Author Accepted Manuscript. Gong, S., Cartlidge, J., Bai, R., Yue, Y., Li, Q., & Qiu, G., (2018), "Automated prediction of shopping behaviours using taxi trajectory data and social media reviews", in Proceedings 3rd IEEE International Conference on Big Data Analysis (ICBDA 2018): Shanghai, China, Mar. 2018, pp. 117-121.
For version of record, see DOI: https://doi.org/10.1109/ICBDA.2018.8367661

Automated prediction of shopping behaviours using taxi trajectory data and social media reviews

Shuhui Gong*†, John Cartlidge‡, Ruibin Bai*, Yang Yue†, Qingquan Li† and Guoping Qiu§
*International Doctoral Innovation Centre & Department of Computer Science
University of Nottingham Ningbo China, Ningbo China, 315100
Email: {shuhui.gong, ruibin.bai}@nottingham.edu.cn
†Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Research Institute for Smart Cities
Department of Urban Informatics, School of Architecture and Urban Planning
Shenzhen University, Shenzhen China, 518060
Email: {yueyang, liqq}@szu.edu.cn
‡Department of Computer Science, University of Bristol, Bristol UK, BS8-1UB
Email: john.cartlidge@bristol.ac.uk
§School of Computer Science, University of Nottingham, Nottingham UK, NG8-1BB
Email: guoping.qiu@nottingham.ac.uk

***Abstract*—The Huff model is a well used mathematical abstraction for predicting shopping centre patronage. It considers two factors: shopping centre attractiveness, and customers' travel costs. Here, taxi trajectory data (more than three million journeys) and social media data (more than eight thousand customer reviews) is used to calibrate the Huff model for five primary shopping centres in the rapidly expanding metropolitan city of Shenzhen, China. The Huff model is calibrated in two ways: globally, to find the single pair of best-fit parameters for attractiveness and travel cost; and locally, using Geographical Weighted Regression to find the best-fit parameters at each spatial location. Results demonstrate that customer reviews on social media provide relatively high prediction accuracy for weekend shopping behaviours when the Huff model is calibrated globally. In contrast, customer footfall, calculated directly from number of taxi journeys, provides higher prediction accuracy when the Huff model is calibrated locally. This suggests that, at weekends, sensitivity to footfall has greater spatial variance (i.e., customers living in some areas have greater preference for shopping at popular centres) than sensitivity to customer reviews (i.e., regardless of where customers live, positive reviews on social media are equally likely to affect behaviour). We present this geographical homogeneity in review sensitivity and heterogeneity in footfall sensitivity as a novel discovery with potential applications in urban, retail, and transportation planning.**

***Index Terms*—Social media review data; Taxi trajectory data; Huff model; Geographically Weighted Regression.**

I. Introduction

Retail is intrinsic to urban development and planning [1]. There are various methods that can be used to analyse retail trading areas, such as gravity assumptions [2], discrete choice models [3], and logit models [4]. One of the most widely used methods is the Huff model [2]. First introduced in 1964, the Huff model follows simple gravity assumptions and estimates the spatial probability distribution of shopping centre patronage based on shopping centre attractiveness and customers' travel costs.

To be applied effectively, the abstract Huff model requires calibration with real-world data. Traditionally, interviews and surveys were used for calibration. However, these approaches are labour intensive and generate relatively limited and low resolution data; as a result, the predictive accuracy of the calibrated model is diminished. With recent developments in sensing technology and the move to smart city infrastructure, as well as the now ubiquitous proliferation of mobile technology and social media applications, new big data streams are available; offering the opportunity for automated and high resolution calibration, at a fraction of the cost. Here, the Huff model is calibrated for the city of Shenzhen, China, by fusing two data sources: taxi trajectory data, and social media shopping reviews. Taxi data has previously been used for Huff model calibration [5], [6], while social media data has been used elsewhere to delimit trade areas [7]. However, as far as the authors are aware, this is the first time these two data sources have been fused to predict shopping behaviours.

Results demonstrate that social media reviews give greater predictive power when the Huff model is calibrated globally, while taxi journeys give greater predictive power when the Huff model is calibrated locally. Since social media review data is freely available and continuously growing, we suggest that social media reviews offer a powerful new opportunity for predicting retail behaviours. Fusing data sources for automatic prediction of shopping behaviours has the potential for significant impact on urban, transport, and retail planning.

II. Related work

The Huff model [8] is a traditional mathematical method to estimate customers' patronage probability distributions to a set of target shopping centres. There are two factors influencing the probability: attractiveness of each shopping centre, S, and the customer's travel cost to get there, C. Accordingly, the classic expression of the Huff model is:

$$P_{ij} = \frac{S_j^{\alpha_i} C_{ij}^{-\beta_i}}{\sum_{j=1}^{m} S_j^{\alpha_i} C_{ij}^{-\beta_i}} \tag{1}$$

where P_{ij} represents the probability that customer from origin i shops at shopping centre j, C_{ij} is the travel cost from origin i to shopping centre j, S_j is the attractiveness of shopping centre j, and α and β (which are empirically estimated from data) are the parameters associated with attraction and cost variables, respectively. Finally, m is the total number of shopping centres considered.

To calibrate the Huff model, O'Kelly introduced four parameter estimation methods [9]. However, in a previous study, the authors have shown that only two of these work well with the taxi data for Shenzhen [6]. Therefore, here, to calibrate the Huff model only two of O'Kelly's equations are considered:

$$K1 : T_{ij} = exp(\alpha S_j - \beta C_{ij}) \tag{2}$$

$$K2 : T_{ij} = exp(\alpha S_j - \beta LnC_{ij}) \tag{3}$$

where T_{ij} is the numerator of (1).

To make this study comparable to previous work [6], [10], we use both estimation methods, K1 and K2, to fit the Huff model globally (such that there is one best-fit pair of values for α and β). Subsequently, K1 and K2 are applied using Geographically Weighted Regression (GWR) (such that best-fit values of α and β are estimated for each geographic region).

GWR is a geographical method used to discover spatially varying relationships [11]. The general form of GWR is:

$$y_i = \gamma_{i0} + \sum_{k=1}^{m} \gamma_{ik} x_{ik} + \epsilon_i \tag{4}$$

where y_i is the dependent variable at location i; x_{ik} is the k^{th} independent variable at location i; m is the number of independent variables (since the Huff model has two independent variables—α and β—therefore $m = 2$); γ_{i0} is the intercept parameter at location i; γ_{ik} (corresponding to α and β in (1)) is the local regression coefficient for the k^{th} independent variable at location i; and ϵ_i is the random error at location i. In this study, two parameter estimation methods, K1 (2) and K2 (3), are used to fit spatially variant parameters of the Huff model via GWR.

III. Data Cleaning

Here, the social media review data and taxi trajectory data used in this study are described.

A. Taxi data pre-processing

Eight days of taxi trajectory data in Shenzhen from 13–20 October 2013 were collected. The dataset includes three million individual journeys from 15,000 taxis. Each journey records data at 30 second intervals, including taxi location (longitude, latitude), speed, direction-angle, and status (0: taxi has no passenger; 1: taxi has passenger).

To calibrate the Huff model, it is necessary to have choice-based samples, such that groups of individuals have chosen to visit a particular destination [9], [12]. Choice-based samples are used to make inferences about the full population, so samples must be representative and unbiased. However, since taxi fares are generally higher than other transport modes, taxi data has a natural bias on customers' income and travel distance. We are aware of this limitation, but believe that the large quantity of taxi data we have available is representative of the major shopping trends in the city.

Shenzhen taxi data is initially segmented into a grid of square cells of side 400 meters, with range boundary 113.80°–114.63° longitude and 22.46°–22.80° latitude. For non-empty cells, the mean number of taxi pick-up points is 67, making 400 meters a suitable minimum resolution. The same steps used in [5] to extract choice-based samples are then followed: (i) taxi drop-off points located near target shopping centres are selected. As anchor stores play an important role in shopping centre attractiveness [13], previous research defined a buffer radius for shopping centre to embed GPS error and human behaviour randomness [5]. Following this method, it is considered that customers aim to visit a target shopping centre if taxi drop-off points are located within a 500 metres buffer radius around the shopping centre (previously shown to be an average walking trip in China [14]); (ii) for each drop-off point, the corresponding taxi pick-up point is collected in order to extract the Origin-Destination (O-D) pairs; (iii) as most of the shopping centres in Shenzhen are open from 10am to 10pm, only taxi O-D pairs in which GPS time is from 10am to 10pm are extracted; (iv) previous research defined the primary trading area of a shopping centre as the region where 75% to 80% of its customers live [15]. In this study, the closet 80% of taxi pick-up points for each shopping centre are selected to represent this area.

To verify model calibration, data is split into two subsets: training data (used for calibration), and testing data (used for verifying the prediction accuracy of the calibrated model). Every tenth O-D pair is selected for the test set, the other 90% of data are used for calibration. Since it has previously been shown that model calibration produces relatively high error rates during working hours [6], in this study shopping behaviours during weekday working hours (10am–5pm) are not considered. As a result, training and testing data are further segmented into two subsets: weekend; and weekday evenings (5pm–10pm).

B. Social media data pre-processing

There are a total of 94 shopping malls located in the five target shopping centres studied in this paper. Reviews of these shopping malls were collected from Dianping.com, a social media platform that enables customers to share their comments and give a rating (from one to five) for each shopping mall. We collected 8,070 reviews in total. For each mall, the number of reviews given each rating score was calculated. A summary of the review data is presented in Table I.

To obtain the review ranking of the 5 shopping centres, Bayesian Average is used to pre-process the data. Bayesian Average is a statistical method based on Bayes' Rules [16]. It

TABLE I: Summary of Dianping shopping mall data.

Shopping Centre	Region	Malls	Reviews	S_{review}
Dongmen	Luohu	33	2495	0.196
Huaqiangbei	Futian	31	1941	−0.367
Futian	Futian	10	1727	−0.020
Nanshan	Nanshan	16	1401	−0.165
Baoan	Baoan	4	506	0.531

has been widely used for calculating the ratings of customer reviews, such as movie rankings.[1]

Bayesian Average is calculated as follows:

$$q = \frac{\sum_{i=1}^{n} r_i + B \times m}{B + n} \tag{5}$$

where q is the rating value of each shopping centre, r is the rating of each mall, n is the number of votes for each shopping mall, B is the mean number of votes across the whole set for each shopping mall, and m is overall average rating in the whole set.

After each shopping centre's rating has been calculated, z-score is used to normalize the social media data. After normalization, the data is ranged from [-1,1]. The form of z-score is calculated as follows:

$$z = (x - \mu)/\sigma \tag{6}$$

where x is the input data, μ is the average value of x, and σ is standard deviation of x. Z-score values are used as an estimation of shopping centre attractiveness, S, in the Huff model. We label this estimation as $S = S_{review}$.

IV. Methodology

There are two variables in the Huff model: shopping centre attractiveness (S) and travel cost (C). Here, two factors are considered as estimations of attractiveness, S: (i) number of journeys ($S_{journey}$) to a shopping centre—equivalent to footfall—calculated directly from the taxi data drop-off points; and (ii) average z-score review rating of shopping centres, calculated from the social media data (S_{review}). To estimate travel cost, C, O-D route distance returned from Baidu.com's Application Program Interface (API) is used. To calibrate the Huff model, K1 (2) and K2 (3) are applied, using R's *spgwr* package to perform local and global GWR calibration.

Following calibration—where the best fit parameters over the training data are selected—the predictive accuracy is verified by comparing model prediction against the test data using Kullback-Leibler (KL)-divergence. KL-divergence is a mathematical statistic used to measure the divergence between two probability distributions [17]. The function of KL-divergence is given as:

$$D = \sum_{i} (P_i) log(P_i/Q_i) \tag{7}$$

where P is the observed patronage probability, Q is the calculated probability, and D is the difference between P and Q. KL-divergence represents high forecasting accuracy when the value is low; and represents poor forecasting accuracy when the value is high.

V. Results

A. Quantitative evaluation of model

Results of model calibration are presented in Table II. Residual Standard Error (lower values indicate better fit), R^2 (higher values indicate better fit), and Sum of Squares (lower values indicate better fit) are used to evaluate the parameter fitting result.

When $S = S_{review}$, K1 gives the best global result with α=0.029 and β=-0.245 on weekends, and α=0.041, β=-0.242 on weekday evenings. While there is little variation in β, it can be seen that α is higher on weekday evenings. This suggests that customers care more about shopping centre reviews on weekday evenings than on weekends. One interpretation of this could be that since customers have less time on weekday evenings, they prefer to choose a shopping centre with higher ratings since they do not want to take the risk of selecting a "bad" mall. Conversely, on weekends, customers have more time to explore, and so are less likely to strictly follow the review ratings of others.

Prediction results for social media reviews are also shown in Table II, measured using KL-divergence, where lower values indicate better prediction. On weekends, it can be seen that there is little change in predictive power when moving from a global to a local model. This is interesting, since under all other conditions a model fitted geographically (i.e., a local model) performs better. From this, it can be inferred that, at weekends, customer sensitivity to social media reviews does not vary geographically. Perhaps this is because customers have more time to consider reviews and are prepared to travel to malls rated highly, regardless of where in the city they live.

When $S = S_{review}$, the prediction at weekends is more accurate than weekday evenings for both a global and local model. This suggests that there is more predictability in customer responses to reviews at weekends than weekday evenings. This is also likely to be a consequence of shoppers having longer to read and digest multiple reviews to make more informed, and therefore predictable, choices.

Overall, under each condition, the Huff model calibrated locally performs better than when calibrated globally. For local calibration, $S = S_{journey}$ offers significantly better predictive power. However, for global calibration, $S = S_{review}$ offers better predictive power than $S = S_{journey}$ at weekends. This intriguing result is investigated further in the following section by looking at the spatial probability distribution of behaviours.

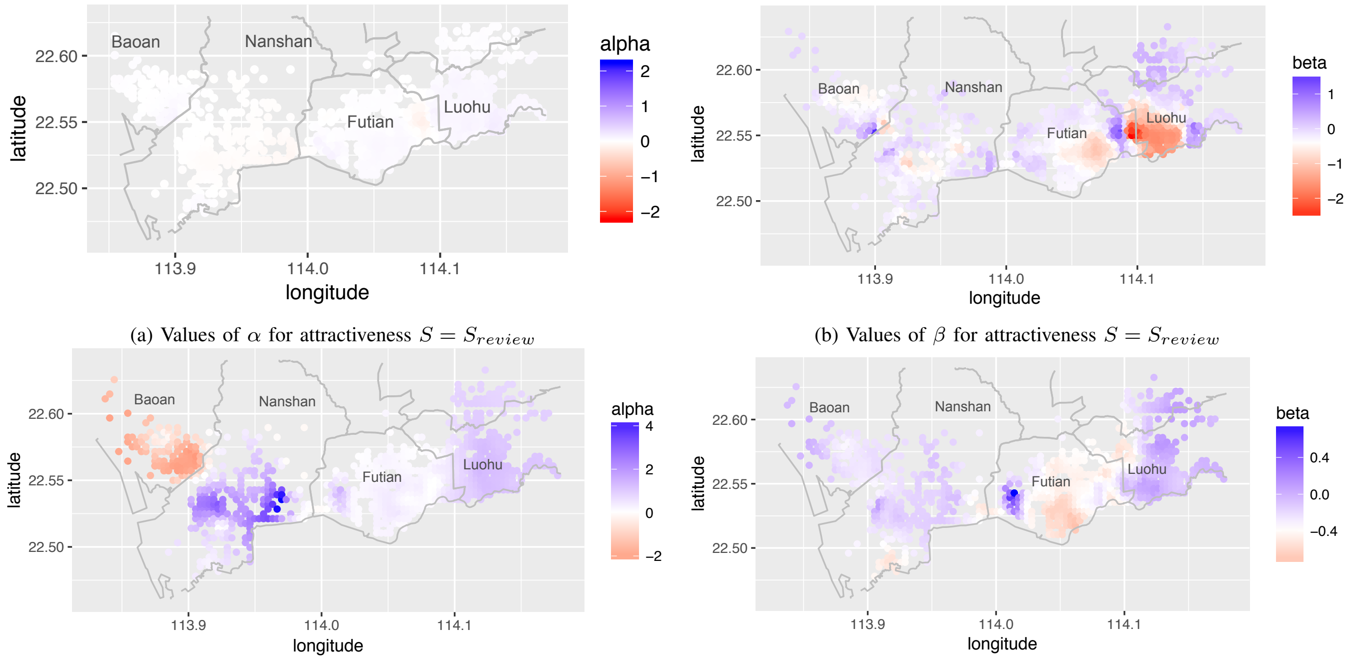
B. Spatial distribution

Fig. 1 presents the geographical parameter distributions of α and β for the local Huff model calibration. The districts mapped are: Baoan, Nanshan, Futian, and Luohu.

[1] For example, Bayesian Averaging is used by Internet Movie Data Base (IMDB), the world's most popular and authoritative source for movie, TV and celebrity information content, containing more than 2 million records.

TABLE II: Model calibration and testing for weekends. Global calibration on training data (highest R^2, lowest sum of squares); and KL-divergence of calibrated models on test data (smaller values indicate greater model prediction accuracy).

Attractiveness	Time	Estimator	Global Huff Calibration					KL-Divergence	
			α	β	Residual S.E.	R^2	Sum of squares	Global	Local
Reviews	weekend	K1	0.029	-0.245	0.024	0.754	0.214	0.21	0.20
	weekday	K1	0.041	-0.242	0.026	0.833	0.167	0.59	0.36
Journeys	weekend	K2	0.198	-0.261	0.028	0.806	0.193	0.42	0.09
	weekday	K2	0.126	-0.281	0.028	0.814	0.185	0.40	0.05



(a) Values of α for attractiveness $S = S_{review}$

(b) Values of β for attractiveness $S = S_{review}$

(c) Values of α for attractiveness $S = S_{journey}$

(d) Values of β for attractiveness $S = S_{journey}$

Fig. 1: GWR calibration on weekend. The four regions are: Baoan, Nanshan, Futian, and Luohu.

1) α distribution: When $S = S_{journey}$ (Fig. 1c), α is positive highest (blue) in Nanshan district and negative highest (red) in Baoan district. This suggests that people who live in Nanshan are more likely to shop at attractive stores (those with the largest footfall), whereas people who live in Baoan prefer the opposite. In Futian (where Huaqiangbei and Futian Shopping Centre are located) and Luohu (where Dongmen Shopping Centre is located), α values are close to 0. In these regions, people pay less attention to the attractiveness of shopping centres when deciding where to shop.

For social media reviews, there is a very different spatial distribution (see Fig. 1a). Throughout all regions, α displays little variation. This suggests that, regardless of where customers live, they pay similar attention to shopping centre ratings on social media. It can therefore be inferred that social media reviews have similar impact on customers who live in different regions. This helps to explain why global and local calibration of the model performs similarly when $S = S_{review}$.

2) β distribution: In Baoan and Nanshan regions, β values are largely positive for both $S = S_{review}$ and $S = S_{journey}$. This suggests that people who live in Baoan prefer to travel far to visit a popular shopping centre. This counter-intuitive result is easier to understand by considering the locations of shopping centres. In total, data from 94 shopping malls were collected. For the top 20 malls with the highest review ratings, only one mall is located in Baoan, and only three malls are located in Nanshan. Starved of choice, therefore, customers in Baoan and Nanshan have need to travel much farther to shop than customers in Futian and Luohu.

In southwest of Luohu and south east of Futian, the values of β are highly negative when $S = S_{review}$, but positive when $S = S_{journey}$. Once again, this can be interpreted via the ranking of malls in the social media reviews: of the top-five ranked malls, four are located in Luohu near the boundary of Futian—exactly in the red area shown in Fig. 1b. Since a large collection of the highest rated malls are located in this region, so customers have no reason to travel far, therefore resulting in negative β scores.

In contrast, when $S = S_{journey}$, β values are positive throughout Luohu. To understand this, customer volumes at each shopping centre are observed directly from the taxi data—it can be seen that there are two shopping centres in Futian with the highest footfall. Therefore, this increased weighting of these popular stores affects the localised β values in Luohu.

VI. Application

The results of Huff model prediction demonstrate that social media data has high performance on global calibration. This suggests that when taxi trajectory data is not available or not large enough, it is very likely to be a good choice to use social media data to calibrate the Huff model instead of taxi data. In particular, shopping centre reviews on social media could be used to estimate attractiveness, while large open source customers' check-in data with location can likely be used to analyse and estimate shopping journeys. By fusing big data sources to tease out the causal factors of shopping behaviours, this approach has the potential for positive impact on urban planning, transportation, and retail applications.

VII. Conclusion

Huff model prediction of shopping behaviours is presented for the city of Shenzhen. The Huff model is calibrated using taxi trajectory data and social media review data of shopping malls. In total, the taxi data contains more than three million individual taxi journeys for 15,000 taxis collected across eight days, with GPS data recording location every 30 seconds. The social media data contains more than 8,000 reviews across 94 shopping malls in Shenzhen. The fusion of these big data sets is used to calibrate the Huff model using Geographical Weighted Regression, in order to determine spatial relationships in shopping behaviours. This is an example of automated data mining and analysis applied to a problem area that traditionally required labour intensive collection of data via surveys and interviews.

Prediction results show that errors in forecasting can be as low as 5% for weekday evenings, and 10% for weekends. This demonstrates the power of the process. Further, it is demonstrated that while there is great spatial variation in the behaviour of shoppers in Shenzhen, the response of customers to mall reviews posted on social media is to a large extent spatially invariant at weekends. This result demonstrates the great geographical reach of social media reviews on customer behaviours.

Despite these successes, some limitations to this study are acknowledged: in particular regarding the use of taxi data as the sole travel mode for a city. To address this, future work will compare results against other travel vectors. Other extensions to be considered in the future include: introducing time-series forecasting methods for shopping behaviour predictions (e.g., [10]); and building agent-based models of shopping behaviours, calibrated using real-world travel data (e.g., [18]).

Acknowledgement

This research was supported by the National Science Foundation of China (No. 41671387, 91546106, 71471092), Shenzhen Scientific Research and Development Funding Program (No. CXZZS20150504141623042), Zhejiang Natural Science Foundation (Grant No. LR17G010001), Ningbo Science and Technology Bureau (Grant No. 2014A35006), UK Engineering and Physical Sciences Research Council (Grant No. EP/L015463/1), and Thomson Reuters.

References

[1] M. Bohnet and J.-M. Gutsche, "Estimating land use impacts on transportation—findings from the Hanover region," in *Proceedings of the European Transport Conference*. Leiden, The Netherlands: Association for European Transport, Oct 2007.

[2] D. L. Huff, "Defining and estimating a trading area," *The Journal of Marketing*, vol. 28, no. 5, pp. 34–38, 1964.

[3] A. Gracia and T. de Magistris, "The demand for organic foods in the south of italy: A discrete choice model," *Food Policy*, vol. 33, no. 5, pp. 386–396, 2008.

[4] C. Chu, "A paired combinatorial logit model for travel demand analysis," in *Transport Policy, Management & Technology Towards 2001: Selected Proceedings of the Fifth World Conference on Transport Research*, vol. 4. Yokohama, Japan: World Conference on Transport Policy, 1989, pp. 295—309.

[5] Y. Yue, H.-d. Wang, B. Hu, Q.-q. Li, Y.-g. Li, and A. G. Yeh, "Exploratory calibration of a spatial interaction model using taxi gps trajectories," *Computers, Environment and Urban Systems*, vol. 36, no. 2, pp. 140–153, 2012.

[6] S. Gong, J. Cartlidge, Y. Yue, G. Qiu, Q. Li, and J. Xin, "Geographical Huff model calibration using taxi trajectory data," in *Proceedings of 10th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, Redondo Beach, CA, USA, November 7–10, 2017. https://doi.org/10.1145/3151547.3151553

[7] Y. Wang, W. Jiang, S. Liu, X. Ye, and T. Wang, "Evaluating trade areas using social media data with a calibrated huff model," *ISPRS International Journal of Geo-Information*, vol. 5, no. 7, p. 112, 2016.

[8] D. L. Huff, "A probabilistic analysis of shopping center trade areas," *Land economics*, vol. 39, no. 1, pp. 81–90, 1963.

[9] M. E. O'Kelly, "Trade-area models and choice-based samples: methods," *Environment and Planning A*, vol. 31, no. 4, pp. 613–627, 1999.

[10] J. Cartlidge, S. Gong, R. Bai, Y. Yue, Q. Li, and G. Qiu, "Spatio-temporal prediction of shopping behaviours using taxi trajectory data," in *Proceedings of IEEE 3rd International Conference on Big Data Analysis (ICBDA'18)*, Shanghai, China, March 9–12, 2018.

[11] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically weighted regression: the analysis of spatially varying relationships*. University of Newcastle, UK: John Wiley & Sons, 2003.

[12] F. A. Stewart and M. E. O'Kelly, "Spatial interaction models: formulations and applications," 1989.

[13] A. Finn and J. J. Louviere, "Shopping center image, consideration, and choice: anchor store contribution," *Journal of business research*, vol. 35, no. 3, pp. 241–251, 1996.

[14] W. T. P. Bureau, "Wuhan traffic impact analysis guideline," Technical report, Wuhan Transportation Planning Bureau, Tech. Rep., 2009.

[15] W. Applebaum, "Methods for determining store trade areas, market penetration, and potential sales," *Journal of Marketing Research*, vol. 3, no. 2, pp. 127–141, 1966.

[16] X. Yang and Z. Zhang, "Combining prestige and relevance ranking for personalized recommendation," in *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, 2013, pp. 1877–1880.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[18] M. Birkin and A. Heppenstall, "Extending spatial interaction models with agents for understanding relationships in a dynamic retail market," *Urban studies research*, Article ID 403969, 2011. https://doi.org/10.1155/2011/403969