# Geographical Huff Model Calibration using Taxi Trajectory Data

Shuhui Gong
International Doctoral Innovation Centre, University of Nottingham
Ningbo China, Ningbo, China 315100
Shuhui.Gong@nottingham.edu.cn

John Cartlidge
Department of Computer Science, University of Bristol
Bristol, UK BS8 1UB
John.Cartlidge@bristol.ac.uk

Yang Yue
Shenzhen Key Laboratory of Spatial Smart Sensing and Services,
Shenzhen University, China 518060
YueYang@szu.edu.cn

Guoping Qiu
Department of Computer Science, University of Nottingham
Nottingham, UK NG8 1BB
Guoping.Qiu@nottingham.ac.uk

Qingquan Li
Shenzhen Key Laboratory of Spatial Smart Sensing and Services,
Shenzhen University, China 518060
liqq@szu.edu.cn

Jingyu Xin
Department of Electrical and System Engineering, Washing University in St. Louis, St. Louis MO, 63130
xinjingyu@wustl.edu

## ABSTRACT

The widely used Huff model is designed to estimate the spatial probability distribution of shopping centre patronage based on a shopping centre's attractiveness and the cost of a customer's travel. Here, we calibrate the Huff model for the city of Shenzhen, China, using GPS taxi trajectory data for one million taxi journeys. Using Geographical Weighted Regression to fit the model, we show that there is significant geographical variation in best estimates of the Huff parameters of attractiveness and cost. To explain this variation, we use open-source house price sales' data as a proxy for customers' wealth in each region. Regression results demonstrate a significant linear relationship between localised house prices and the Huff model parameter of attractiveness, suggesting that wealthy customers are more sensitive to shopping centre attractiveness than customers with less wealth. We present this as a novel discovery.

## CCS CONCEPTS

• **Information systems application** → **Geographic information systems**; • **Information systems applications** → *Data analytics*; • **Operations research** → Transportation;

## KEYWORDS

Shopping behavior; Taxi data; House price data; GWR; OLS

**ACM Reference format:**
Shuhui Gong, John Cartlidge, Yang Yue, Guoping Qiu, Qingquan Li, and Jingyu Xin. 2017. 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 (IWCTS'17),* 6 pages.
https://doi.org/10.1145/3151547.3151553



## 1 INTRODUCTION

Retail is an important driving force of urban development [4] and customers' shopping patterns affect both business and urban developers. There are many models that can be used to analyse retail trading areas, like gravity assumptions [11], discrete choice models [9], and logit models [7]. However, the calibration of spatial interaction models is a significant challenge. Without accurate calibration, a model is almost useless for either description or prediction [17]. Traditionally, researchers used survey and interview methods to calibrate models. However, these traditional methods are labour intensive, time consuming, and costly. As such, the amount of data that can be obtained is often limited and of low resolution. This leads to reduced model prediction accuracy.

One example of an automatically generated data stream is GPS taxi trajectory data. Taxis are a widely used mode of transport in most major cities, and their journeys can be automatically tracked. By regularly recording a taxi's GPS location and status (*occupied* or *free*), a significant coverage of a city's transportation network and behaviour can be automatically captured in high resolution, and at relatively little cost. GPS taxi trajectory data has previously been used to explore patterns in travel behaviour [24], urban design [21], human behaviour analysis [20], and taxi service promotion, such as customer searching model [22].

First introduced in 1964, the Huff model [11] is one of the most widely used models in retail trading analysis. Following simple gravity assumptions, the Huff model estimates the spatial probability distribution of shopping centre patronage based on a shopping centre's attractiveness (often considered as *size*) and the customer's cost of travel. Despite the simplicity of the model, if calibrated accurately, the Huff model has strong explanatory power [20, 23].

It has previously been shown that taxi trajectory data can be used to successfully calibrate the Huff model [23]. Here, we perform spatial calibration using taxi trajectory data from the emerging metropolitan city of Shenzhen. Using Geographical Weighted Regression (GWR) to fit Huff model parameters, we demonstrate that exponents of attractiveness, $\alpha$, and distance, $\beta$, are spatially variant. That is, in different regions of Shenzhen, shoppers exhibit different shopping behaviour tendencies.

To explain this variation, we hypothesise that more affluent shoppers are more sensitive to attractiveness and distance (i.e., shoppers with more wealth are: (1) more likely to be more selective in the centres they choose to patronise; and (2) more likely to be prepared to travel farther to patronise a centre). We test these hypotheses by using house price data for each local region of Shenzhen as a proxy of wealth, and perform Ordinary Least Squares (OLS) regression to explore whether there is a significant relationship. Results show that, while there appears to be no significant relationship between distance travelled and wealth, there *is* a strongly significant linear relationship between $\alpha$ (sensitivity to attractiveness) and wealth (average house price). That is, wealthy shoppers are more sensitive to (are more likely to choose) shopping centres that are large and popular—see equations (10) and (11). To our knowledge, this is the first time this relationship has been shown using real data.

We believe this is a very interesting new finding which can potentially have major impacts on city planning. Given that property prices around the world and in particular in booming economic regions have experienced rapid changes (rises) over the years, our finding suggests that similar studies should be carried out in other major metropolitan regions to include house prices as an important factor in planning city infrastructures and facilities, such as the choice of brands in the store, choice of target customer, and investment options for infrastructure constructions.

In Section 2, we review related work. In Section 3, the taxi data is introduced and our method of cleaning and categorising the data is described. In Section 4 we detail our first study: calibrating the Huff model using taxi data. In Section 5 we detail findings using factor analysis of house price data. Finally, Section 6 concludes.

## 2 RELATED WORK

The Huff model [10] is designed to estimate the spatial probability distribution of shopping centre patronage. The classic form is often written as follows:

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

where $P_{ij}$ represents the probability that a customer from origin $i$ will visit shopping cente $j$; $C_{ij}$ is the travel cost from $i$ to $j$; $S_j$ is the attractiveness of shopping centre $j$; $\alpha_i$ and $\beta_i$ are, respectively, the parameters of attractiveness and distance decay at each origin $i$, estimated from empirical observation; and $m$ is the total number of shopping centres.

Previous studies have shown that shopping centre size and proximity to competition are of significant importance [3], while distance is able to explain approximately 70% of the variation in actual retail sales at regional shopping centres [18]. Therefore, in this study, we use size as one of the factors considered for shopping centre attractiveness, and distance as the factor considered for cost.[1]

To estimate Huff model parameters $\alpha$ and $\beta$, four attraction and cost function combinations were introduced by O'Kelly [16], covering exponential and power influence of both variables. We label these, below, as $K1$ to $K4$ (equations (2) to (5)):

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

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

$$K3 : T_{ij} = exp(\alpha LnS_j - \beta C_{ij}) = S_j^{\alpha} exp(-\beta C_{ij}) \tag{4}$$

$$K4 : T_{ij} = exp(\alpha LnS_j - \beta LnC_{ij}) = S_j^{\alpha} C_{ij}^{-\beta} \tag{5}$$

where $T_{ij}$ equates to the numerator in equation (1). Note that K4 is identical to the traditional Huff model (1), with $\alpha$ and $\beta$ parameters a power of $S$ and $C$, respectively. In K1, which O'Kelly found to best fit the data [16], $\alpha$ and $\beta$ parameters are both exponents.

In addition, Nakanishi and Cooper [15] proposed the following equation (6), which is the log-transformed-centred form of OLS, to estimate parameters in Huff:

$$OLS : Ln(P_{ij}/\bar{P}_i) = \alpha_i Ln(S_j/\bar{S}) + \beta_i Ln(C_{ij}/\bar{C}_i) \tag{6}$$

where $\bar{P}_i$, $\bar{S}$ and $\bar{C}_i$ are the geometric means of $P_{ij}$, $S_j$ and $C_{ij}$ over j. To enable comparisons with the literature, we use all five estimation methods to fit one pair of Huff model parameters ($\alpha$ and $\beta$)—i.e., one global fit—and make comparisons between different time periods.

Subsequently, equations (2) to (6) are applied using Geographically Weighted Regression (GWR) to fit spatially variant parameters (i.e., to estimate best fit values for $\alpha$ and $\beta$ at each locale; rather than one global estimate for each). GWR is a non-stationary technique that models spatially varying relationships. Compared with a global regression, the coefficients in GWR are functions of spatial location [8]. The general form of a GWR model is:

$$y_i = \gamma_{i0} + \sum_{k=1}^{m} \gamma_{ik} x_{ik} + \epsilon_i \tag{7}$$

Where $y_i$ is the dependent variable at location $i$; $x_{ik}$ is the $k^{\text{th}}$ independent variable at location $i$; $m$ is the number of independent variables; $\gamma_{i0}$ is the intercept parameter at location $i$; $\gamma_{ik}$ is the local regression coefficient for the $k^{\text{th}}$ independent variable at location $i$; and $\epsilon_i$ is the random error at location $i$. Since the Huff model has two parameters to estimate, so $k = 2$, and $\gamma_{ik}$ in equation (7) corresponds to $\alpha$ and $\beta$ for each location $i$.

After fitting the Huff model using GWR (Section 4), we attempt to address the following issues:

(1) How do parameters $\alpha$ and $\beta$ change over time?
(2) How do $\alpha$ and $\beta$ vary geographically?
(3) How to explain the spatial-temporal variation of $\alpha$ and $\beta$?

## 3 DATA CLEANING

The taxi data includes taxi location (longitude, latitude), speed, direction angle, and status (0: taxi has no passenger; 1: taxi has passenger). For each taxi, the data collection time interval is between 30 seconds to 1 minute. Eight-days of taxi trajectory data (24 hours/day) were collected between 13–20 October 2013, capturing more than one million journeys from approximately 15,000 taxis. Although Sunday 13 October 2013 was a Chinese holiday, we confirm taxi data is consistent with Sunday 20 October 2013, so consider the holiday as a normal Sunday.

### 3.1 Taxi data and choice-based samples

Choice-based samples—groups that have chosen to visit a particular destination; i.e., shopping centres—are normally used in calibrating spatial interaction models [16, 19]. The ultimate goal of using

[1]Since we only consider one travel mode—metered taxi—we assume journey costs per unit distance are approximately constant (subject to fluctuations in traffic congestion).

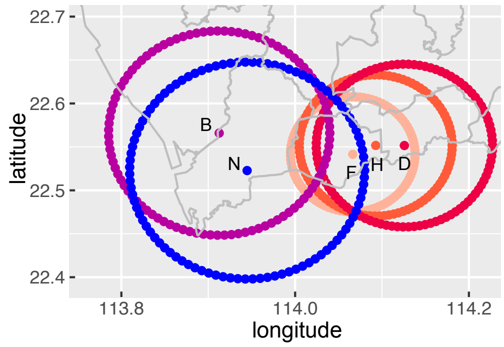



**Figure 1: Shopping centre trading areas, radii equal to maximum distance of closest 80% of journeys, for: Nanshan (N), Baoan (B), Dongmen (D), Huaqiangbei (H), and Futian (F).**

choice-based samples is to make inferences about the full population, therefore the samples must be representative and unbiased. As taxi fares are relatively expensive compared to other transportation modes, we are aware that the data are naturally biased on customer income and distance travelled. While we acknowledge this limitation, we believe the large data sample is representative enough to explain general shopping behaviours in Shenzhen.

## 3.2 Extracting choice-based samples

As the raw taxi data includes a large proportion of journeys that are not shopping activities, it is necessary to extract choice-based samples. We initially segment Shenzhen and all taxi data 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. Choice-based samples are then extracted using the method employed in [23]. First, we select all taxi drop-off points located within a 500 metres *buffer radius* of the five target shopping centres. It has been shown that the average walking trip of customers when they go to a shopping centre in China is 500 metres [6]. While this distance may seem large, and not immediately applicable to taxi journeys, it enables us to include journeys to smaller boutique shops and restaurants that tend to aggregate near shopping centres. Elsewhere, 300 metres has been used as a buffer radius [18]. We have repeated our analysis using a buffer radius of 300 metres and, although the model performs less well, results show no substantial difference.

Once drop-off (or *destination*) points have been selected, we then extract the respective taxi pick-up point (or *origin*) for each. We refer to these as Origin-Destination (O-D) pairs. As most of the shopping centres in Shenzhen open from 10am to 10pm, we filter to extract taxi O-D pairs in which GPS time is from 10am to 10pm.

Finally, we order all O-D pairs by distance of origin from target shopping centre, and then keep only the closest 80%. This follows Applebaum's process of defining a shopping centre's primary trading area [2]. Fig. 1 displays the primary trading areas—plotted as circle with radius equal to the maximum distance travelled for the closest 80% of journeys—of the five major shopping centres of Shenzhen that we consider in this paper. Plots are generated using R's *ggplot2* and *ggmap* packages. We see that Nanshan has the largest trading area, while Futian's trading area is smallest.

## 3.3 Training data and testing data

To verify the prediction accuracy of the calibrated Huff models, we split the data into two subsets: training data (used for calibrating the Huff model), and testing data (used to verify the prediction accuracy of the calibrated model). The full set of extracted choice-based data (see Section 3.2) is initially sorted by GPS time. Every tenth O-D pair is then selected for the test set, the other 90% of data are used for the Huff model calibration. Both sets are further segmented into five subsets categorised by time of day and day of week. These are: weekend, weekday all, weekday 10am–1pm, weekday 1pm–5pm, and weekday 5pm–10pm.

# 4 CALIBRATING THE HUFF MODEL

## 4.1 Methodology

The Huff model, equation (1), has two variables representing shopping centre attractiveness ($S$) and travel cost ($C$). To calculate $C$, we use O-D route distance returned from *Baidu.com*'s API. To calculate $S$, we consider two factors: (a) size of shopping centre ($S = S_{size}$), the traditional method used in the literature, e.g., [23]; and (b) number of journeys ($S = S_{journey}$), equivalent to *footfall*, calculated directly from the data.

To calibrate these models, we use the five parameter estimation methods (K1–K4, and OLS) introduced in Section 2. For example, we first use K1 to calibrate the Huff model. Five factors are considered: shopping centre attractiveness ($S = \{S_{size}, S_{journey}\}$), route distance, pick-up time, and pick-up locations. K1 is then linearised, as:

$$Ln(T_{ij}) = \alpha S_j - \beta C_{ij} \tag{8}$$

Weekend training data is then transformed into this format and substituted into the GWR model, shown in equation (7). We used R's *spgwr* package to perform GWR fitting. One global pair of best-fit parameter values are calculated for $\alpha$ and $\beta$, and a set of local best-fit parameter values $\alpha_i$ and $\beta_i$ are also calculated for each geographic location, $i$.

After calibrating the Huff model, the prediction accuracy is verified using Kullback-Leibler (KL) divergence on the test data [12]. KL-divergence is a statistical technique used to quantify the divergence between an expected probability distribution, $P$, and a generated probability distribution, $Q$. The general form of KL-divergence is expressed as:

$$D(P||Q) = \sum (P_i)log(P_i/Q_i) \tag{9}$$

Where $P$ is the observed patronage probability (in the test dataset) and $Q$ is the patronage probability predicted by the calibrated Huff model. A KL-divergence value of 0 represents perfect prediction accuracy; values close to 1 represent very poor prediction accuracy.

## 4.2 Results and Discussion

A summary of results of Huff model calibration and testing are presented in Table 1. For data during weekday working period (10am–5pm), the model calibration produces relatively high error

**Table 1: Huff model calibration and testing. Global calibration on training data using best fit model (highest $R^2$, lowest sum of squares); and KL-divergence of calibrated models on test data (smaller values indicate greater accuracy of model prediction).**

| Attractiveness | Time | Estimator | Global Huff Calibration | | | | | KL-Divergence | |
|---|---|---|---|---|---|---|---|---|---|
| | | | $\alpha$ | $\beta$ | Residual S.E. | $R^2$ | Sum of squares | Global | Local |
| Size | weekend | K1 | 0.062 | -0.236 | 0.025 | 0.763 | 0.206 | 0.51 | **0.22** |
| | weekday | K1 | 0.199 | -0.234 | 0.028 | 0.814 | 0.186 | 0.50 | **0.11** |
| Journey | weekend | K2 | 0.134 | -0.289 | 0.024 | 0.773 | 0.197 | 0.42 | **0.09** |
| | weekday | K2 | 0.126 | -0.281 | 0.028 | 0.814 | 0.185 | 0.40 | **0.05** |

rates. This is likely due to the proliferation of journeys during this period that are non-shopping related. Therefore, calibration results are only considered for weekend and weekday data after 5pm.

*4.2.1 Global parameter calibration.* Table 1 shows best fit calibration results when $S = S_{size}$ and $S = S_{journey}$.

When $S = S_{size}$, K1 estimator gives the best global calibration, with $\alpha = 0.062$ and $\beta = -0.236$ on weekends, and $\alpha = 0.199$ and $\beta = -0.234$ on weekday evenings. While there is little variation in $\beta$, we see that $\alpha$ is much higher on weekday evenings. Since $S = S_{size}$, this suggests that customers care more about shopping centre size on weekday evenings than they do at weekends. One interpretation of this could be that because customers have less time on weekday evenings, they prefer large shopping centres that are more likely to contain all of the items they would like to purchase. Conversely, at weekends, shopping centre size is less important as shoppers have more time to browse.

When $S = S_{journey}$, K2 estimator gives the best global calibration, with $\alpha = 0.134$ and $\beta = -0.289$ on weekends, and $\alpha = 0.126$ and $\beta = -0.281$ on weekday evenings. This suggests that customers' sensitivity to shopping centre popularity does not vary between weekday evenings and weekends.

Under both measures of attractiveness, K3 and K4 estimators have high error rates and return unintuitive negative estimates for $\alpha$. This suggests that the functional forms of K3 and K4 (in particular assuming $\alpha$ is a power rather than exponent of $S_j$) is *not* a good descriptor of our taxi data. The OLS estimator (6) also performs poorly. Therefore, we only consider calibration results from K1 and K2 regression estimators.

*4.2.2 Local parameter calibration.* Table 1 also shows the predictive accuracy (KL-divergence) of the calibrated models on the test data. We see that: (1) the Huff model with local parameters (i.e., calibrated using GWR; right-hand column) has greater predictive accuracy (lower KL-divergence) than Huff calibrated with one pair of global parameters; (2) the prediction accuracy of the model using attractiveness $S = S_{journey}$ is better than the model using $S = S_{size}$, suggesting that calculating the attractiveness of shopping centres directly from the number of taxi journey destinations (i.e., footfall) is more accurate than the traditional method of estimating attractiveness using shopping centre size; (3) the prediction accuracy is better for weekday evenings than weekends.

We consider each of the above results in turn. (1) Using GWR to fit parameter values locally offers much higher resolution than fitting one pair of global parameters. Therefore, unless the taxi data exhibits uniformity, GWR is always likely to give a more accurate result. (2) Since $S = S_{journey}$ is a direct measure of footfall measured from the taxi data, it is perhaps unsurprising that this direct measure provides more accurate prediction than the traditional low-resolution technique of using a shopping centre's size. We present this as direct evidence of the utility of calibrating the Huff model from taxi trajectory data. (3) At weekends, people generally have more time available for leisure activities than during weekday evenings. Therefore, at weekends, people may travel between multiple shopping centres while browsing for goods and/or integrate other leisure activities into their schedule. For this reason, with fewer shopping journeys originating at home (an explicit assumption of the Huff model), the weekend data set is more noisy and hence less predictable.

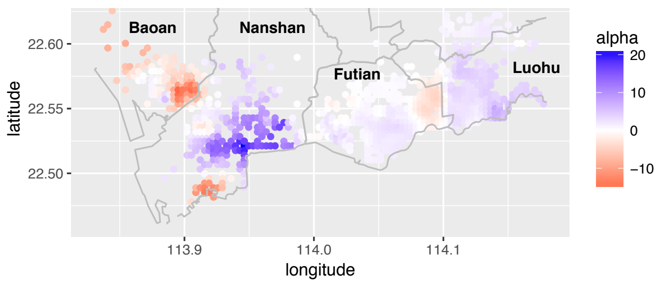
Fig. 2 presents the results of GWR calibration graphically. The districts mapped are: Baoan, Nanshan, Futian, and Luohu. Fig. 2a and Fig. 2c display qualitatively similar geographic variation in $\alpha$. In particular, $\alpha$ is positive highest (blue) in Nanshan district and negative highest (red) in Baoan district. This suggests that people who live in Nanshan are more likely to prefer shopping at attractive stores (the largest, and the most popular), whereas people who live in Baoan prefer the opposite—that is *unattractive* stores—possibly because they offer cheaper, low-quality goods. In Futian (where Huaqiangbei and Futian shopping centres are located) and Luohu (where Dongmen shopping centre is located), $\alpha$ values are close to 0. In these regions, shoppers pay less attention to the attractiveness of shopping centres when deciding where to shop.

In Futian the value of $\beta$ is consistently negative, indicating that people living in this area prefer to shop a short distance from home. Given it is a city centre district, this is perhaps unsurprising. In other regions, we see that $\beta$ tends to be negative when $S = S_{size}$, and positive when $S = S_{journey}$. This suggests that, in these regions, people are not prepared to travel far to visit a large shopping centre; but *are* prepared to travel far to visit a popular shopping centre. Perhaps this suggests that there is a greater homogeneity in large shopping centres, whereas popular shopping centres have more individuality.

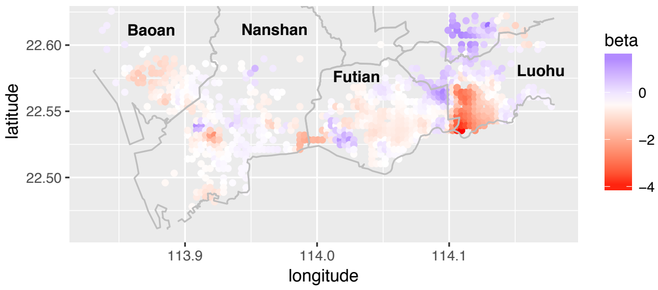
# 5 HOUSE PRICE ANALYSIS

## 5.1 Methodology

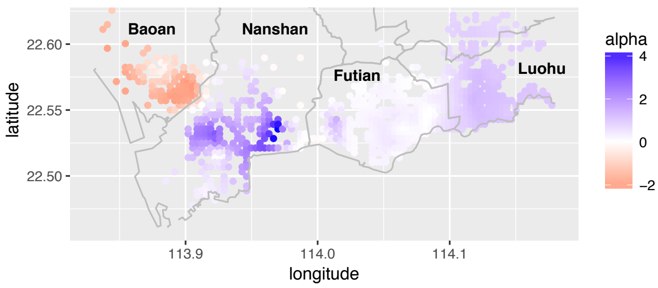
GWR demonstrates high geographic variability in sensitivity to $\alpha$ and $\beta$. To explain this variability, we hypothesise that more affluent shoppers are more sensitive to $\alpha$ and $\beta$ (i.e., shoppers with more wealth are (1) more likely to be more selective in the centres they choose to patronise, and (2) more likely to be prepared to travel



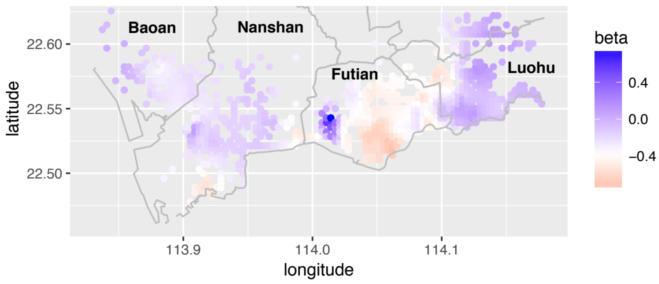

(a) Values of $\alpha$ for attractiveness $S = S_{size}$.




(b) Values of $\beta$ for attractiveness $S = S_{size}$.




(c) Values of $\alpha$ for attractiveness $S = S_{journey}$.




(d) Values of $\beta$ for attractiveness $S = S_{journey}$.

**Figure 2: GWR calibration (weekends) for Shenzhen. The four regions shown are: Baoan, Nanshan, Futian, and Luohu.**

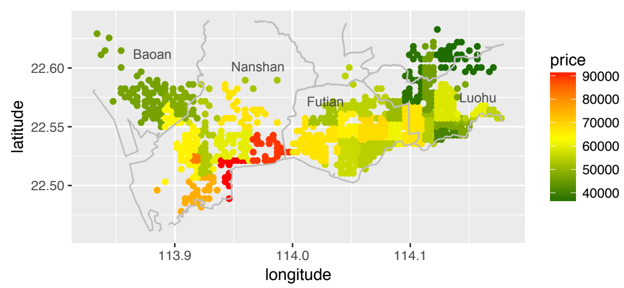



**Figure 3: House price distribution, $\pi$ (RMB/$m^2$), across four districts of Shenzhen: Baoan, Nanshan, Futian, and Luohu.**

farther to patronise a centre). Here, we test this hypothesis by using house price data for each region as a proxy of wealth, and perform OLS regression to explore whether there is a significant relationship.

House price data has previously been used in urban design, and a relationship between house prices and retail trade has been observed [1, 5, 13, 14]. However, to our knowledge, there is no previous research using house price data to provide an explanation for the spatial variance of people's shopping behaviours.

We collect second-hand house price sales' data for Shenzhen during the first quarter of 2017, and calculate average house price corresponding to each geographical cell in the segmented taxi data (1324 cells in total). All house price data were retrieved on 01 April 2017 from *Fang.com*, the largest and most comprehensive open-source repository for house price sales in China. The average house

**Table 2: OLS regression of best fit geographical Huff model on house prices**

| S | OLS regression | Value | S.E. | $\lvert T$ value$\rvert$ | $P$ value |
|---|---|---|---|---|---|
| size | house price ($\pi$) | 0.37 | 0.11 | 3.31 | $9.71 \times 10^{-4}$ |
| size | intercept | 0.25 | 0.03 | 8.17 | $<7.37 \times 10^{-16}$ |
| journey | house price ($\pi$) | 1.50 | 0.17 | 8.58 | $<7.37 \times 10^{-16}$ |
| journey | intercept | 0.56 | 0.05 | 11.73 | $<7.37 \times 10^{-16}$ |

price distribution for each area is presented in Fig. 3. There appears to be a clear visual correlation between values of $\alpha$ and house prices throughout Shenzhen. We attempt to quantify this relationship in the next section.

## 5.2 Results and discussion

We test whether local values of house price, $\pi$, are positively correlated to $\alpha$. Table 2 presents results of OLS regression, used to explore the linear relationship between these two variables. OLS regression was performed using R's *sp* package, with normalised $\alpha$ and $\pi$ values.

In all cases, $p$ and $T$ values evidence a statistically significant linear relationship between distribution of parameters $\alpha$ and $\pi$. The best-fit relationships are given by:

$$\alpha_{size} = 0.37\pi + 0.25 \tag{10}$$

and

$$\alpha_{journey} = 1.50\pi + 0.56 \tag{11}$$

As an individual's wealth increases—and therefore *capacity for consumption* grows—the attractiveness of a shopping centre becomes increasingly important when deciding where to shop. When $S = S_{journey}$ (equation (11)), $\alpha$ is much more sensitive to wealth—i.e., the multiplier of $\pi$, or gradient of best fit, is larger—than when $S = S_{size}$ (equation (10)). Therefore, as wealth and income increase, sensitivity to *popular* shopping centres grows more quickly than sensitivity to *large* shopping centres. This may suggest that wealthy individuals prefer more fashionable places.

The strong correlation that we have identified between wealth, $\pi$, and parameter of attractiveness, $\alpha$, has potential for considerable positive impact. For data scientists, transportation modellers, and urban planners, it suggests open source house price data can be used for accurate modelling of city-wide shopping behaviours. In particular, we believe urban planners could use this information for optimal location of new retail and residential centres. In addition, by monitoring the fluctuation of house prices in regions of a city, one may be able to predict changes in traffic flow and congestion, by modelling the expected resultant changes in shopping behaviour. City regions that are undergoing a period of *gentrificiation* may be likely to offer the best opportunities for this type of modelling.

## 6 CONCLUSIONS

We have presented a Huff model calibration for shopping behaviours in the city of Shenzhen, using taxi trajectory data containing approximately one million journeys over an eight day period. Results demonstrate that fitting the model geographically using GWR—returning local calibration values for each local region—provides much higher predictive power than using one global value for each parameter in the model. Where sufficient data are available for Huff model calibration, we therefore suggest that GWR is a superior regression technique that should be used more widely.

We also demonstrated that using the number of journeys to each shopping centre as a measure of a *attractiveness* provides greater predictive power than using *size* to measure attractiveness. Although size is easy to calculate and generally available, we suggest that where data are available, more direct measurements of a shopping centre's attractiveness should be used.

Finally, we presented results demonstrating that the geographical variation is consumers' behaviour can be largely explained by consumers' *wealth*. Using house price sales' data for each local region as a proxy for wealth, we showed a significant linear relationship between wealth and sensitivity to attractiveness. As far as we are aware, this is the first time this relationship has been shown using real world transportation data. We provide this as evidence that fusing data sources can enhance model interpretation and improve predictive power.

However, we acknowledge that there are limitations to the research presented here. In particular, we consider taxis as the only mode of transport in the city. This is clearly unrealistic. In future we would like to compare results against other transportation modes. We would also like to perform the following extensions: (1) data filtering to observe how people journey *between* shopping centres; (2) modelling changing house prices and the effects that has on urban and transportation planning; (3) using social media data (shopping reviews) as a measure of attractiveness.

## ACKNOWLEDGMENTS

This research is supported by Shenzhen key laboratory of Spatial Smart Sensing and Services in Shenzhen University that provided the taxi trajectory data set.

## REFERENCES

[1] Alex Anas. 1995. Capitalization of urban travel improvements into residential and commercial real estate: simulations with a unified model of housing, travel mode and shopping choices. *Journal of Regional Science* 35, 3 (1995), 351–376.

[2] William Applebaum. 1966. Methods for determining store trade areas, market penetration, and potential sales. *Journal of Marketing Research* 3, 2 (1966), 127–141.

[3] David R Bell, Teck-Hua Ho, and Christopher S Tang. 1998. Determining where to shop: Fixed and variable costs of shopping. *Journal of Marketing Research* 35, 3 (1998), 352–369.

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

[5] João O Borba and Tomaz Ponce Dentinho. 2016. Evaluation of urban scenarios using bid-rents of spatial interaction models as hedonic price estimators: an application to the Terceira Island, Azores. *The Annals of Regional Science* 56, 3 (2016), 671–685.

[6] Wuhan Transportation Planning Bureau. 2009. *Wuhan traffic impact analysis guideline*. Technical Report. Technical report, Wuhan Transportation Planning Bureau.

[7] Chaushie Chu. 1989. 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. World Conference on Transport Policy, Yokohama, Japan, 295—309.

[8] A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. 2003. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, University of Newcastle, UK.

[9] Azucena Gracia and Tiziana de Magistris. 2008. The demand for organic foods in the South of Italy: A discrete choice model. *Food Policy* 33, 5 (2008), 386–396.

[10] David L Huff. 1963. A probabilistic analysis of shopping center trade areas. *Land economics* 39, 1 (1963), 81–90.

[11] David L Huff. 1964. Defining and estimating a trading area. *The Journal of Marketing* 28, 5 (1964), 34–38.

[12] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[13] J. D. Landis. 2012. Modelling Urban Systems. In *The Oxford handbook of urban planning*, Randall Crane and Rachel Weber (Eds.). Oxford University Press, Oxford, UK, 323–351.

[14] Ming-Long Lee and R Kelley Pace. 2005. Spatial distribution of retail sales. *The journal of real estate finance and economics* 31, 1 (2005), 53–69.

[15] Masao Nakanishi and Lee G Cooper. 1982. Technical Note—Simplified Estimation Procedures for MCI Models. *Marketing Science* 1, 3 (1982), 314–322.

[16] Morton Edward O'Kelly. 1999. Trade-area models and choice-based samples: methods. *Environment and Planning A* 31, 4 (1999), 613–627.

[17] Jean-Paul Rodrigue, Claude Comtois, and Brian Slack. 2013. *The geography of transport systems*. Routledge, New York, NY 10017.

[18] Anja Simma, Pietro Cattaneo, Myriam Baumeler, and Kay W Axhausen. 2004. *Factors influencing the individual shopping behaviour: The case of Switzerland*. Technical Report. ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau, Zürich.

[19] Fotheringham A Stewart and Morton E O'Kelly. 1989. Spatial interaction models: formulations and applications. (1989).

[20] Jinjun Tang, Han Jiang, Zhibin Li, Meng Li, Fang Liu, and Yinhai Wang. 2016. A Two-Layer Model for Taxi Customer Searching Behaviors Using GPS Trajectory Data. *IEEE Transactions on Intelligent Transportation Systems* 17, 11 (2016), 3318–3324.

[21] Wei Tu, Qingquan Li, Zhixiang Fang, Shih-lung Shaw, Baoding Zhou, and Xiaomeng Chang. 2016. Optimizing the locations of electric taxi charging stations: A spatial–temporal demand coverage approach. *Transportation Research Part C: Emerging Technologies* 65 (2016), 172–189.

[22] RCP Wong, WY Szeto, and SC Wong. 2014. A cell-based logit-opportunity taxi customer-search model. *Transportation Research Part C: Emerging Technologies* 48 (2014), 84–96.

[23] Yang Yue, Han-dong Wang, Bo Hu, Qing-quan Li, Yu-guang Li, and Anthony GO Yeh. 2012. Exploratory calibration of a spatial interaction model using taxi GPS trajectories. *Computers, Environment and Urban Systems* 36, 2 (2012), 140–153.

[24] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, ACM, New York, NY, USA, 312–321.