

# Spatio-temporal prediction of shopping behaviours using taxi trajectory data

John Cartlidge\*, Shuhui Gong<sup>†‡</sup>, Ruibin Bai<sup>†</sup>, Yang Yue<sup>‡</sup>, Qingquan Li<sup>‡</sup> and Guoping Qiu<sup>§</sup>

\*Department of Computer Science, University of Bristol, Bristol UK, BS8-1UB

Email: [john.cartlidge@bristol.ac.uk](mailto:john.cartlidge@bristol.ac.uk)

<sup>†</sup>International Doctoral Innovation Centre & Department of Computer Science  
University of Nottingham Ningbo China, Ningbo China, 315100

Email: {shuhui.gong, ruibin.bai}@nottingham.edu.cn

<sup>‡</sup>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

<sup>§</sup>School of Computer Science, University of Nottingham, Nottingham UK, NG8-1BB

Email: [guoping.qiu@nottingham.ac.uk](mailto:guoping.qiu@nottingham.ac.uk)

**Abstract**—Taxi trajectory data (GPS data collected for 15,000 taxis at intervals of 30 seconds across three million journeys over eight days) is used to generate a spatio-temporal prediction of shopping behaviours in the emerging metropolitan city of Shenzhen, China. Two approaches are compared: time-series forecasting using ARIMA; and a gravity model approach, using the Huff model calibrated with Geographical Weighted Regression. Results demonstrate that ARIMA performs with significantly higher accuracy than the more traditional Huff model method. Further, it is demonstrate that while the accuracy of the Huff model is constrained by model assumptions, applying time-series methods to the underlying data directly (i.e., the ARIMA method) has no such constraints, and is limited only by the amount of data available. This suggests that, as richer data sets become available, spatio-temporal modelling of this kind will become more accurate.

**Index Terms**—Taxi trajectory data; time-series; ARIMA; Huff model; Geographically Weighted Regression; shopping behaviour

## 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. In a previous study, the authors calibrated the Huff model using Geographical Weighted Regression (GWR) over taxi trajectory data for the city of Shenzhen, China [5]. By using GWR, the Huff model was calibrated independently at each geographic location, rather than globally. Results showed that GWR calibration of Huff performed with much higher accuracy than global calibration, and evidenced the spatial variation in shopping behaviours

across the city of Shenzhen. One factor contributing to this variation was shown to be wealth of customers [5].

Here, previous work is extended by introducing time-series analysis of the taxi data to generate a spatio-temporal model of shopping behaviours directly from the underlying data. Time series analysis can be used to discover time regularity of data. First developed in 1976, Auto-Regressive Integrated Moving Average (ARIMA) is one of the most commonly used methods for time-series analysis [6]. It consists of three parts, auto regression (AR), moving averages (MA), and differencing in order to strip off the integration (I). ARIMA has been used to successfully forecast tourist demand [7], wholesale market demand [8], and local market sales [9]. In this paper, ARIMA is used to develop a spatio-temporal model of shopping behaviours and compared with GWR-calibrated Huff model predictions. Results show that using ARIMA to forecast customers' shopping behaviours provides significantly higher accuracy than using a spatially-calibrated Huff model.

## II. RELATED WORK

The Huff model [10] 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 Huff model is classically expressed as:

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

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  $\alpha$  and  $\beta$  (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.

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

$$y_i = \gamma_{i0} + \sum_{k=1}^m \gamma_{ik} x_{ik} + \epsilon_i \quad (2)$$

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 there are two variables  $\alpha$  and  $\beta$  in the Huff model,  $m = 2$ );  $\gamma_{i0}$  is the intercept parameter at location  $i$ ;  $\gamma_{ik}$  (corresponding to  $\alpha$  and  $\beta$  in (1)) is the local regression coefficient for the  $k^{th}$  independent variable at location  $i$ ; and  $\epsilon_i$  is the random error at location  $i$ .

Previously, the Huff model was calibrated using GWR over taxi trajectory data to fit the model spatially [5], [12]. Here, to discover time regularity of the taxi data, a time series method—ARIMA [6]—is used to forecast shopping probability.

ARIMA contains three parameters and is represented formally as a ternary function  $ARIMA(p, d, q)$ ; where  $p$  is the autoregressive order,  $d$  is the order of the difference, and  $q$  is the order of the lagged forecasting errors. Here,  $R$ 's *forecast* package is used to determine  $p$ ,  $d$ , and  $q$  automatically. Once  $d$  is determined, the linear model becomes:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \dots - \theta_q e_{t-q} \quad (3)$$

where  $\hat{y}_t$  is the forecast result at time  $t$ ,  $\mu$  is the constant,  $\phi$  and  $\theta$  are parameters defined according to the historical time series,  $y_{t-1}, \dots, y_{t-p}$  are historical time series data, and  $e_{t-1}, \dots, e_{t-q}$  are lagged forecasting errors [13]. ARIMA can be used for short-run estimation over time intervals ranging from hours to years [14].

Here, ARIMA and the Huff model are used to forecast spatio-temporal shopping probability distributions for the city of Shenzhen. Results are compared to discover which method generates the greatest predictive power, and why.

### III. DATA CLEANING

Eight days of taxi trajectory data were collected for the city of Shenzhen between 13–20 October 2013. The dataset includes three million individual journeys across 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).

#### A. Extract choice-based samples

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 [15], [16]. 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

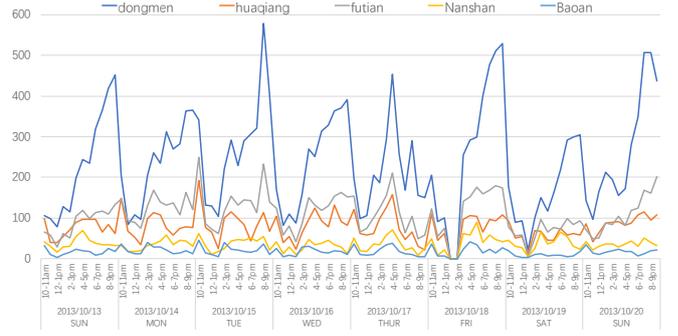


Fig. 1: Hourly customer volume (Dongmen: blue; Huaqiangbei: orange; Futian: grey; Nanshan: yellow; Baoan: cyan).

large quantity of taxi data available is representative of the major shopping trends in the city.

Initially, all Shenzhen taxi data is segmented into a grid of square cells of side 400 meters, with range boundary  $113.80^\circ$ – $114.63^\circ$  longitude and  $22.46^\circ$ – $22.80^\circ$  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 identified by [17] to extract choice-based samples is then followed: (i) taxi drop-off points located within 500 metres of target shopping centres are selected; (ii) for each drop-off point, the corresponding taxi pick-up point is extracted in order to generate Origin-Destination (O-D) pairs; (iii) as most of the shopping centres in Shenzhen are open from 10am to 10pm, taxi O-D pairs with GPS time outside of these hours are discarded; (iv) previous research defined the primary trading area of a shopping centre as the region where 75% to 80% of its customers live [18]. In this study, the closet 80% of taxi pick-up points for each shopping centre are used to represent this area; (v) finally, to enable representative time-series forecasting, cells that do not have shopping records in all eight days are rejected.

Hourly time-series volume of choice-based samples for each shopping centre are plotted in Fig. 1. It can be seen that Dongmen has much greater volume than the other centres, and correspondingly higher volatility (difference between peak and off-peak) throughout the day; in comparison, customer volumes in Nanshan and Baoan vary much less over time.

#### B. Training data and Testing data

Since two forecasting methods on customers' shopping behaviour (ARIMA and Huff model) are tested, the data are treated in two ways: (i) For ARIMA forecasting, taxi O-D pairs from 13–19 October are used to train the model, including origin location (longitude, latitude), date (from 13 to 19, inclusive), and destination probabilities to all five shopping centres. After training the model, O-D pairs for 20 October are used as test data to evaluate the prediction accuracy of ARIMA. (ii) For Huff model forecasting, only taxi O-D pairs for 13 October and 19 October (both days are weekend dates) are considered

as the training set to calibrate the Huff model.<sup>1</sup> The dataset includes origin location (longitude, latitude), each shopping centre's history of customer volume, route distance to each shopping centre, and shopping probability to each shopping centre. In order to compare the Huff model's performance against ARIMA, we also use taxi data of 20 October as test data to evaluate the predication accuracy.

#### IV. METHODOLOGY

As described in Section III, we only utilise data from geographical cells (each a square of side 400 metres) that have shopping records in all 8 days. As such, we extract 369 cells (each cell has at least one shopping journey origin every day), and then apply ARIMA (using R's *forecast* package) to forecast each cell's shopping probability distribution for 20 October. Previously, the authors used GWR to calibrate the Huff model for the taxi data [5]. These best-fit parameters are reused here to predict spatio-temporal shopping behaviours on 20 October.

To evaluate and compare the forecast performance of Huff and ARIMA, four measures are used to determine errors of forecast customer volume against real customer volume (taken directly from taxi journeys to each centre) for 20 October. These error measures are described in turn, below. For each measure, larger absolute values indicate greater error. Likewise, lower absolute values indicate lesser error, with 0 indicating a perfect forecast against the real ground truth data.

Root Mean Square Error (RMSE) is a commonly used method to calculate the differences between forecast and observed data. The classic format for RMSE is:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y - f(x))^2} \quad (4)$$

Mean Error (ME) is used to calculate mean difference between forecast and observed data. The equation is:

$$ME = \frac{1}{n} \sum_{i=1}^n y - f(x) \quad (5)$$

Mean Percentage Error (MPE) is used to calculate the percentage difference between forecast and observed data. The equation is:

$$MPE = \frac{1}{n} \sum_{i=1}^n \frac{y - f(x)}{y} \quad (6)$$

Mean Average Percentage Error (MAPE) usually expressed as a percentage, is defined as:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y - f(x)}{y} \right| \quad (7)$$

For all equations (4–7), above,  $y$  represents the real value (i.e., the actual ground truth data value),  $f(x)$  represents the forecast value, and  $n$  is the number of data pairs ( $y$  and  $f(x)$ ).

<sup>1</sup>Previously, the Huff model has been shown to have greatest predictive power when considering weekend and weekday data separately [5]. Therefore, here only weekend data is used for Huff model calibration.

## V. RESULTS AND DISCUSSION

### A. ARIMA forecasting

Table I presents a forecast comparison between ARIMA and Huff. For ARIMA, ME, MPE, and MAPE return values around 10% or less, suggesting high forecast accuracy. Errors for RMSE are larger, suggesting that there are some outliers in the real data (likely due to irregular and unanticipated shopping trips across fairly large distances). ME forecasts are consistently extremely low, suggesting the forecasts are unbiased about the mean. Comparing the forecasts between the five shopping centres, it can be seen that values of RMSE, MPE and MAPE are lowest in Baoan and highest in Dongmen. More specifically, the list of shopping centres ordered by forecast errors from highest to lowest is: Dongmen, Futian, Huaqiangbei, Nanshan, and Baoan. Referring to Fig. 1, it can be seen that this is also the ordering of shopping centres in terms of retail volume and volatility (i.e., the difference between peak-time crests and off-peak lows). Therefore, this is an indication that customer volumes in Dongmen have less regularity than in other shopping centres. Since ARIMA forecasts are based on historical shopping records, it is likely that ARIMA forecasting performs more accurately when previous records have regular, periodic features.

### B. Huff model prediction

Table I presents results of forecast prediction for the Huff model. Throughout, forecast prediction errors are higher for Huff than for ARIMA, suggesting that ARIMA is a better predictor than Huff. Considering Huff in isolation, it can be seen that Huff has better performance on Dongmen, Huaqiangbei, and Futian. Errors in prediction forecasts for Nanshan and Baoan are very high. This contrasts with the behaviour of ARIMA.

In Fig. 1, customer volumes for Nanshan and Baoan are much lower than for the other shopping centres. Since the Huff model is calibrated geographically (using GWR) using historical taxi journeys, it is likely that a lack of data is negatively impacting Huff's prediction accuracy. Therefore, more taxi data could help improve the prediction accuracy of the model.

### C. Spatial comparison between ARIMA and Huff

Results in Table I demonstrate that ARIMA outperforms Huff on the taxi data available. Therefore, overall, it can be inferred that historical shopping records in Shenzhen have strong time regularity. As a result, it appears that time-series methods are suitable for analysing and forecasting shopping behaviours. In contrast, the Huff model is steady-state and does not consider time. It has previously been shown that Huff performs differently at weekends compared with weekdays [5]. Once again, this suggests that time-series forecast approaches to shopping behaviours are more suited than steady-state models.

In Table I it can be seen that ARIMA forecasts have greatest error on Dongmen shopping centre, while the equivalent Huff model forecasts have the lowest error. To better understand

TABLE I: Forecast prediction results. ARIMA outperforms Huff on all metrics (absolute values closest to zero).

Shopping centre	ARIMA				Huff			
	RMSE	ME	MPE (%)	MAPE (%)	RMSE	ME	MPE (%)	MAPE (%)
Dongmen	0.23	0.01	-1.10	11.14	0.34	0.01	-4.13	19.70
Huaqiangbei	0.18	-0.01	-2.24	9.90	0.29	0.11	4.71	14.49
Futian	0.21	-0.01	-2.23	10.66	0.29	-0.19	-18.78	22.42
Nanshan	0.14	-0.01	-1.11	5.00	0.64	-0.56	-57.22	59.99
Baoan	0.07	-0.01	-0.41	1.58	0.38	-0.32	-32.73	34.31

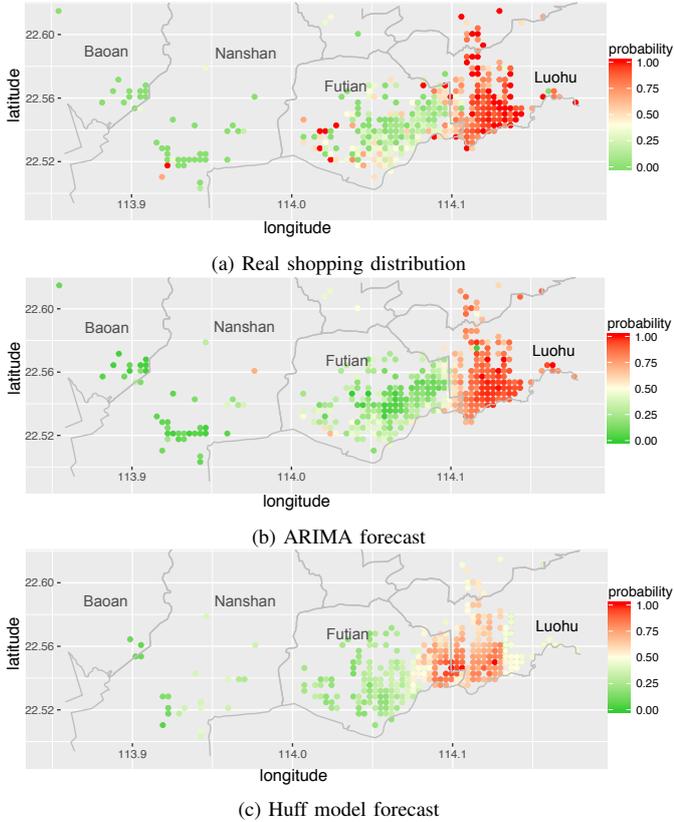


Fig. 2: Forecast for Dongmen shopping centre in Luohu. Districts mapped are: Baoan, Nanshan, Futian, and Luohu.

these results, geographical shopping forecasts to Dongmen using both methods are compared. Results are presented in Fig. 2. The districts mapped (from left to right, or west to east) are: Baoan, Nanshan, Futian, and Luohu. Dongmen shopping centre is located in the south of Luohu, near the coast. The observed shopping probabilities for Dongmen (the ground truth) are presented in Fig. 2a. It can be seen that, unsurprisingly, the majority of shopping trips originate from near Dongmen’s location in Luohu, with further high-probability origins spreading north and also east along the coast. There are also small pockets of high-probability origins further to the west in Futian, and occasional journeys originating very far to the west in Nanshan. No shopping journeys to Dongmen begin in the

western region of Baoan. This distribution of shopping origins to Dongmen helps to explain the relatively high RMSE error values for ARIMA forecasts, with squared errors of anomalous long-distance journeys affecting the prediction accuracy more than the other error metrics.

Fig. 2b shows the prediction of shopping behaviours for Dongmen shopping centre using ARIMA. Visually, the prediction matches the real data (Fig. 2a) well, particularly in the high-probability regions of Luohu (where Dongmen is located) and Futian, immediately to the west. In particular, the forecast matches the spatial structure of the real data in these regions, indicating that it accounts for areas of high residential density such as along the coast to the east of Dongmen, and to the north. The most noticeable differences between the ARIMA forecast and real data are the missing long-distance journeys beginning in the west of Futian and Nanshan. This is likely due to the lack of data in these areas. With more taxi data, it is likely that spatial shopping patterns throughout all of Shenzhen could be predicted with similar accuracy as the regions with closest proximity to Dongmen shopping centre.

Fig. 2c shows the shopping prediction for Dongmen shopping centre generated by the calibrated Huff model. While the Huff prediction has spatial similarities with the real data at the macro scale (specifically, high probability in the immediate vicinity of Dongmen and decreasing with distance), it visibly matches the real data less well than the ARIMA prediction. Most noticeably, the Huff model assumptions of exponential distance decay means that the real spatial structure of Shenzhen is not accounted for. Therefore, the high probability regions stretching away to the north of Dongmen and along the coast to the east are not observed; neither is there evidence of any long distance journeys. This is not surprising given the constraints of the Huff model assumptions. By calibrating the Huff model using GWR, some of the constraints of the Huff model are overcome as the model is individually calibrated at each spatial location. However, despite this, there is still a tendency (clearly visible in the figure) for spurious spatial regularity.

Overall, the results presented in Fig. 2 give a clear visual indication of the relative merits of using time-series predictions such as ARIMA for spatio-temporal shopping behaviour predictions over more constrained spatial models such as the Huff model. This is also evidenced quantitatively in the greater predictive accuracy of ARIMA over Huff, as shown in Table I.

Therefore, this appears to be a promising area of investigation for further work. In future, it will be interesting to extend the spatio-temporal predictive modelling of shopping behaviours by applying different time-series approaches and gathering more extensive taxi trajectory data. This will enable testing of the hypothesis that the majority of the forecast errors predicted by ARIMA is a result of limited data. In contrast, the restrictive assumptions of the Huff model mean that it is impossible to achieve similar forecast accuracy irrespective of the amount of data available.

## VI. SUMMARY OF FINDINGS

The results in this paper demonstrate that spatio-temporal modelling of shopping behaviours using time-series approaches such as ARIMA have much higher prediction accuracy than steady-state models such as the Huff model; likely due to the structural constraints of the mathematical assumptions imposed on the Huff model. Free of these constraints, in comparison, it is expected that the performance accuracy of ARIMA will improve given more data to fit the model.

The potential impact of these findings are two-fold. First, a move to spatio-temporal modelling using time-series methods should be considered where large datasets are available. Second, such unconstrained time-series models will enable better predictive accuracy than the simplifying assumptions of traditional mathematical models (such as Huff) will allow. Improved accuracy of models to predict shopping behaviours have the potential to positively impact urban, retail, and transportation planning; particularly in the burgeoning era of smart cities.

## VII. CONCLUSION

A spatio-temporal model for predicting shopping behaviours using taxi trajectory data for the emerging metropolitan city of Shenzhen, China, has been presented. A comparison between a steady state gravity model (the Huff model) and a time-series method (ARIMA) was undertaken on the data. Results showed that while both models have some expressive power, the Huff model is limited in prediction accuracy due to assumption constraints of the mathematical model. Contrastingly, having freedom from such constraints, ARIMA predictions are much more accurate; and in regions of high density data, accuracy grows. It is therefore suggested that increasing the input data to the model can increase the accuracy of the ARIMA prediction.

Future work will test to see whether a larger taxi journey dataset can improve prediction accuracy. Other time-series methods combined with taxi trajectory data to form spatial-temporal models will also be considered. Once completed, the aim is to introduce an agent-based model to predict individual behaviours at the micro-scale, and the relationship with the macro-scale behaviour of the city (e.g., [19]). This work should hopefully enable what-if scenario testing for locating new retail and residential centres, as well as corresponding traffic infrastructure.

## 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] 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>
- [6] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
- [7] C. J. Lin, H. F. Chen, and S. L. Tian, "Forecasting tourism demand using time series, artificial neural networks and multivariate adaptive regression splines:evidence from taiwan," *International Journal of Business Administration*, vol. 2, no. 2, 2011.
- [8] M. Shukla and S. Jharkharia, "Applicability of arima models in wholesale vegetable market: An investigation," *International Journal of Information Systems and Supply Chain Management*, vol. 6, no. 3, pp. 105–119, 2013.
- [9] D. R. Munro, "Consumer behavior and firm volatility," 2016.
- [10] D. L. Huff, "A probabilistic analysis of shopping center trade areas," *Land economics*, vol. 39, no. 1, pp. 81–90, 1963.
- [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] S. Gong, J. Cartlidge, R. Bai, Y. Yue, Q. Li, and G. Qiu, "Automated prediction of shopping behaviours using taxi trajectory data and social media reviews," in *Proceedings of IEEE 3rd International Conference on Big Data Analysis (ICBDA'18)*, Shanghai, China, March 9–12, 2018.
- [13] A. Parkantz, "Forecasting with univariate box-jenkins models," 1983.
- [14] B. Petrevska, "Predicting tourism demand by arima models," *Economic Research-Ekonomska Istraživanja*, vol. 30, no. 1, pp. 939–950, 2017.
- [15] M. E. O'Kelly, "Trade-area models and choice-based samples: methods," *Environment and Planning A*, vol. 31, no. 4, pp. 613–627, 1999.
- [16] F. A. Stewart and M. E. O'Kelly, "Spatial interaction models: formulations and applications," 1989.
- [17] 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.
- [18] 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.
- [19] 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. [10.1155/2011/403969](https://doi.org/10.1155/2011/403969)