Validating Gravity-Based Market Share Models Using Large-Scale Transactional Data

Yoshihiko Suhara · Mohsen Bahrami ·
Burçin Bozkaya · Alex 'Sandy' Pentland

Abstract Customer patronage behavior has been widely studied in market share modeling contexts, which is an essential step towards modeling and solving competitive facility location problems. Existing studies have conducted surveys to estimate merchants' market share and factors of attractiveness to use in various proposed mathematical models. Recent trends in Big Data analysis allow us to better understand human behavior and decision making, potentially leading to location models with more realistic assumptions. In this paper, we propose a novel approach for validating Huff gravity market share model, using a large-scale transactional dataset that describes customer patronage behavior in a regional scale. Although the Huff model has been well-studied and widely used in the context of competitive facility location and demand allocation, this paper is the first in validating the Huff model with a real dataset. Our approach helps to easily apply the model in different regions and with different merchant categories. Experimental results show that the Huff model fits well when modeling customer shopping behavior for a number of shopping categories including grocery stores, clothing stores, gas stations, and restaurants. We also conduct regression analysis to show that certain features such as gender diversity and marital status diversity lead to stronger validation of the Huff model. We believe we provide strong evidence, with the help of real world data, that gravity-based market share models are viable assumptions for competitive facility location models.

Y. Suhara, M. Bahrami, A. Pentland
The Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
E-mail: suhara@mit.edu

M. Bahrami, B. Bozkaya
Sabanci University, Istanbul, Turkey

1 Introduction

During past decades and especially by the advent of the new machinery and technologies, the number of companies has been increasing dramatically, which has led to a highly competitive business environment. For example, in the UK there has been a sustained growth in total business population with a 64% growth rate since 2000, and the number of companies has continuously increased during recent years. In 2016, it has increased by 197,000, which is equal to 4% growth [1]. To compete in such an environment, perhaps the biggest challenge for companies is to "optimally" locate their facilities to capture more demand and market share, while trying to alleviate the burden of their fixed and operational costs. This makes facility location decisions of critical importance to companies as such decisions must take into account the market environment to operate in and consumers' preferences.

For decades, companies have been trying to understand how customers are attracted to retail businesses so as to make effective decisions about where to open a new store of their chain. To address this challenge, a vast literature on facility location models has emerged. Many facility location models focus on locating retail stores in a competitive environment. An overall aim of such models is to maximize the market share captured as a result of the new locations opened, and consequently to maximize the profitability of the company for its shareholders [Jensen(2001)][Wernerfelt(1986)]. Hence, decision makers must understand and model the underlying processes for retail patronage before the facility location models can be solved effectively.

Among various models of competitive facility location that are developed and available in the literature [Berman et al(2009)Berman, Drezner, Drezner, and Krass][Eiselt et al(1993)Eiselt, Laporte, and Thisse][Plastria(2001)][Drezner(1995)], we are especially interested in those with underlying market share models that consider customer shopping behavior and retail patronization. The main goal of patronization models is to derive a realistic estimate of how and where people shop, and consequently a retail facilitys market share. These models assume that the patronage behavior is influenced by multiple factors such as the retail facilitys attractiveness to customers, distance from customers location, and customers' purchasing power [Drezner(2006)]. Among various market share estimation approaches proposed, five main ones include proximity [Hotelling(1929)], deterministic utility [Drezner(1994)], random utility [Leonardi and Tadei(1984)] [Drezner and Drezner(1996)], cover-based [Drezner et al(2010)Drezner, Drezner, and Kalczynski], and gravity-based [Huff(1964)] approaches.

The first and the simplest approach is the proximity approach, which only considers the distance factor. Hotelling was the first to propose and use this model [Hotelling(1929)]. Based on this model, a customer is more likely to patronize the facility closer to his or her location. The second approach is the deterministic utility approach introduced by Drezner [Drezner(1994)], which

---

[1] http://www.fsb.org.uk/media-centre/small-business-statistics/

suggests that customers are attracted differently to retail facilities. Therefore, proximity only is not sufficient anymore and a utility value is defined for allocation of the customers to the facilities. Yet, customers are assumed to spend most at the facility that is most attractive to them. The third kind of model was introduced to address the problem of all or nothing in deterministic utility models. The random utility model is an extension of the deterministic utility model, where the utility of the customer is drawn from a multivariate normal distribution of utility function [Leonardi and Tadei(1984)][Drezner and Drezner(1996)]. The fourth is the cover-based approach, where for each facility an influence circle with a certain radius is defined based on its attractiveness. Customers inside the circle are fully attracted by the facility in the center and those outside the influence circles of all the facilities are considered as 'lost sales' [Drezner et al(2010)Drezner, Drezner, and Kalczynski].

The fifth and the most extensively used approach is the gravity-based approach. Estimating market share based on this approach was first introduced by Reilly [Reilly(1931)] and further extended by Huff [Huff(1964)][Huff(1996)]. The Huff model approximates the probability of a customers patronization of a particular retail facility based on two factors: attractiveness and distance. This means that the more attractive shops (based on various relevant criteria) draw more customers, and people tend to visit shops closer to where they live or work. It is common in Huff-based models to approximate the market share of each facility based on the total number of visits or the total money customers spend, which translate into the calculated probabilities of patronizing each facility. Nakanishi and Cooper [Nakanishi and Cooper(1974)] further propose an improved Huff model by developing a multiplicative competitive interaction (MCI) model that combines multiple dimensions of attractiveness into a single measure. Many extensions of the Huff model proposed by other researchers using different attractiveness factors and distance decay functions are also proposed [Bozkaya et al(2010)Bozkaya, Yanik, and Balcisoy][Berman and Krass(2002)][Aboolian et al(2007)Aboolian, Berman, and Krass][Hodgson(2007)].

To explore the nature of customer behavior and patronization choice, Drezner applies the Huff model as part of a behavioral analysis based on manually collected survey information [Drezner(2006)]. The survey uses a set of merchants in Orange County, California and tracks subjects to analyze how and why these customers patronize these merchants. Her metric for verifying the estimated attractiveness levels derived from the survey data is the correlation between the theoretical model's results and the estimation based on the survey, where a high correlation was reported.

In this study, we take an approach similar to Drezners to model customer retail patronization, but this time relying on real transaction data collected from tens of thousands of customers credit card activities. The recent rise of Big Data analysis has led to many similar studies trying to model and understand urban-scale human behavior based on call records [Isaacman et al(2011)Isaacman, Becker, Cáceres, Kobourov, Martonosi, Rowland, and Varshavsky][Blondel et al(2012)Blondel, Esch, Chan, Clerot, Deville, Huens,

Morlot, Smoreda, and Ziemlicki][Bogomolov et al(2014)Bogomolov, Lepri, Staiano, and Oliver], credit card transactions [Singh et al(2015)Singh, Bozkaya, and Pentland][Hasan et al(2012)Hasan, Schneider, Ukkusuri, and González], GPS traces [Zheng et al(2013)Zheng, Liu, and Hsieh], etc. The Huff model is a very popular model used both in research, appearing in 27 out of 55 articles on competitive facility location modeling as reported by [Drezner(2014)], and in business/retail applications [Huff and McCallum(2008)], yet to the best of our knowledge, there is no research on using real transactional data sets to test or validate Huff or similar gravity-based models. In contrast to the previous studies based on survey data for understanding shopping behavior, this paper presents a novel data-driven approach for patronage behavior analysis based on real-world transaction records. We believe such an approach also alleviates the limitations of survey-based studies related to data collection and data quality, by using readily available data that reflect real human behavior as opposed to drawing conclusions based on stated preferences of consumers.

The contribution and advantages of our approach include the following:

- Competitive facility location models can now benefit from the validated use of Huff or similar gravity based models for better representation of reality in retail patronization and market share estimation.
- One can consider using the Huff model for distinct merchant categories to compare its performance across various categories.
- Our analysis reveals that the presence or lack of certain demographic features, such as gender diversity or marital status diversity, lead to better validation of the Huff model.
- Merchants and business owners can implement our validation approach in different geographical regions with different settings so that retail location decisions can be used with higher reliability.
- It is computationally inexpensive to perform our validation approach on a different transactional data set. This eliminates the need and associated costs to conduct surveys for data collection under different settings.

The rest of our paper is organized as follows. In Section 2, we present our validation methodology. Next we present the results of our validation using the transactional dataset. Finally, we provide concluding remarks and directions for future research.

2 Methodology

2.1 The Huff Model

The Huff model [Huff(1964)] is an economic model for estimating market share in relation to customer retail patronization decisions. This model is based on gravity models [Reilly(1931)][Anderson(2010)], which describe the magnitude of interaction by two factors, namely mass and distance. The Huff model uses merchant *attractiveness* for the mass factor and the *distance* between a customer and a merchant for the distance factor. The utility of customer i visiting

merchant j can be formally defined as follows:

$$H_{ij} = \frac{A_j^\alpha}{D_{ij}^\beta}, \tag{1}$$

where A_j is the attractiveness of merchant j, D_{ij} is the distance between customer i and facility j, and the parameters α and β are used to adjust the sensitivity of the model to the two factors. To obtain the probability of customer i visiting merchant j, H_{ij} is normalized by the sum of all utility values for possible visits:

$$P_{ij} = \frac{H_{ij}}{\sum_{j' \in c(i)} H_{ij'}}, \tag{2}$$

where $c(i)$ denotes the set of merchants that customer i could potentially visit.

2.2 Data

In this study, a large amount of credit card transaction records were used for the designed experiment. The dataset was collected from a major bank in a major city of an OECD country between July 2014 and June 2015. The transaction records contain customer IDs, spending amounts, merchant IDs, and their merchant categories. We filtered customers who have at least 10 transactions in the dataset. The statistics of the dataset are shown in Table 2.2.

Table 1 Dataset summary.

Period	July 1, 2014 to June 30, 2015
# of transactions	4,254,652
# of customers	62,392
Avg. transactions / customer	68.19

Customer information. For each customer in the dataset, we have the following demographic information:

– age;
– gender;
– marital status;
– education level;
– work status (employed by private sector, self-employed, etc.);
– income (as estimated by the bank);
– home location;
– work location.

Merchant Category. Table 2 shows the selected merchant categories and their corresponding number of transactions and descriptions. We have chosen these categories to compare patronage behavior over different types of merchants. Customers tend to visit grocery stores more often than other categories. During our experiment, we evaluated the consistency and inconsistency between these categories.

Table 2 Basic statistics of the top-4 most frequented merchant categories.

Merchant category	# of transactions
Grocery store (Grocery)	1,089,614
Gas station (GS)	482,178
Clothing store (Clothing)	437,760
Restaurants	185,595

2.3 Experimental Setting

We split the dataset into 17 regions based on the administrative districts of the city of interest. The Huff model was fitted to each region for each merchant category. Therefore, we created 68 (17 regions x 4 merchant categories) Huff models for the experiment. For each region, we had a set of merchants belonging to the corresponding categories and a set of customers who visited these merchants. This resulted in the creation of a visit-count matrix $V_{ij}^{(r,c)}$, which consists of the visit count of customer i to merchant j that belongs to merchant category c and is located in region r.

Revenue Estimation for Attractiveness. We use the total revenue of merchant j in the dataset as the magnitude of its overall economic presence, and hence as its attractiveness measure (A_j.) Although total transaction count of merchant j is an alternative option, the total revenue is more appropriate from the facility location perspective. In other words, a company tries to choose the right location for a new facility in order to maximize the profitability. As a result, the revenue information well represents the profitability and attractiveness of a merchant. We aggregate the transaction amount of a merchant by all customers as an approximated revenue of each merchant.

Parameter Estimation. Parameters α and β are optimized to maximize the evaluation metric through Particle Swarm Optimization (PSO) technique [Kenndy and Eberhart(1995)] within the range of $\alpha, \beta \in [0, 100]$. We also tested the OLS method [Nakanishi and Cooper(1974)], which is commonly used for the Huff model parameter estimation. However, the method significantly under-performed the PSO method. Therefore, we used PSO for parameter estimation in this experiment. To the best of our knowledge, this paper is the first to use PSO or any kind of derivative-free optimization technique to optimize the α and β parameters of the Huff model.

Evaluation metric. We use Pearson's correlation between the estimated visit distribution calculated by the Huff model and the actual transaction-based visit distribution of a region as an evaluation method. The fitted model outputs the probability of a customer visiting a merchant, resulting in a visiting probability matrix $P_{ij}^{(r,c)}$ whose (i, j)-element is the probability of customer i visiting merchant j of category c in region r.

We also conduct regression analysis to find significant indicators of the computational model performance. In this part of our analysis, we compare the following indicators and conduct regression analysis to verify their contributions to the computational model performance.

Mobility diversity. We define the mobility diversity of a district i as the entropy value of visited districts for shopping. That is, for a given district i, we aggregate the transactions of all customers in that district by all districts in the region where the customers purchased items. A higher entropy value indicates that customers living in a district visit diverse areas for shopping.

Demographic diversity. For demographic diversity, we use *gender*, *marital status*, *education level*, and *job status* attributes of customers living in a district. For each district, we aggregate the demographic attribute counts to calculate the diversity of each attribute. We use entropy as a diversity metric.

Merchant diversity. We calculate the entropy value of merchant category distribution for each district. If a district has exactly the same number of merchants for each merchant category, the entropy takes the highest value. We prepared this merchant diversity metric following the intuition that a skewed distribution of merchant categories possibly affects patronage behavior in a district or region.

Merchant share bias. We calculate merchant share bias based on the market share of the top-5 merchant shares in a district. We calculate the total transaction amount of merchants for each district and then we divide the total amount of the top-5 merchants by the total amount of all transactions in a district.

Income inequality. Based on the income information reported to the bank, we calculated the Gini coefficient of income distribution for each district for income inequality. As some customers reported their income as 0, we filter them out in order to get a reliable analysis.

We consider the indicators described above as independent variables and the Huff model performance value as a dependent variable. We concatenated all district results to create a dataset with 68 (17 regions x 4 merchant categories) samples for the regression analysis. All the variables were standardized by converting into z-scores for easier interpretation.

3 Results and Discussion

3.1 Model Performance

Table 3 summarizes the basic statistics of the results and Figure 1 shows the boxplots of the model performance distributions for each merchant category. Detailed results including Pearson's correlation and the optimized parameter of each district are shown in the Appendix.

Table 3 Huff model performance for each merchant category.

Merchant category	Mean	Std	Max	Min
Grocery	0.8935	0.1068	0.9850	0.6753
GS	0.9050	0.1011	0.9928	0.6595
Clothing	0.8852	0.0930	0.9924	0.6916
Restaurant	0.7586	0.3418	1.0000	-0.0370

Fig. 1 Distribution of model performance (Pearson's correlation) values for each merchant category.

The models in all categories perform well as their mean/median values of Pearson's correlation values exceed 0.7 regardless of merchant categories. Except for the Restaurant category, the Pearson's correlation metric of the worst-performing district of each category is still above 0.65. The results indicate that the Huff models based on transaction data robustly capture customer patronage behavior in these categories.

However, the Huff model in the Restaurant category has relatively unstable performance compared to the other categories. Four districts have less than 0.5 Pearson correlation values and the worst performance shows -0.037. We

consider that the main reason of the unstable performance of the Huff model in the Restaurant category arising from the fact that customers' patronage behaviors do not fully follow the Huff model's assumption. That means that people often choose to go to restaurants in distant locations with various attractiveness measures (other than the total revenue of the merchant) that are not captured in our model. One can view restaurant patronage as a more hedonic way of "shopping" experience, where customers with a variety of tastes and expectations may choose to patronize a variety of places around the city to fulfill their expectations.

To verify our interpretations, we analyze the distribution of the distance between visited merchants of these four categories and customers' home/work locations (closer location is taken.) The distributions are shown in Figure 2. As shown in Figure 2(a), 2(b), and 2(c), the visited merchants of the Grocery, GS and Clothing categories are basically located close to the customers' home/work locations while the distance distribution of restaurants contains long distance values as shown in Figure 2(d). The results support our interpretation of the Huff model performance and also show a limitation of modeling patronage behavior with the Huff model based on transaction data. Despite this argument, we see that the model performance for the Restaurant category still suggests that the Huff model based on transactional data still performs reasonably well in several districts since the Huff model's performance in 13 out of 17 districts is higher than 0.5.

The patronage behavior of gas stations is a great example of the Huff model being used on transaction data, among the four categories considered. We observe that the most frequented gas stations are in close proximity to the customers' home/work locations. Moreover, the mean value of the Huff model performance in the GS category is highest (0.905.) The result confirms the fact that customers often do stop by their popular gas stations in the vicinity of their home/work locations.

3.2 Regression Analysis

Table 4 shows the adjusted R^2 values of regression analysis for all four merchant categories. As shown in the table, the diversity measure indicates reasonably high Huff model performance for the Grocery and Clothing categories. On the other hand, the regression models do not perform well in establishing a link between diversity measures and Huff model performance for the GS and Restaurant categories. Furthermore, the regression models do not show any statistical significance in the β coefficient values of the diversity indicators for these two categories.

Table 5 shows the β coefficient output of the regression models for (a) Grocery and (b) Clothing categories. The table summarizes the β coefficient value of each indicator with 95% confidence interval. We show the regression analysis results of the GS and Restaurant categories in the Appendix since
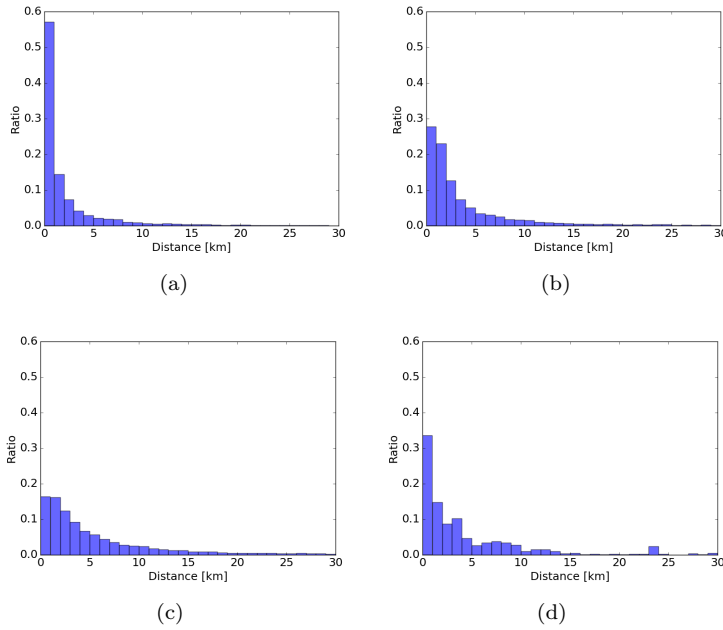
Fig. 2 Distribution of the distance between visited merchants and customer's home/work locations (closer location is taken) of each merchant category: (a) Grocery, (b) GS, (c) Clothing, and (d) Restaurant

Table 4 Adjusted R^2 values of the regression analysis results.

Merchant category	Adjusted R^2 score
Grocery	**0.638**
GS	-0.281
Clothing	**0.557**
Restaurant	0.150

we did not confirm any statistical significance in all the indicators for these categories.

The bold-faced values (i.e., statistically significant coefficients) in Table 5 suggest that gender diversity is positively correlated with the Huff model performance while marital status diversity is negatively correlated. A high gender diversity value means that males and females are equally distributed in a district. The gender diversity takes the highest value when the male/female ratio is one. In other words, a skewed distribution of male/female customers in a district makes the gravity model difficult to fit. On the other hand, marital status diversity is negatively correlated with the Huff model performance. The result follows our intuition that single and married customers have significantly different shopping styles. That is, the Huff model cannot simply generalize the patronage behavior in a district as the marital status diversity increases within the district.

No significant statistical results can be seen in other indicators like mobility, merchant, and income diversity. Originally, we hypothesized that the mobility diversity and the merchant diversity would correlate with the Huff model performance. For instance, the mobility diversity is a direct indicator of the lifestyle of customers living in a district. Therefore, we would hypothesize that a high mobility diversity value of a district would make the Huff model difficult to fit.

Although the Huff model works well for the GS category, the regression analysis does not perform well in the GS category. Our interpretation of this is that the diversity features we propose are simply not indicative of the model performance across various districts in the region.

Table 5 OLS regression model between Huff model performance (i.e., the Pearson's correlation values) and indicators. *, ** denote $p < 0.05, 0.01$ respectively. Bold face denotes the β coefficient is statistically significant.

(a) Grocery

Indicator	β coefficient	Confidence interval (95%)
Mobility diversity	-0.1799	$[-0.6516, 0.2917]$
Merchant diversity	-0.2038	$[-0.7601, 0.3524]$
Merchant monopoly	0.0586	$[-0.3650, 0.4822]$
Gender diversity	**2.5007****	$[1.1776, 3.8239]$
Marital status diversity	$-$**2.4411****	$[-3.6434, -1.2388]$
Education level diversity	-0.3585	$[-0.8686, 0.1516]$
Job status diversity	**0.5106***	$[0.0858, 0.9355]$
Income inequality	0.2643	$[-0.2817, 0.8103]$

(b) Clothing

Indicator	β coefficient	Confidence interval (95%)
Mobility diversity	0.1355	$[-0.3742, 0.6453]$
Merchant diversity	0.5748	$[-0.0263, 1.1760]$
Merchant monopoly	-0.4263	$[-0.8842, 0.0315]$
Gender diversity	**1.4881***	$[0.0581, 2.9181]$
Marital status diversity	$-$**1.3081***	$[-2.6076, -0.0087]$
Education level diversity	$-$**0.8321****	$[-1.3834, -0.2808]$
Job status diversity	0.2080	$[-0.2511, 0.6672]$
Income inequality	-0.3274	$[-0.9174, 0.2627]$

We have also conducted another type of mobility analysis to understand the differences regarding customers' shopping behavior for each merchant category. Fig. 3 shows the distributions of mobility patterns for each category. The x-axis and y-axis of these figures represent the district of a merchant and the district of a customer's home location, respectively. The numbers are normalized by row. For instance, the cell (i, j) is the normalized transaction frequency of merchants located in district j by customers who live in district i. It is intuitive that the diagonal line basically has the highest values as customers mostly visit merchants in the same district as they live. However, Fig. 3 (d) shows

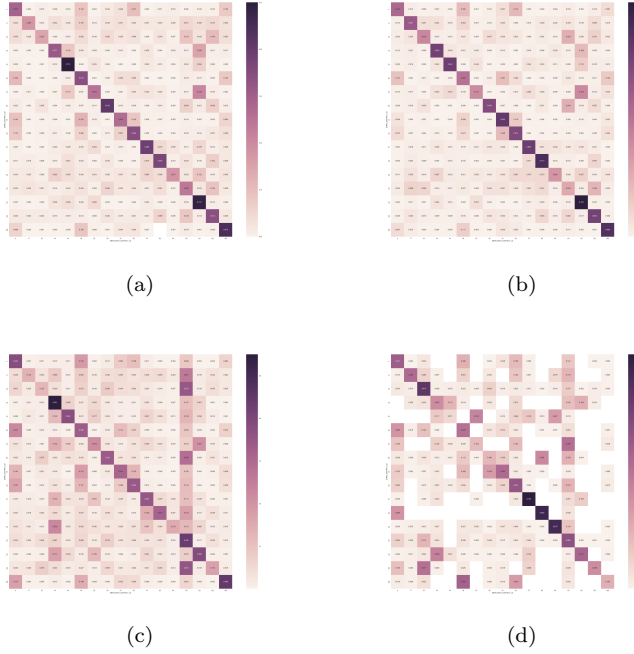that merchants in the Restaurant category have more biased distributed with respect to the mobility patterns.

(a)

(b)

(c)

(d)

Fig. 3 Mobility patterns of each merchant category: (a) Grocery, (b) GS, (c) Clothing, and (d) Restaurant.

4 Summary and Conclusions

In this paper, we have proposed a novel approach in validating a widely used customer retail patronage and market share estimation model, namely the gravity-based Huff model, using transactional data. Our approach applies the Huff model that consists of the attractiveness and distance factors to explain customer behavior. Our computational results have shown that the Huff model performs well in terms of the Pearson's correlation value calculated between the predicted market share and real market share.

Our study is the first to validate the Huff model with a large-scale transactional dataset to produce realistic representation of customer patronage behavior. Since gravity-based models such as Huff are widely used in competitive facility location optimization, our study provides key insight for reliable use of Huff or other gravity-based models in facility location optimization. Compared to the conventional survey-based approaches, the major advantages of

our transaction-based approach include: (a) no requirement for surveys where data collection cost and data quality might be an issue, (b) the ability to directly compare different categories of shopping, and (c) ease of computational implementation in terms of computational complexity and time so that the model parameters can be updated in a periodic manner (e.g., daily, weekly, quarterly etc.) and also with different data sets.

As we have shown in our analysis, the performance of the Huff model varies between categories. For certain categories, additional criteria may need to be included in attractiveness calculations, or human behavior may simply be too complex to model using a gravity-based approach. However, we believe that our approach provides various benefits which cannot be obtained from the conventional (survey-based) approaches. On the other hand, we would like to emphasize that survey-based approaches can collect more fine-grained information and these two approaches can be complementary to one another. In this regard, combining survey-based information and transaction-based information to build a sophisticated shopping behavior model would be a future direction. Another possibility for future study might be validating and comparing the performance of other market share estimation models with the Huff model.

Acknowledgements

Supporting Information

References

Aboolian et al(2007)Aboolian, Berman, and Krass. Aboolian R, Berman O, Krass D (2007) Competitive facility location and design problem. European Journal of Operational Research 182(1):40–62

Anderson(2010). Anderson JE (2010) The gravity model

Berman and Krass(2002). Berman O, Krass D (2002) Locating Multiple Competitive Facilities - Spatial Interaction Models with Variable Expenditures. Annals OR

Berman et al(2009)Berman, Drezner, Drezner, and Krass. Berman O, Drezner T, Drezner Z, Krass D (2009) Modeling Competitive Facility Location Problems: New Approaches and Results. In: INFORMS 2009 Tutorials in Operations Research: Decision Technologies and Applications, pp 156–181

Blondel et al(2012)Blondel, Esch, Chan, Clerot, Deville, Huens, Morlot, Smoreda, and Ziemlicki. Blondel VD, Esch M, Chan C, Clerot F, Deville P, Huens E, Morlot F, Smoreda Z, Ziemlicki C (2012) Data for Development: the D4D Challenge on Mobile Phone Data. arXivorg 1210.0137v2

Bogomolov et al(2014)Bogomolov, Lepri, Staiano, and Oliver. Bogomolov A, Lepri B, Staiano J, Oliver N (2014) Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of the 16th International Conference on Multimodal Interaction

Table 6 Grocery

districtid	Avg. distance	α	β	Pearson r	p-value
4	8.073920	44.092704	54.290707	0.957693	1.815968e-50
17	7.193814	50.098677	100.000000	0.833162	6.116027e-14
33	9.475784	25.459851	100.000000	0.450857	2.355955e-04
34	8.946319	0.776089	0.026872	0.985043	5.379011e-46
42	9.842346	0.869327	0.486844	0.968569	1.247863e-61
50	8.526481	0.879940	0.000000	0.978794	3.661174e-134
53	10.068181	44.430748	82.520949	0.906168	2.923771e-14
59	9.293074	0.813904	0.000000	0.981842	3.553790e-55
61	8.975816	0.727393	1.654360	0.884556	3.198309e-44
63	8.391520	40.762860	100.000000	0.923845	3.488153e-63
70	7.241697	3.612020	31.386009	0.632991	1.179105e-08
82	6.566708	0.524779	1.277479	0.898500	5.617862e-30
84	9.907927	0.581690	0.923136	0.715644	6.018784e-14
115	8.304005	29.783773	100.000000	0.667452	2.556054e-17
121	7.396679	0.858752	0.044139	0.965963	1.453555e-80
123	6.220508	0.277481	1.423520	0.781789	1.194598e-22
160	8.681894	34.661848	56.752011	0.953576	1.277798e-69

Table 7 Gas stations

districtid	Avg. distance	α	β	Pearson r	p-value
4	11.006358	0.592850	0.023663	0.940351	2.206558e-09
17	10.139055	0.988860	1.424924	0.978365	1.124570e-11
33	12.166465	40.288062	100.000000	0.978458	3.783059e-03
34	12.719812	0.502571	0.851630	0.992763	1.772337e-11
42	15.457192	0.983946	0.106719	0.974016	2.275157e-14
50	11.539085	0.497945	0.310459	0.952412	2.336259e-13
53	12.977294	0.698021	0.000000	0.867928	1.132502e-02
59	11.643731	35.637848	100.000000	0.837396	2.501177e-03
61	11.052237	0.711459	0.069953	0.896547	1.335805e-09
63	10.527793	23.877870	100.000000	0.659504	4.554565e-04
70	7.641760	0.507755	0.000000	0.930629	4.876992e-07
82	9.296232	23.872853	100.000000	0.980732	7.087527e-10
84	11.699089	0.698974	11.551928	0.966626	1.264309e-06
115	11.732630	3.930043	2.705459	0.902851	5.778257e-05
121	11.479532	0.453450	0.000000	0.895071	1.881786e-08
123	8.481359	0.037286	0.478554	0.666772	9.202447e-03
160	11.430354	0.748378	0.112061	0.964259	7.938458e-12

Bozkaya et al(2010)Bozkaya, Yanik, and Balcisoy. Bozkaya B, Yanik S, Balcisoy S (2010) A GIS-Based Optimization Framework for Competitive Multi-Facility Location-Routing Problem. Networks and Spatial Economics 10(3):297–320

Drezner(1994). Drezner T (1994) Locating A Single New Facility Among Existing, Unequally Attractive Facilities. Journal of Regional Science 34(2):237–252

Drezner(2006). Drezner T (2006) Derived attractiveness of shopping malls. IMA Journal of Management Mathematics 17(4):349–358

Drezner(2014). Drezner T (2014) A review of competitive facility location in the plane. Logistics Research 7(1):114–12

Drezner and Drezner(1996). Drezner T, Drezner Z (1996) Competitive Facilities: Market Share and Location With Random Utility. Journal of Regional Science 36(1):1–15

Table 8 Clothing

districtid	Avg. distance	α	β	Pearson r	p-value
4	9.158209	0.815865	1.191384	0.933204	3.323991e-63
17	11.612848	1.129118	0.670492	0.970486	4.455518e-25
33	11.131178	1.111029	2.365670	0.726212	1.762807e-16
34	12.620926	1.014079	0.438777	0.854472	3.350070e-74
42	17.460296	2.551279	3.436768	0.964759	2.092163e-39
50	11.768580	0.799089	1.801190	0.811081	2.119634e-49
53	13.393870	1.251408	3.273314	0.931657	1.266878e-27
59	13.045309	0.863173	0.000000	0.691551	1.686600e-13
61	9.828976	0.939626	0.000000	0.864188	1.346881e-37
63	9.904244	0.701541	0.000000	0.908179	1.142956e-53
70	9.030380	0.792165	0.918037	0.924041	6.768797e-26
82	7.563128	1.052447	0.000000	0.977592	2.047455e-31
84	11.273849	0.906613	2.291778	0.741179	4.978748e-10
115	13.782368	1.065176	0.000000	0.878257	7.645000e-97
115	13.782368	1.065176	0.000000	0.878257	7.645000e-97
121	10.076457	1.356190	0.256458	0.992374	7.172498e-50
123	8.647519	1.047006	0.000000	0.972279	3.175114e-26
160	9.861500	0.826446	0.044827	0.907266	2.403986e-33

Table 9 Restaurants

districtid	Avg. distance	α	β	Pearson r	p-value
4	8.246086	1.694699e+00	14.005347	0.705706	3.513858e-04
17	12.973318	3.609896e+01	100.000000	-0.037036	9.306203e-01
33	12.393137	2.284407e-01	2.121636	0.620474	1.584134e-03
34	12.342009	1.000000e+02	46.690396	0.956793	1.102762e-06
42	14.321486	0.000000e+00	100.000000	0.521633	4.783672e-01
50	10.063284	4.376173e+01	100.000000	0.051864	7.932417e-01
53	11.459094	7.655047e+01	0.403219	1.000000	0.000000e+00
59	10.507480	1.200341e+00	11.290621	0.906397	1.909024e-03
61	9.491864	8.194770e+01	48.867638	0.957679	1.333397e-05
63	10.540970	3.694519e+01	100.000000	0.903863	1.347535e-04
70	11.275803	1.482044e+00	0.415506	0.996231	2.128370e-05
82	15.038343	6.727115e+01	97.742040	1.000000	0.000000e+00
84	6.991360	3.884061e+01	19.634595	0.997704	4.843408e-07
115	12.531356	3.360911e-08	0.000000	0.365079	3.992169e-02
121	6.372688	1.603767e+00	0.000000	0.980718	3.204930e-03
123	11.206271	3.784310e+01	30.296372	0.998778	3.147883e-02
160	9.775328	7.847438e+01	41.454500	0.970033	1.562465e-01

Drezner et al(2010)Drezner, Drezner, and Kalczynski. Drezner T, Drezner Z, Kalczynski P (2010) A cover-based competitive location model. Journal of the Operational Research Society 62(1):100–113

Drezner(1995). Drezner Z (1995) Facility Location: A Survey of Applications and Methods. Springer, New York

Eiselt et al(1993)Eiselt, Laporte, and Thisse. Eiselt HA, Laporte G, Thisse JF (1993) Competitive Location Models - A Framework and Bibliography. Transportation Science 27(1):44–54

Hasan et al(2012)Hasan, Schneider, Ukkusuri, and González. Hasan S, Schneider CM, Ukkusuri SV, González MC (2012) Spatiotemporal Patterns of Urban Human Mobility.

Table 10 OLS regression model between Huff model performance and indicators. gas station (5541)

Indicator	β coefficient	Confidence interval (95%)
Mobility diversity	-0.3949	$[-1.2822, 0.4924]$
Merchant diversity	0.0228	$[-1.0236, 1.0692]$
Merchant monopoly	0.2326	$[-0.5643, 1.0295]$
Gender diversity	-0.0024	$[-2.4916, 2.4868]$
Marital status diversity	0.1781	$[-2.0838, 2.4400]$
Education level diversity	-0.5003	$[-1.4600, 0.4593]$
Job status diversity	0.2611	$[-0.5382, 1.0604]$
Income inequality	0.0928	$[-0.9343, 1.1200]$

Table 11 OLS regression model between Huff model performance and indicators. Restaurants (5812)

Indicator	β coefficient	Confidence interval (95%)
Mobility diversity	0.2624	$[-0.4601, 0.9849]$
Merchant diversity	-0.2362	$[-1.0883, 0.6159]$
Merchant monopoly	-0.0775	$[-0.7264, 0.5715]$
Gender diversity	-0.6746	$[-2.7015, 1.3524]$
Marital status diversity	0.1780	$[-1.6638, 2.0199]$
Education level diversity	0.2504	$[-0.5310, 1.0319]$
Job status diversity	-0.1828	$[-0.8337, 0.4680]$
Income inequality	-0.4394	$[-1.2758, 0.3970]$

Journal of Statistical Physics 151(1-2):304–318

Hodgson(2007). Hodgson MJ (2007) A location—allocation model maximizing consumers' welfare. Regional Studies 15(6):493–506

Hotelling(1929). Hotelling H (1929) Stability in Competition. The economic journal 39(153):41

Huff and McCallum(2008). Huff D, McCallum BM (2008) Calibrating the Huff Model Using ArcGIS Business Analyst. ESRI White Paper

Huff(1964). Huff DL (1964) Defining and Estimating a Trading Area. Journal of Marketing 28(3):34

Huff(1996). Huff DL (1996) A Programmed Solution for Approximating an Optimum Retail Location. Land Economics 42(3):293–303

Isaacman et al(2011)Isaacman, Becker, Cáceres, Kobourov, Martonosi, Rowland, and Varshavsky. Isaacman S, Becker R, Cáceres R, Kobourov S, Martonosi M, Rowland J, Varshavsky A (2011) Identifying Important Places in People's Lives from Cellular Network Data. In: Proceedings of the International Conference on Pervasive computing, pp 133–151

Jensen(2001). Jensen MC (2001) Value Maximization, Stakeholder Theory, and the Corporate Objective Function. Journal of applied corporate finance 14(3):8–21

Kenndy and Eberhart(1995). Kenndy J, Eberhart RC (1995) Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks

Leonardi and Tadei(1984). Leonardi G, Tadei R (1984) Random utility demand models and service location. Regional Science and Urban Economics 14(3):399–431

Nakanishi and Cooper(1974). Nakanishi M, Cooper LG (1974) Parameter estimation for a multiplicative competitive interaction model: least squares approach. Journal of marketing research 11(3):303

Plastria(2001). Plastria F (2001) Static competitive facility location - An overview of optimisation approaches. European Journal of Operational Research

Reilly(1931). Reilly WJ (1931) The law of retail gravitation. Knickerbocker Press, New York

Singh et al(2015)Singh, Bozkaya, and Pentland. Singh VK, Bozkaya B, Pentland A (2015) Money Walks: Implicit Mobility Behavior and Financial Well-Being. PLoS ONE 10(8):e0136628–17

Wernerfelt(1986). Wernerfelt B (1986) The Relation Between Market Share And Profitability. Journal of Business strategy 6(4):67–74

Zheng et al(2013)Zheng, Liu, and Hsieh. Zheng Y, Liu F, Hsieh HP (2013) U-Air: When Urban Air Quality Inference Meets Big Data. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp 1436–1444