Attraction and radiation range: The impact of spatial awareness on the choice of business district in Shanghai

Kelin Li[1], Yina Li[2], Hong Yin[3], and ChangBo Wang[4]
[1]School of Computer Science and Software Engineering, East China Normal University, Shanghai,China
[2]Nankai University, Tianjin, China
[3]School of Computer Science and Software Engineering, East China Normal University, Shanghai,China
[4]School of Computer Science and Software Engineering, East China Normal University, Shanghai,China

The measurement and evaluation of the attraction and radiation range of business districts in Shanghai is examined in this paper, the city which had experienced rapid development of public transport in recent years. Convenient public transport has changed our shopping habits, which become more irregular, and business districts have different modes of development. However, the research methods based on statistics, experience and sampling have not been able to adapt to the era of big data. Here we establish an attractiveness model focus on Shanghai's business districts by mining traffic data and apply a new method with customer-oriented to division scope of radiation range. Visual analysis is used to draw the conclusion that 1)Spatial awareness is no longer the important factor negative influence shopping, 2)the factors that affect the attractiveness already from single factor (commercial area) to mixed factors (commercial area, reputation, goods grade, etc.), and 3)psychological expectations will make customer behavior more regular. We evaluate these studies and anticipate that this work can provide decision-making assistance to governments and retail businesses.

***Index Terms*—Business district, Visual analysis, Radiation range, Attraction model.**

I. Introduction

BUSINESS district is a geographical concept generated by aggregation of retail stores. Business district in a broad sense is a business concept which made up of shopping centers, department stores and related infrastructure, and in a narrow sense is a geographical concept with several stores.This article explores the attraction and radiation range of business districts (broad sense) in Shanghai by analyzing and mining urban public transport data and business data.

The concept of attraction of business was proposed by Reilly in 1930s, and Huff put forward the attraction model of the business district. Until now, most of the relevant research is an extension of these studies. However, different regions have their own economic level, which result the conclusions of business district research to not be universal.Now, the rapid development of public transport makes commerce more prosperous and intensive in Shanghai, and many non-commercial functions (e.g. catering, banks, government organs, etc.) have gradually moved over from individual area to the business districts. Like densely populated cities, business districts have become the center of business activities in Shanghai.

As the business activity is people-centred, a considerable part of the research on the business districts is conducted through questionnaires and sample surveys. Many valuable and interesting conclusions have been found in the existing research, but we need to use new ideas to study the business district and business activity in the era of big data. Due to balance development in Shanghai economy and commerce, customer characteristics, market competition and other factors have little discrepancy. If we fully follow the existing conclusions, we find that it does not always satisfy our research. Therefore, this work is an exploration of business district under circumstances of big data.

Manuscript received December 1, 2012; revised August 26, 2015. Corresponding author: M. Shell (email: http://www.michaelshell.org/contact.html).

Business districts in Shanghai become "energetic organs", comprised of entertainment, shopping and leisure, and the public transport system (PTS) connects the whole body. The objective of this paper is to measure the attractiveness of the business district in Shanghai, with an emphasis on building an attraction model, a summary measure of centre attraction for potential customers, and a new angle about commercial radiation range. The methodological approach is based on information visualization, visual analysis, statistics and evaluation for a variety of different models. On the other hand, the specifics of PTS introduce an original element to the study of attraction. In addition, the variables used in this research are atypical of the most of big cities, that the process can be repeated in similar research. Thus the results of the study can be used for broader scope, but only suitable for big cities at this stage.

In the first place, our contribution is mainly focused on the following points:

- Put forward a attraction model of business district with more accurate and practical, which could calculate the degree of attractiveness from one business district to anywhere.
- Propose a method to division and drawing the radiation range of business district, and explain its advantages by comparison with the human-centered method.
-

And this paper is trying to answer the following questions:

- Q1: Is the attractiveness of business districts can express in numerical terms? Is this the existing result and model represents the status quo of business development of Shanghai?

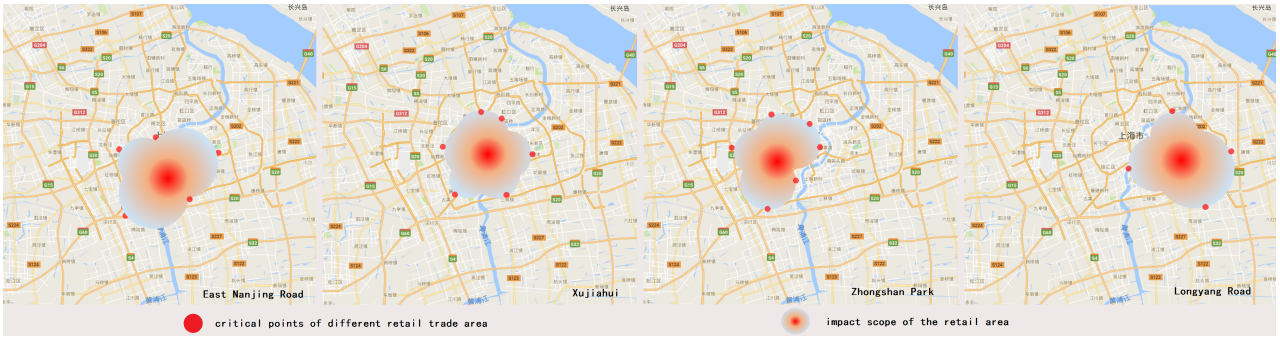

Fig. 1. Radiation range based on the resulys calculated by Reilly Rule of four different types of business districts in Shanghai.

- Q2: Is the attractiveness depends on a variety of factors with high degree of influence? Is there any advantage compared with previous research results? and
- Q3: Is there a way to represent the radiation range of a business district with another view? Whether it has good application value if it exists?

This paper is divided into four sections. After the data and theoretical introduction, the crowd classification and behavior based on traffic data are discussed. The methods and attraction model used in the analysis are then designed and discussed, followed by the case studies and user surveys were conducted. In addition, the summary of the overall work and some limitations were finally proposed, and ideas for future research are discussed.

II. Relate Work

III. Business District Theory and Data

A. Attraction Model

Business district is the range of retail transactions, in this paper, it is the space range of shopping malls and department stores. The most widely used theory of business district are Reilly Rule and Huff Model, which are the earliest model.

Detailed business information is difficult to obtained in most cases, which results a challenge about how to invest for enterprise, and Reilly Rule first provides theoretical guidance. Reilly believes that commerce also has the characteristics of mutual attraction, and then he put forward the method of the critical range of attractiveness of the business district based on the Law of Universal Gravitation. However, Reilly Rule has great limitations, and requires more stringent preconditions. A Reilly Rule calculation of business districts in Shanghai is then studied, which has complex errors due to the differences and complexity of business districts and customers behavior. The Reilly Rule we use IS as follows:

$$D_{ab} = d/(1 + \sqrt{P_a/P_b}) \tag{1}$$

where D_{ab} is the radiation range of business district A, d is the distance between A and B, and P_a, P_b are the number of customers in this two business districts respectively.

Reilly Rule can establish the radiation range of business district, but due to it does not consider the customer uncertainty in the choice led to significant errors, and the results are shown in Fig.1. The four business districts shown in this figure have completely different features: Xujiahui also has the position of transportation hub and financial center, Nanjing East Road has a large number of small shopping malls and attracts a large number of tourists, Zhongshan Park is an important transportation hub and Longyang Road connects the suburb and the urban area of the Pudong region. In addition, the premise of using Reilly Rule are as follow: (1) the traffic are similar; (2) the hardware properties are similar; and (3) the customer (population type and flow) are similar. In this paper, we analyze the business districts that meet these requirements.

Based on Reilly's research, Huff conducts work from the customer's point of view, quantifying the attractiveness of a business district as a probability value, which can indicate the degree that one people going to a business district. However, due to the limitations of the times, Huff only proposed a few factors for calculation, some of which are not suitable in the present society like space distance or commercial area.

Huff believes that the fundamental reason for affecting the size of the business district is that consumers engaged in shopping behavior of the psychological identity. Huff Model is as follows:

$$P_{ab} = \frac{\frac{S_j^{\mu}}{T_{ij}^{\lambda}}}{\sum_{j=1}^{n} \frac{S_j^{\mu}}{T_{ij}^{\lambda}}} \tag{2}$$

where P_{ij} is the probability where region i to the business district j, S_j is the attractiveness of business district j and the resistance that people in region i to the business distrrict j is expressed by T_{ij}. In addition, μ and λ are the revised values estimated on experience, n is the count of business districts that have competitive relationships.

We can get the value of the attractiveness of the business district by Huff Model which is similar to the actual value. However, if n (the business districts that have competitive relationships in calculation) is too large, so the accuracy will be reduced with the size of the n which is shown in Fig.2. In addition, more factors will have impacts on the attractiveness of the business district in the current economic context, and the resistance is not only based on space distance. In this paper, we extracted a dozen factors that may have an impact on the attractiveness and resistance of the business district after interacting with a marketing manager, and a correlation study (5.1) and a attraction model (5.2) of business district are then carried out and designed.

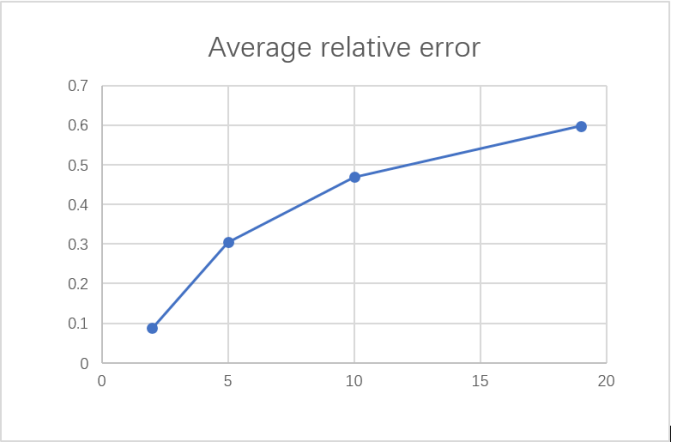

Fig. 2. The calculation accuracy of the number of business districts which have competitive relationships in Huff Model

B. Input Data Set

A large amount of data and seldom subjective factors are used in this paper, and the dataset is as follows:

RFID public transportation card data records passenger journeys in the Shanghai public transportation system over a typical day. This includes the metro system and the public bus network, where passengers use their personalized RFID cards to tap on card readers. The card reader system records every tap in/out action, which contains time, address, the cost and the mark of tap in/out. This dataset is about one billion and six hundred million lines in a total of four months.

Budiness districts data includes all the Central Business District (CBD) in Shanghai, which contains the commercial area, malls rent and etc. Commercial area is one of the most important criteria for measuring conditions, we get this data through the official website of each shopping mall or department store.

Travel time cost in this paper refers to the time that customer spend from a place to a CBD. The PTS in Shanghai is well-developed, so the spatial distance will cause error to the calculation results. Therefore, we introduce the concept of travel time cost instead of spatial distance as a part of the resistance factors. But it's hard to get the exact value due to each people is unique.

At first we crawled the official website data of the Shanghai Metro, and some error is found after calculating in the model. In official point, time always have to considers the worst case in PTS which causes the calculation overflowed. Therefore, we statistically analyze PTS data, and sort all the data for any two locations. As some people stay in station for some reason, we extracte the first 80 percent and calculated the average, which is the travel time cost of this two locations.

Real attractiveness data is a value obtained by statistical analysis and calculation. A research of people classification and shopping behavior is used to extract the customer group, a probability value is then obtained through the analysis of the classification and the traffic.

In addition, some of the data used in the calculation like the business district level, traffic station level and so on are from the government public data or network data.

IV. People Classification and Shopping Behavior

Due to traffic jams and sparse parking spaces at the peak of the trip, most people rely on PTS, which is the artery of a city. This section focuses on exploring and analyzing human mobility behaviors from the passenger RFID card data in a PTS with a family of analytical tasks. In particular, we present an interactive visual analytics system that deals with the major challenge discovering interactively behaviors of different groups and displaying changes of the time-series traffic flow.

The metro system plays an important role in the modern public transportation system. Salaried people rely on the metro, due to its convenience without the traffic jams. The potential movement patterns can be funded from the massive passenger RFID card data. The changes of the traffic flow with time need be displayed for a better understanding of movement trajectories. Especially, we focus on the traffic flow of office workers and analyze the residence location and the work location according to RFID card data. Our goal is that reveals features of the metro traffic flow by combining with visualization techniques. Our main analytical tasks are shown as follows:

T.1: How to explore and display the changes of the traffic flow with time in different subway stations? This allows end users to understand the features of the traffic flow between different lines.

T.2: Measuring the traffic flow trend of between subway stations, which has large passenger flow? These subway stations reflect the main features of the subway network and movement patterns of office workers.

The above tasks mainly focus on analyzing the traffic flow of the subway system. The solution to these tasks will empower users to select different subway lines and subway stations at a given time period. The following tasks aim at exploring characteristics and movement patterns of different groups.

T.3: Discovering residence location and work location of office works. The location information provides a better way to understand the movement patterns of office workers.

T.4: Distinguishing different groups from the RFID card data according to its own movement patterns. For example, we can infer a person whether work overtime or not according to the frequency of going to work location.

Millions of records contain trajectories of different groups. In this paper, we focus on analyzing mobility behaviors of office works. First, trajectories of office workers need be acquired from passenger RFID card data. These office workers who take subway usually have the same journey. All journeys of an office worker have the same origin and destination during the working days. Finding office workers is modeled as follows:

$$W = \{w_i| \ |S_{in}| \geq 4, |S_{out}| \geq 4\} \tag{3}$$

where W is the set of all office workers in Shanghai, for each office worker w_i, there must be at least continous 4

records during working days with the same origin station and the same destination station. Here, we define normal and abnormal office worker for further analyzing the metro flow of office workers. The normal office worker is who works for at most consecutive 5 work days. The abnormal office worker is who works at least consecutive 6 days, especially works on Saturday or Sunday or both as well.

Office workers are more willing to take the metro, due to its convenience and low-carbon lifestyle. Furthermore, at the peak of the trip, it is difficult to take a taxi or find a parking space in Shanghai. In order to explore movement patterns of office workers, we infer the residence location and the work location of office workers based on our assumption. We suppose that office workers often choose the nearest subway station from his/her home as the origin and the nearest subway station from his/her work location as the terminal. This makes sense in line with personal experience. According to this finding, the work location and the residence location are modeled as follows:

$$L_r = \{L_r^i |\ 6:00 < T_{in} < 9:00, Dur_i \geq 5\} \quad (4)$$

$$L_w = \{L_w^i |\ 17:00 < T_{in} < 20:00, Dur_i \geq 5\} \quad (5)$$

where L_r is the set of residence locations, and L_r^i is the residence location for the ith card, if its tap-in time T_{in} is between 5:00 and 10:00 a.m. and it appears for at least consecutive 5 working days with the same station. Where L_w is the set of work locations, and L_w^i is the work locations for the ith card, if its tap-in time T_{in} is between 17:00 and 20:00 and it appears for at least consecutive 5 working days with the same station. Once the residence location and the work location are inferred, more features of different groups could be found.

Hence, the flow between two stations was displayed in flow snapshot view. Here, the thickness of the line indicates the size of flow. As illustrated in Fig.3, the line with the color yellow, metro line 2, carries more traffic loads that other lines. Besides, a pie chart was used to display flow changes of office workers that appear at least 5 continues working days at the same residence and work location. The blue part indicated the number of going into this station; the red part indicated the number of going out this station. Then, distribution of residence location and work location was shown in Fig. 3. The region with many pie charts that have a larger number of going into this station probably contained certain residence communities. The region with pie charts that have a larger number of going out this station probably contained certain sci-tech parks.

The research in this chapter provides support for the validation of the computational results of the attraction model.

V. The Attraction Model of Business District

This section introduces the design and analysis of the attractiion model of the business district.

A. Correlation and Coefficient

The significance and correlation of the calculated variables proposed by Huff are shown in TABLE I. The correlation

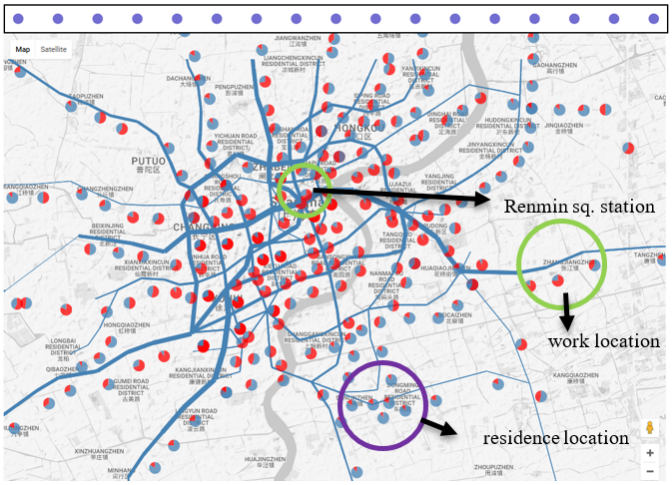

Fig. 3. Flow snapshot view from 6:00 pm to 6:30 pm. This view shows the metro traffic flow at a given time and pie chart indicates a number of going into and out this station.

coefficient for time cost is -0.489 and for commercial area is 0.149, and both have significant correlations. After discussing with the manager, the reason we think for the small correlation coefficient is follows: The PTS in Shanghai is high developed, which result the space distance and time cost have smaller impact on people's travel; The research samples are all central business districts, which have similar commercial areas.

TABLE I
THE CORRELATION OF TIME COST AND COMMERCIAL AREA

Variable	Probability	Time cost	Commercial area
Correlation	1.000	-0.489	0.149
Significance	.	.000	.000
Sample	5111	5111	5111

As shown in formula 2, μ and λ are the revised values estimated on experience in Huff Model. We invite experts in related fields to help us give two values as a subjective index to make a better comparison. Meanwhile, we obtain the correlation coefficient value as an objective index through big data analysis. In addition, a constraint is proposed as $\mu + \lambda = 2$ for better calculation. We got two sets of index numbers after normalization and amplification, and then we calculate the attractiveness of the business districts and the range of radiation. The two adjustment index values are like TABLE II.

TABLE II
TWO ADJUSTMENT INDEX VALUES

	subjective index		objective index	
Regulatory factor	$\lambda(t)$	$\mu(s)$	$\lambda(t)$	$\mu(s)$
Value	1.5	1.2	0.454	0.120
Normalized	0.556	0.444	0.791	0.209

The results are shown by visualization techniques in Fig.4, where the regulatory factors were normalized and magnified.

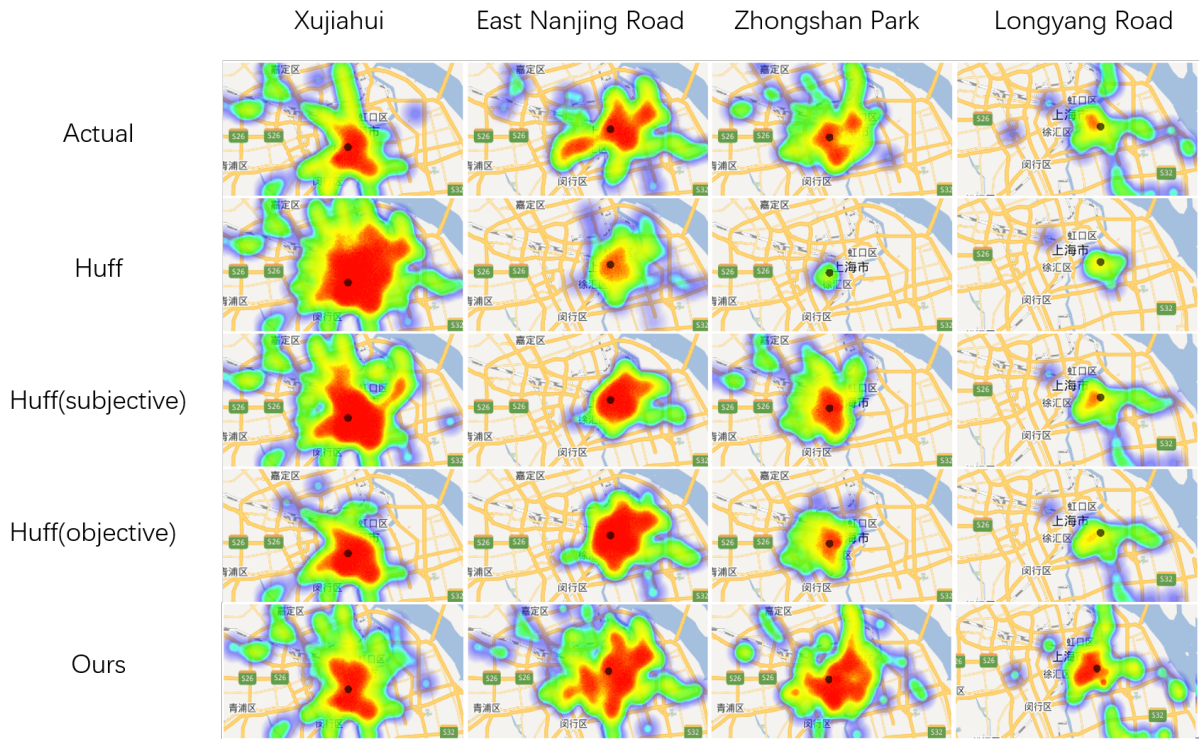

Fig. 4. A visual view of the radiation range, which are calculated from four models and raw data.

Through the visual comparison of the model calculation results, we can clearly see that the precision of the exponentially adjusted model has been significantly improved, but the two index adjustment methods do not have obvious advantages and disadvantages. We think this is due to Huff Model uses only commercial area and space distance for calculation, and in practice, the determination of charm and resistance is more complicated. In addition, in order to get more accurate results, we train the data using machine learning and get a set of values of influence factors. The possible influence factors and training results are shown in Table III.

B. Model Design

The basis of our model design is also the law of universal gravitation. A larger commercial area and a wider variety of products can attract more consumers, which need to be considered in related research. However, in this paper, the samples are the central business districts and all are very prosperous, which result a smaller differences in attribute. Meanwhile, we can get similar conclusions: Factors such as commercial area and district level have no significant impact on attractiveness, but other ffactors, such as time cost and the count of changeovers, have a much higher impact on selection.

After many validations and analyzes, we propose a model of attractiveness for Shanghai's business districts:

$$Attraction = Grade^{\alpha} * Mall^{\beta} * Area^{\gamma} * Reputation^{\delta} \quad (6)$$

$$Attractive = \frac{Attraction}{Time^{\mu}} \quad (7)$$

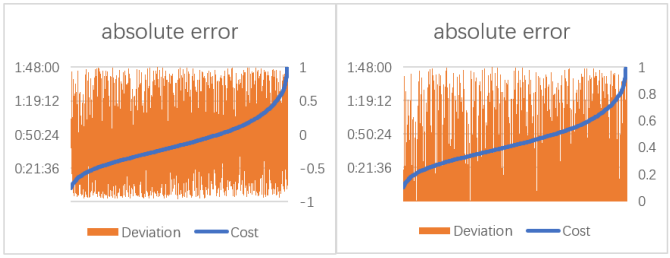

Fig. 5. Absolute error 1

where α, β, γ, δ and μ are adjusted index based on data analysis and mining. We can get $Attraction$ to empress the charm, and $Attractive$ to represent the degree of attraction.

In our opinion, the main resistance factor affecting the customer's choice of business district is the time cost. Although it is influenced by certain subjective emotions (factors such as the distance in the metro map), it is still the most important factor. In addition, the commodity grade, the number of shopping malls, the commercial area and the popularity of the district are the main attractiveness factors.

And the probability of a customer to the business district is:

$$Probability = \frac{Attractive}{\sum_{i=1}^{n} Attractive_i} \quad (8)$$

where i is the number of business districts that affect each other and n is the totle number of samples.

In the current study, it is not possible to verify that the value of the adjustment factor applies to all similar business districts

TABLE III
THE TRAINING RESULTS OF MACHINE LEARNING

	time cost	commercial area	commodity grade	market number	popularity
influence value(%)	27.64570	36.85034	33.19354	28.09620	26.56424

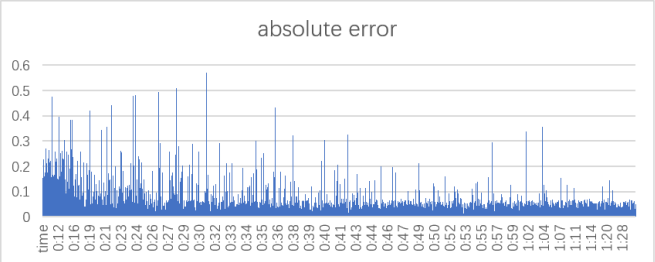

Fig. 6. Absolute error 2

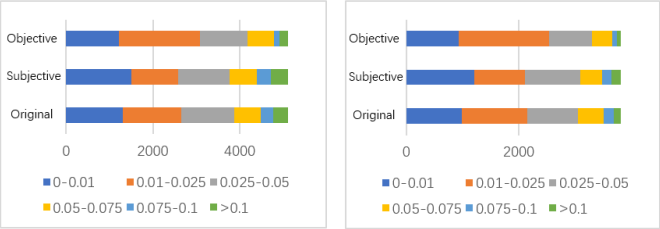

Fig. 7. Absolute error 3

in large cities. Therefore, the research on the adjustment factor we made through a variety of analytical methods is only applicable to Shanghai.

In this paper, we use the method of statistical correlation analysis and the method of machine learning to determine the training factors. We have separately obtained the corresponding values of different factors and conducted a visual comparison study. The result is shown in the Fig.4.

However, we find that there are large errors in the calculation results in some cases (As shown in Fig.6 and Fig.7). After discussion, we think there are mainly two possible causes of these errors, one is the impact of buses on data statistics and the other is the influence of the number of changeovers on the resistance of model accuracy. In order to calculate the attractiveness and radiation range of the business district more accurately, we divided the experimental samples and optimized the resistance value. And then we get the final model:

$$Attractive = \frac{Grade^{\alpha} * Mall^{\beta} * Area^{\gamma} * Reputation^{\delta}}{Time^{\mu} * \sqrt{Turn}} \tag{9}$$

where $\sqrt{Turn}$ is resistance correction value through the transfer calculation. We use the optimization model for calculation, the results of the comparison shown in Fig.8.

C. Deviation Analysis

We use the original model to calculate, and then calculate the average error with the actual value, as shown in Fig.5.

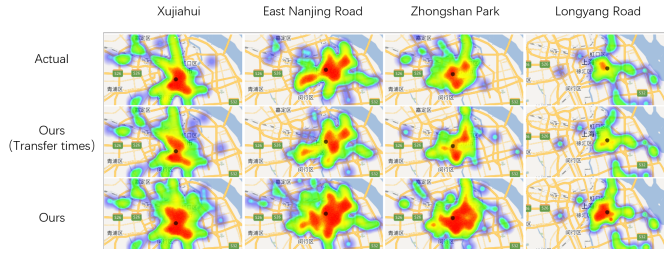

Fig. 8. Comparison of model and real value

Since the result of the calculation is a probability value between -1 and 1, we use the absolute error display.

We can find that in overall, the time cost and the positive or negative of error are no necessary connection. The reason for this result is that the underlying model does not apply to the current study, and after that we also perform an error study on the results of our model (formula 9), as shown in Fig.6

It can be clearly seen in the diagram that the lower the time cost, the greater the error. At the same time, we compare the calculated values of Huff Model with different adjustment indices, as shown in the Fig.7. Among this results, the error in the calculation is less than 0.025 that is much higher than that of the others, while the subjective index adjusted value can get more error than 0.01, which means that the application of objective index adjustment with better accuracy.

In a separate analysis of locations with larger errors, we found that the time cost for most of these sites to go to a business district was less than 10 minutes, which we believe is due to data error. We use the subway credit card data, but do not use the bus data which also occupy a large proportion of public transport. Customers near the business district prefer to take the bus to the nearest shopping district, which leads to errors in the actual probability values we measure. This error is mainly reflected in the over-estimation of the attraction of the nearest shopping district to residents, resulting in a larger error in subsequent calculations. However, we are temporarily unable to solve this problem in our work at this stage. In order to improve the model's applicability again and to optimize and improve it better, we removed the time-consuming sites less than 10 minutes and re-calculated the model. The results of the experiment are shown in the right of Fig.7, of which there are 201 locations with 3819 groups data.

From the figure above, we can see that after we remove the locations with small time cost, the result has less error. Therefore, we think that the data errors are objective, but if we temporarily remove these locations with small time cost, the results will be more accurate.

By plotting the radiation profile with the actual values, we found a characteristic that was neglected in previous studies, which is the number of changeovers has a significant

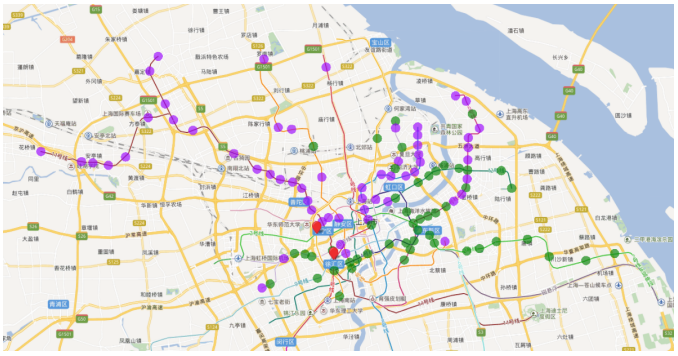

Fig. 9. Transfer times

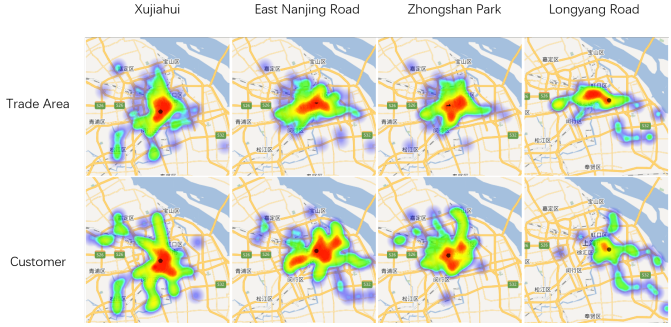

Fig. 10. Real range

impact on the results. After analyzing the statistical results, we find it more attractive to go to the business district with fewer changeovers if the two business districts have similar resistance and attractiveness but obvious and different characteristics. We conducted an in-depth analysis of some of the locations and business districts with the above characteristics (business district: Zhongshan Park and Xujiahui, time cost: difference is less than 5min, which is considered consistent) as shown in Fig.9. Zhongshan Park is the intersection of 2,3,4 lines and Xujiahui is the intersection of 1,9,11 lines.

Through visual analysis, we can clearly see that there is a connection between two locations with the same time cost. Among them, the purple spots are those who prefer Xujiahui, the green spots are those who prefer Zhongshan Park. As can be clearly seen from the figure, the closer geographical position does not mean to be more attractive, and the customer prefers to the business district with fewer changeovers with the same time cost. In other words, the more changeovers, the less attractive(more resistance) to customers. Hence, in our research, we need to consider the number of changeovers, whcih will improve the accuracy.

VI. Case Study

VII. Discussion

VIII. Conclusions and Future Work

mds
August 26, 2015

A. Subsection Heading Here

Subsection text here.

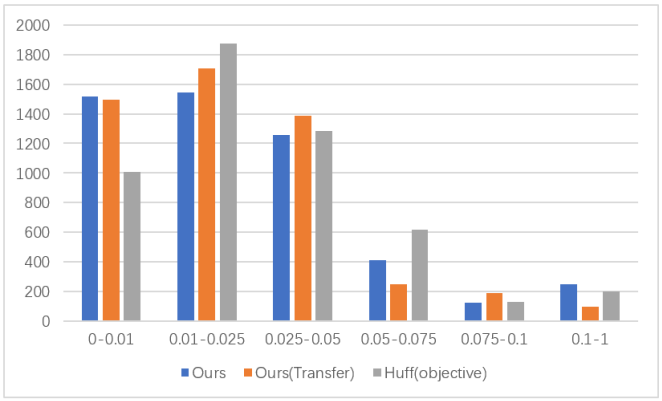

Fig. 11. Error comparison

1) Subsubsection Heading Here

Subsubsection text here.

IX. Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank...