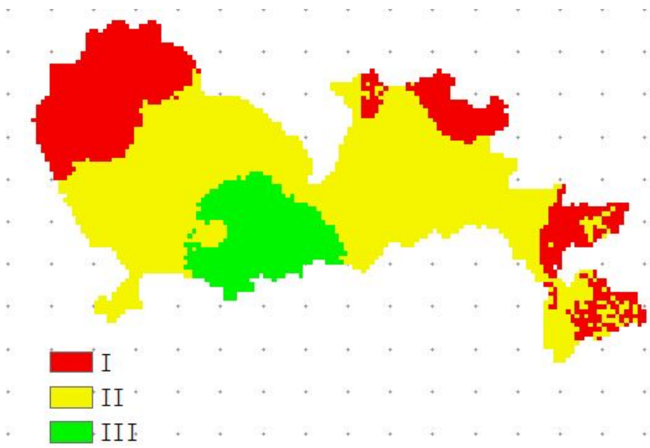


Figure 16 Site suitability level distribution maps

Which we quantified based on site suitability maps for each grade standards, quantization table as shown in table 9:

Table 9 Quantitative table of suitability rank

Distribution rank	Quantization（κ）
Very suitable	1
More appropriate	0.67
Inadequate	0.33
Very inappropriate	0

(3)Direct response: A dynamic evaluation model of the site real time data

We construct the public transportation site indirect display of service evaluation and the rationality of the site location before, we make a direct response to the level of service level of the dynamic evaluation model construction. Model of evaluation index mainly includes two aspects, one is on real time index, on the other hand is I/O flow ratio, the first indicator is the i site in accordance with the operating reliability of punctuality updated vehicle data obtained, its meaning in the pure running time k i punctuality site vehicle reliability. This index can be divided into two parts, one is a pure real-time indicators of punctuality, another part of the other sites affected by the reliability index.

A.pure punctuality indicators in real time

The mathematical model of the index is:

$$PY_{k\to i} = P\{TD_i - T_0 \in [\theta_1, \theta_2]\} \qquad (16)$$

According to the static punctuality reliability model to define the time range of parameters. Here rules： $\theta_1 = 0$, $\theta_2 = bT_0$ Here said： $\theta_2 = bT_0$ as large deviation running on time. To sum up, the pure punctuality reliability evaluation index in evaluation model under real time can be expressed as:

$$PY_{k\to i} = P\{TD_i - T_0 \in [0, bT_0]\} \qquad (17)$$

In the formula, the value of b is fixed with the actual situation.

B.affected by other sites of the reliability index

These indicators can be passed on the reliability of other sites to evaluate the reliability of real-time evaluation of the weight of the right to express other sites on the current impact of the current site, the site is closer to the site, the greater the weight it get.Graph shows the other site on the current site reliability index of the influence diagram.

According to the above diagram, the reliability index of other sites can be expressed as:

$$\alpha_1 \bullet PY_{k_1\to i-1} + \alpha_2 \bullet PY_{k_2\to i-1} + \cdots + PY_{k_{i-1}\to 1} \qquad (18)$$

Which is $\alpha_1, \alpha_2, \cdots, \alpha_i$ the weight coefficient; $PY_{k_1\to i-1}$ is K1 moment I-1 site real-time reliability index, $PY_{k_2\to i-2}$ is K2 moment I-2 site real-time reliability index, $PY_{k_{i-1}\to 1}$ is Ki-1 time 1 site real-time reliability index.

According to the expressions of the 17 can be seen in the evaluation by the influence of other sites, the need for real-time reliability of pure punctuality value at any time according to the I I-1 site before the site is calculated. The correlation calculation of K values is given below.

In this paper, the data is defined as a time window, and other sites on the reliability of the current I site data sources should be derived from the same time window, that is the same as the 5 car, so in consideration of k_{i-1}, k_{i-2},..., k_1 time, should be based on the same data for the same 5.cars.

When calculating the k_{i-1} moments of the I-1 station, the k_{i-1} moment =k moment -t_{i-1}, where t_{i-1} represents the average running time of the 5 cars in the same time window as the I-1 station to the I site.

$$T_{i-1}=\frac{\sum_{n=1}^{5}(t_i^{\,n}-t_{i-1}^{\,n})}{5} \qquad (19)$$

Based on the above analysis, the dynamic evaluation model can be expressed as,:

$$PY_{i,k}=\alpha_0 PY_{k\to i}+\alpha_1 PY_{k_{i-1}\to i-1}+\alpha_2 PY_{k_{i-2}\to i-2} \qquad (20)$$

Which $\alpha_0,\alpha_2,\cdots\alpha_i$ as the weight coefficient; PY$_i$, K said K moment I site service evaluation index, for other sites on the current site, the site is close to the current site, the site, the impact on the current site is more, that is, the weight coefficient is greater. In this paper, the weight coefficient of the model is$a_0=(1/2)^1$,$a_1=(1/2)^2$,$a_{i-1}=(1/2)^i$, which is brought into the weight coefficient, so the dynamic evaluation model of the comprehensive site is:

$$PY_{i,k}=\frac{1}{2}PY_{k\to i}+\frac{1}{2^2}PY_{k_{i-1}\to i-1}+\frac{1}{2^3}PY_{k_{i-2}\to i-2},\ldots,\frac{1}{2^6}PY_{k_i\to 1} \qquad (21)$$

D.Evaluation model construction of I/O flow

The so-called I/O flow in here refers to the ratio of the number of people on the train and the number of people who get off the number, the number of people and the number of people who are equal and two in a reasonable range, the site layout is more reasonable, in this we use a mathematical formula that is:

$$C=\begin{cases}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{I}{O}\right)_i (O>I)\\ \frac{1}{n}\sum_{i=1}^{n}\left(\frac{O}{I}\right)_i (I>O)\end{cases} \quad (I,O\in[\sigma_1,\sigma_2]) \qquad (22)$$

Which δ_1,δ_2 refers to the standard range of the number of people on the train or the number of people to get out of the way, it is generally adopted a flexible way to deal with the common way of the value of the standard. If the value of I and O exceeds the maximum load, or at least one person is not in the range, the O value is set to zero.

(4) the dynamic evaluation model for the final site service is:

the construction process(1) is similar to the three party's judgment matrix is shown in table 10:

Table 10 Dynamic evaluation of each index judgment matrix

	The suitability of the location	the site on time	I / O streams
The suitability of the location	1	1/3	1/3
the site on time	3	1	1
I / O streams	3	1	1

So the weight of the three are:

Location suitability: 0.1420, the site is on time: 0.4290; I / O flow: 0.4290

So the final dynamic evaluation model for the site:

$$\zeta=0.142*\kappa+0.429*(\frac{1}{2}PY_{k\to i}+\frac{1}{2^2}PY_{k_{i-1}\to i-1}+\frac{1}{2^3}PY_{k_{i-2}\to i-2},\ldots,\frac{1}{2i}PY_{k_i\to 1})+0.429*C \qquad (23)$$

5.4.2 Solution of model five

In this paper, regard the M322 line as the research route, the evaluation of the site is 6, the evaluation time is 10:15, according to the Shenzhen Municipal Transportation Bureau released the bus schedule, this paper is a normal bus interval positioning 6-10 minutes, according to the GPS data combined with IC card data obtained in the 8:00-10:15 interval of 5 vehicle data, table 12 shows the actual arrival time interval table.

Table11 Actual vehicle arrive station timetable on M322-6 site

station6	10:15	10:06	9:55	9:43	9:35	9:25	9:16	9:05	8:59	8:50	8:39
station5	10:08	9:57	9:46	9:37	9:27	9:18	9:07	9:02	8:51	8:41	8:31
station4	9:59	9:48	9:38	9:30	9:20	9:08	9:03	8:52	8:42	8:36	8:25
station3	10:03	9:57	9:48	9:21	9:10	9:05	8:53	8:45	8:38	8:31	8:22
station2	9:59	9:51	9:23	9:13	9:06	8:55	8:47	8:39	8:33	8:26	8:21
station1	9:52	9:24	9:15	9:08	8:57	8:48	8:40	8:32	8:29	8:24	8:19

Table12 Actual vehicle arrive station time intervals on M322-6 site

station6	9	11	12	8	10
station5	11	9	10	9	11
station4	12	10	12	5	11
station3	11	5	12	8	7
station2	11	8	8	6	7
station1	8	8	3	8	5

After investigation, we can see that the area of the planned departure time of 10 minutes, so before and after the vehicle in the actual arrival time interval of each site in the 6-10 minutes to think reliable. For example, the 6 sites, respectively, 3, 2, 3, 4, 5, 3 actual arrival time interval reliability. The site for the evaluation results on time:

$$PY_6=3/5=0.6, PY_5=3/5=0.6, PY_4=2/5=0.4, PY_3=3/5=0.6, PY_2=4/5=0.8, PY_1=5/5=1$$

So the real-time evaluation value of each site is:
6 site after the addition of other sites in the 10:15 service evaluation results for:

$$PY_{6,1}=1/2*0.6+1/2^2*0.6+1/2^3*0.4+1/2^4*0.6+1/2^5*0.8+1/2^6*1=0.5781$$

Other site analysis method and 6 similar sites, then according to volume flow. To evaluate the weighted method on the whole line of L moment 10:15 "bus station.

The I/O flow evaluation results for 6 of the site are:

Table 13 People counting on M322-6 site

	1	2	3	4	5	6
Number of people off	9	12	14	6	9	8
Number of car	13	6	12	12	7	13

So, C=1/5（9/13+6/12+12/14+6/12+7/9+8/13）= 0.7885
At the same time according to table 9 the location of the region suitability for 0.67
The overall evaluation results for the 6 points are:

$$\xi = 0.142 * 0.667 + 0.429 * 0.5781 + 0.429 * 0.7885 = 0.6810$$

The results of real-time evaluation of the route are shown in table 14, using the same method as for the 6 site:

Table 14 Final evaluation results of each site on M322 route

Site name	evaluation of estimate	Site name	evaluation of estimate	Site name	evaluation of estimate
1	0.6713	15	0.5187	29	0.5327
2	0.5789	16	0.6327	30	0.5267
3	0.7168	17	0.6893	31	0.6326
4	0.6710	18	0.6716	32	0.4256
5	0.5718	19	0.7162	33	0.6231
6	0.6810	20	0.5827	34	0.6627
7	0.6351	21	0.6528	35	0.5926
8	0.6189	22	0.4527	36	0.6327
9	0.5178	23	0.6104	37	0.6251
10	0.6128	24	0.6418	38	0.5271
11	0.6123	25	0.5832	39	0.4271
12	0.5728	26	0.6314	40	0.5719
13	0.6085	27	0.5967	41	0.6056
14	0.6715	28	0.7826	42	0.6318

So the bus site that is not reasonable arrangement is 39,38,32, the site is mainly due to the location for business to go nearby, people flow, the number of people on the bus and get off the bus more, obviously more than the effective load, and the crowd caused by the follow-up site easy to occur late, to extend the time, causing some negative influence, proposed in these areas to add bus station, and the bus to divert, so that passengers civilized car.

6.Advantages and disadvantages of the model

When data preprocessing is carried out, the SOM is used to detect the error data of the neural network, and the classification results will be very different. Therefore, the accuracy of the step size is directly affected by the setting of the results, so the relevant experiments should be carried out to determine the reasonable step size. In this paper, the selection of reasonable step size is carried out several times, and the final step is set to 10. The results show that this step can be used to classify the data and the classification accuracy is high, the error data can be easily obtained at the same time. In the construction of the OD matrix, the main consideration in the process of model construction is the creation of a single line OD matrix. Based on the theoretical basis is the two site model, but the reality is that the travel of residents vulnerable to many factors, the next place on the car and it is not before time to get off the site, we should add other models related correct results which are only more reliable accuracy. The advantage of the model construction is based on the characteristics of the original data to build the OD structure, based on the tail number, time, latitude and longitude, the maximum use of the original data. In the construction of the public transportation station dynamic service level evaluation model, three aspects have been considered, but the actual situation is the site's service evaluation index has much, this paper does not have a thorough discussion, and the I/O flow of the judgment standard is no practical reference. The merits of the model are that the direct response index and the indirect response index of the site service, and the influence of other sites is also considered in the direct reaction index, thus the results have higher accuracy.

7.References

[1] Yu Fan, Shoubin Chen. Reasonable evaluation of the distribution of Zhengzhou subway station[J]. Science & Technology Information,2013(26).

[2] Deming Wang, Jun Wang. Risk analysis of coal spontaneous combustion based on the unsupervised neural network[J]. Journal of China Coal Society | J Chin Coal Soc,1999,24(2).

[3] Yongle Quan, Jifu Guo, Huimin Wen, Jianping Sun. Theory and technology research on the real time dynamic evaluation of urban road network[J]. ENGINEERING SCIENCE,2011,13(1).

[4] Sheng Zhao. Research on dynamic evaluation and control strategy of freeway traffic[D]. South China University of Technology,2013.

[5] Kefu Tang. Genetic algorithm for traffic flow assignment and scheduling[D]. Dalian University of Technology,2004.

[6] Jing Xie, Zongming Wang, Yangsheng Zhang, Huijuan Feng. Dynamic evaluation of regional highway transportation development[J]. Journal of Northwest University (Philosophy and Social Sciences Edition,2013,43(1).