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

Big Data
Volume 9, Number 3, 2021

DOI: 10.1089/big.2020.0161

ORIGINAL ARTICLE

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

Yoshihiko Suhara,[1] Mohsen Bahrami,[1,2,*] Burcin Bozkaya,[2,3] and Alex 'Sandy' Pentland[1]

Abstract

Customer patronage behavior has been widely studied in market share modeling contexts, which is an essential step toward estimating retail sales and finding new store locations in a competitive setting. 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 article, we propose a novel approach for validating the Huff gravity market share model, using a large-scale transactional dataset that describes customer patronage behavior at a regional level. Although the Huff model has been well studied and widely used in the context of sales estimation, competitive facility location, and demand allocation, this article 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 retail sales estimation and competitive facility location models.

Keywords: market share; Huff model; customer patronage behavior; Big Data analysis; behavioral analytics

Introduction

During the 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 United Kingdom 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 accurately estimate retail sales by location and then "optimally" locate new 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 what type of a new store to add to their chain. To address this challenge, a vast literature on retail sales estimation and facility location models has emerged. Many of these models focus on retail stores operating in a competitive environment.[2–9] An overall aim of such models is to understand how consumers are attracted to store locations, maximize the market share captured as a result of the new location(s) opened, and, consequently, to maximize the profitability of the company for its shareholders.[10,11] Hence, decision makers must understand and model the underlying processes for

[1]*The Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.*
[2]*Sabanci Business School, Sabanci University, Istanbul, Turkey.*
[3]*New College of Florida, Sarasota, Florida, USA.*

*Address correspondence to: *Mohsen Bahrami, The Media Lab, Massachusetts Institute of Technology, 20 Ames Street, Cambridge, Massachusetts, USA, Turkey,* E-mail: bahrami@mit.edu

retail patronage before sales estimation and facility location models can be solved effectively.

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.[12] 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 Drezner's 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,[13–15] credit card transactions,[16,17] GPS traces,[18] 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,[19] and in business/retail applications[20]; however, 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 article 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 contributions and advantages of our approach include the following:

- Retail sales estimation and 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, leads 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 article is organized as follows. Background section provides background information, including a literature review and a detailed description of the Huff model. In the Methodology section, we present our validation methodology, followed by the experimental results and discussion in the Results and Discussion section. Finally, we provide concluding remarks and directions for future research.

Background

Literature review

Among various models of competitive facility location that are developed and available in the literature,[4,5,7,21] 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 facility's market share. These models assume that the patronage behavior is influenced by multiple factors such as the retail facility's attractiveness to customers, distance from customers' location, and customers' purchasing power.[12] Among various market share estimation approaches proposed, five main ones include proximity,[22] deterministic utility,[23] random utility,[24,25] cover-based,[26] and gravity-based[27] 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.[22] 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,[23] which 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. However, 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.[24,25] 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."[26]

The fifth and the most extensively used approach is the gravity-based approach.[9] Estimating market share based on this approach was first introduced by Reilly[28] and further extended by Huff.[27,29] The Huff model approximates the probability of a customer's 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[29] further propose an improved Huff model by developing a multiplicative competitive interaction 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.[30–33]

Since the original Huff model,[27] which used the facility square footage as the attractiveness metric, other metrics have been introduced in the literature. For example, by conducting a survey with shopping mall customers, Drezner[34] identified the variety of stores, the mall appearance, and brand names as the three most important attractiveness measures. Other examples of attractiveness measures in the literature are the availability/size of parking area,[30,35,36] proximity to other stores and/or attractions,[30,36] buying power and price levels,[19,30,35–37] and product variety.[30,35,36] Also, various types of distance decay functions (e.g., logistic, exponential, and hybrid) have been proposed for the Huff model.[38–42] Sevtsuk and Kalvo[43] used survey data, including 1088 households providing information about their shopping trips, and developed a variant of the Huff model that uses street network-based distance. Their model uses the exponential decaying function for better modeling the shopping behavior in the urban area. Recently, Busu et al.[44] have conducted regression analysis and have shown that the current assets, the fixed assets, and the number of employees are significant predictors of the net profit of a store by using the financial data of 68 stores. This indicates that such financial attributes (i.e., current asset, fixed asset, the number of employees) can be good options to be considered as attractiveness measures.

Another line of work uses machine-learning techniques for location prediction[45–49] and site selection.[50–52] The techniques can be categorized by the types of features used for prediction. Commonly used features are cell-phone call records,[48,49] GPS trajectories,[45] check-in logs of the location-based social network,[46,50–52] or a combination of GPS trajectories and check-in information.[47] The main aim of these studies is to leverage historical data with machine-learning techniques to accurately predict the user location or future market potential. Some recent work has used machine learning techniques to predict future market potential of various locations for site selection. Ouyang et al.[52] use location-based social network check-in data and historical sales data to predict the market demand by using a neural network model. Wang et al.[51] use a modified version of the Huff model to calculate the spatial accessibility of digital signage using multiple machine-learning methods for site selection. Those techniques are designed for predicting future demand based on past histories. In contrast, the main focus of this article is modeling customers' patronage behaviors for market share estimation and validating the goodness of fit of the Huff gravity model using credit card transaction data.

The Huff model

The Huff model[27] is an economic model for estimating market share in relation to customer retail patronization decisions. This model is based on gravity models,[28,53] which describe the magnitude of interaction by two factors, namely mass and distance. The Huff model uses merchant *attractiveness* for the mass factor (the square footage of a merchant facility is used in the original Huff model) 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:

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

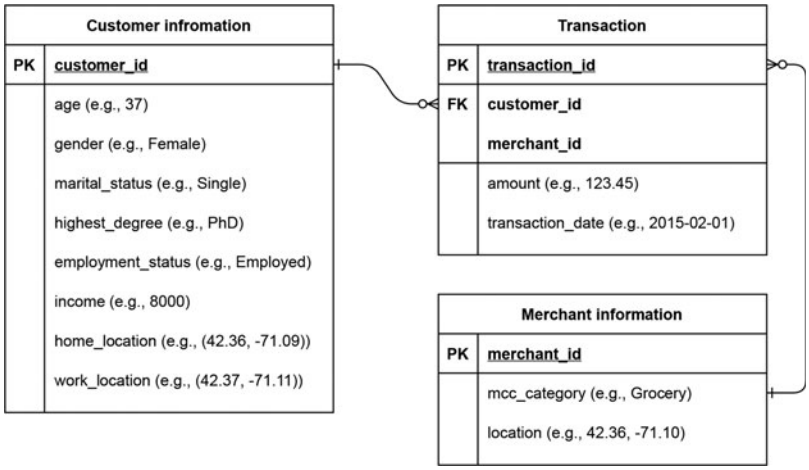

FIG. 1. Schematic diagram and example values of three tables in the credit card transaction dataset used in the article. FK, foreign key; PK, primary key.

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. The two parameters help the model better fit a target region. For instance, urban and rural regions should have different land values and transportation facilities. Thus, the square footage of a merchant and the distance between the merchant and a customer in a region are not always directly comparable with those of a merchant in a different region. The α and β parameters need adjustment to account for the impacts of the attractiveness and distance factors for each region. We will describe how to optimize the parameters in the Experimental Setting section. To obtain the probability of customer i visiting merchant j, H_{ij} is normalized by the sum of all utility values for possible visits:

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

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

Methodology

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 Organisation for Economic Co-operation and Development country between July 2014 and June 2015. The dataset has three tables: customer information table, merchant information table, and transaction table. Figure 1 shows a schematic diagram of the three tables, and the statistics of the dataset are shown in Table 1.

Customer information table. This table provides demographic information of anonymized customers. The demographic information includes age, gender, marital status, education level, employment status, estimated income by the bank, and their home and work locations.

Merchant information table. This table contains merchant information, including merchant category code and locations (geo coordinates).

Transaction table. This table includes all credit card transactions made by the customers in the customer table. Each transaction record contains customer ID and merchant ID, which can be linked to the customer and merchant information stored in the customer and merchant information tables. The table also includes transaction amount and transaction date for each

Table 1. Dataset summary

Period	July 1, 2014 to June 30, 2015
No. of transactions	4,254,652
No. of customers	62,392
Avg. transactions/customer	68.19

record. We also separately calculate and store the total merchant revenue for each merchant, which is the aggregated transaction amount of all customers who transacted at that merchant.

Data preprocessing

To retain the robustness of the experiment, we filtered the customers who have at least 10 transactions in the dataset. We also filtered the merchants based on their business categories. 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.

Model

We now describe how we use credit card transaction data for the Huff model. Specifically, we explain how we calculate the attractiveness A_j and estimate the α and β parameters.

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). Specifically, we approximated the revenue of a merchant with the total purchase amount of transactions made by all customers at that merchant. 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 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.

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

Merchant category	*No. of transactions*
Grocery store (Grocery)	1,089,614
GS	482,178
Clothing store (Clothing)	437,760
Restaurants	185,595

GS, Gas station.

Parameter estimation. Parameters α and β are optimized to maximize the evaluation metric through the particle swarm optimization (PSO) technique[54] within the range of α, β in [0, 100]. The PSO does not require any derivative information for optimization, and thus it is commonly used for model selection in machine learning.[55] A recent study[56] has applied it to a facility location problem. We also tested the ordinary least-squares method,[57] which is a common choice for the Huff model parameter estimation. However, the method significantly under-performed compared with the PSO method. Therefore, we used PSO for parameter estimation in this experiment. To the best of our knowledge, this article 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.

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 $\times$ 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.

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. Then, we aggregate $P_{ij}^{(r,c)}$ for all customers to obtain the estimated market share S_i for each merchant i. From the transaction data, we calculate the actual market share of each merchant based on the number of transactions:

$$S_i = \frac{N_i}{N} \tag{3}$$

where N_i and N denote the number of transactions made at merchant i and all merchants, respectively.

We calculate the Pearson's correlation value between the estimated market shares $\hat{S}$ and the actual market shares S for each district.

Regression analysis

We conduct regression analysis to find significant indicators of the performance of our models. Specifically, we use linear regression and consider the model performance (i.e., Pearson's correlation values) as dependent variables and the following indicators as independent variables.

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 to get a reliable analysis. This is mainly because we are not sure whether those customers who reported their income zero did not report their income or did not really have any income (e.g., homemaker). Thus, we decided to exclude such information for the income inequality calculation.

We consider the indicators described earlier 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 × 4 merchant categories) samples for the regression analysis. All the variables were standardized by converting into z-scores for easier interpretation.

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

Results and Discussion

Model performance

Table 3 summarizes the basic statistics of the results, and Figure 2 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 Tables A1–A6.

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

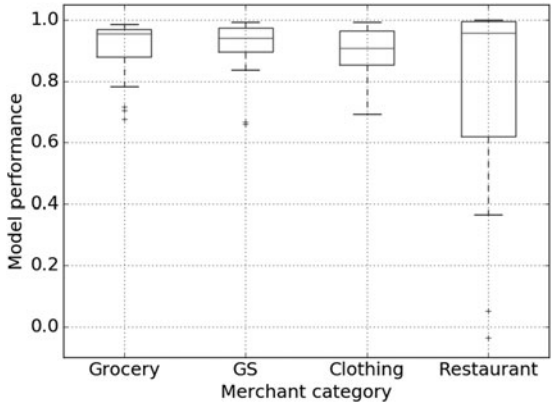

FIG. 2. Distribution of model performance (Pearson's correlation) values for each merchant category. GS, gas stations.

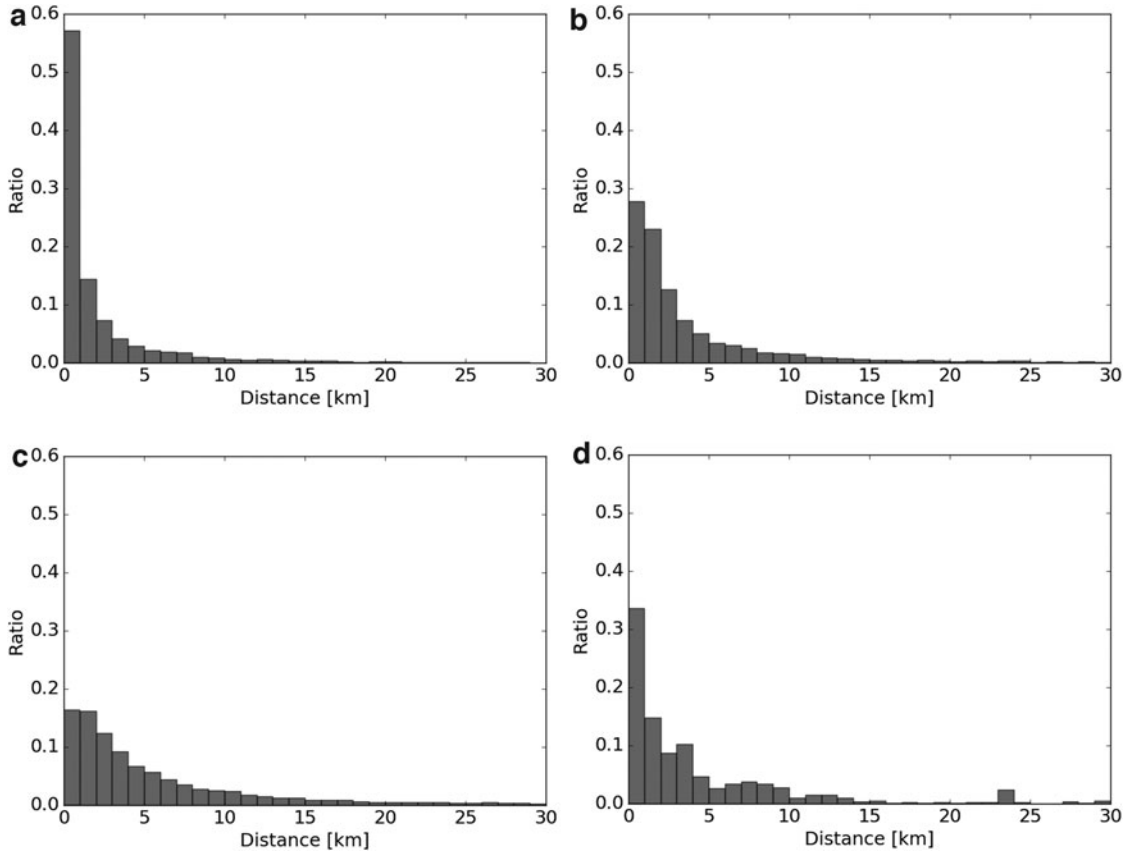

FIG. 3. Distribution of the distance between visited merchants and customer's home/work locations (closer location is taken) of each merchant category. **(a)** Grocery, **(b)** GS, **(c)** Clothing stores, and **(d)** Restaurants.

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 with 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

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

Table 5. Ordinary least-squares regression model between Huff model performance (i.e., the Pearson's correlation values) and indicators

Indicator	β *coefficient*	*Confidence interval (95%)*
(1) Grocery		
Mobility diversity	−0.1799	−0.6516 to 0.2917
Merchant diversity	−0.2038	−0.7601 to 0.3524
Merchant monopoly	0.0586	−0.3650 to 0.4822
Gender diversity	**2.5007****	1.1776 to 3.8239
Marital status diversity	**−2.4411****	−3.6434 to −1.2388
Education level diversity	−0.3585	−0.8686 to 0.1516
Job status diversity	**0.5106***	0.0858 to 0.9355
Income inequality	0.2643	
(2) Clothing		
Mobility diversity	0.1355	−0.3742 to 0.6453
Merchant diversity	0.5748	−0.0263 to 1.1760
Merchant monopoly	−0.4263	−0.8842 to 0.0315
Gender diversity	**1.4881***	0.0581 to 2.9181
Marital status diversity	**−1.3081***	−2.6076 to −0.0087
Education level diversity	**−0.8321****	−1.3834 to −0.2808
Job status diversity	0.2080	−0.2511 to 0.6672
Income inequality	−0.3274	−0.9174 to 0.2627

*, ** denote $p < 0.05$, 0.01 respectively. Bold face denotes that the β coefficient is statistically significant.

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 3. As shown in Figure 3a–c, the visited merchants of the Grocery, GS and Clothing categories are basically located close to the customers' home/work locations whereas the distance distribution of restaurants contains long distance values as shown in Figure 3d. 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

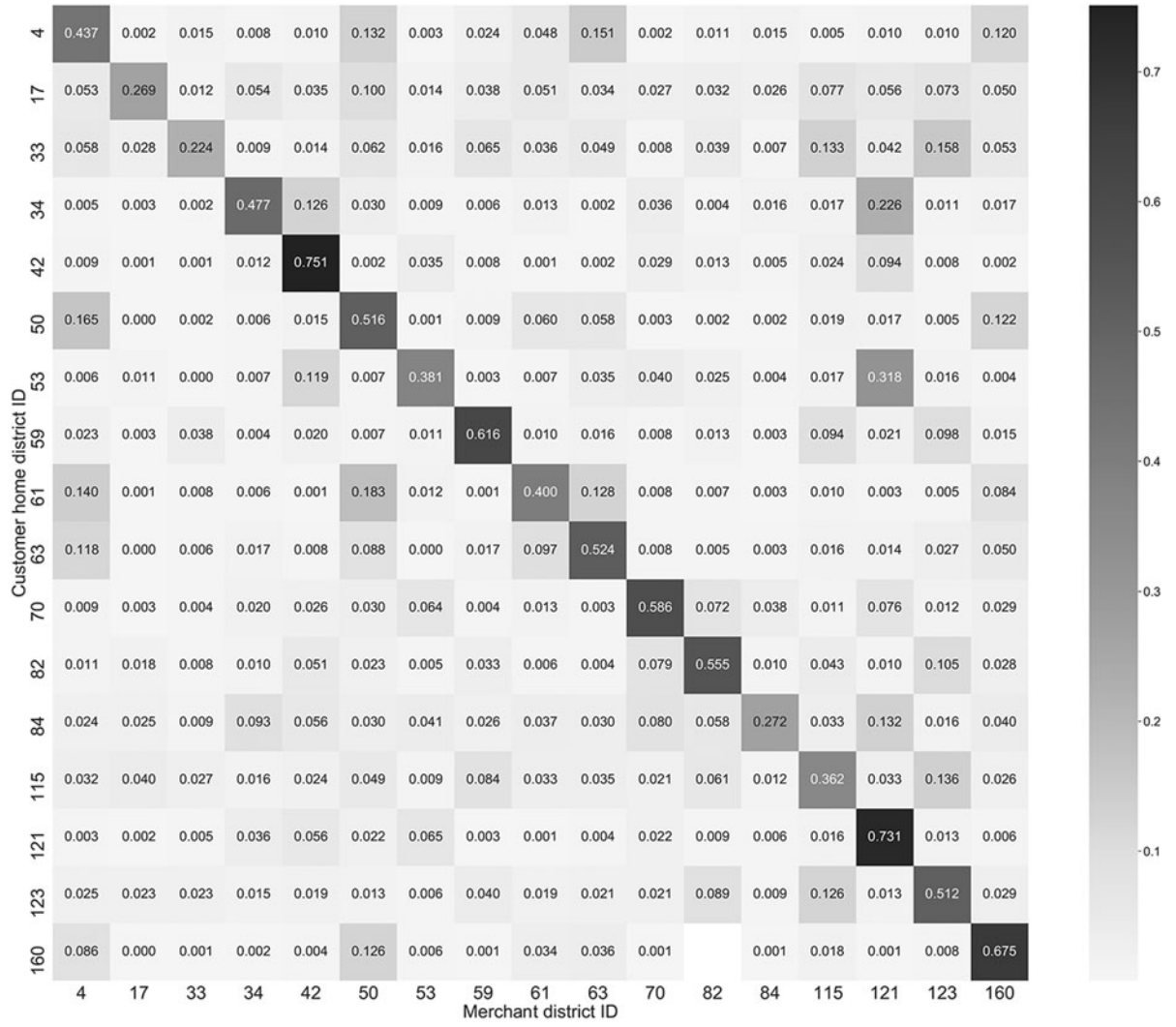

FIG. 4. Mobility patterns of customers for the Grocery category. It is more likely that customers visit grocery stores in the same district as their home locations.

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 (GS) is a great example of the Huff model being used on transaction data, among the four categories considered. We observe that the most frequented GS 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 GS in the vicinity of their home/work locations.

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. Further, the regression models do not show any statistical significance in the β coefficient values of the diversity indicators for these two categories.

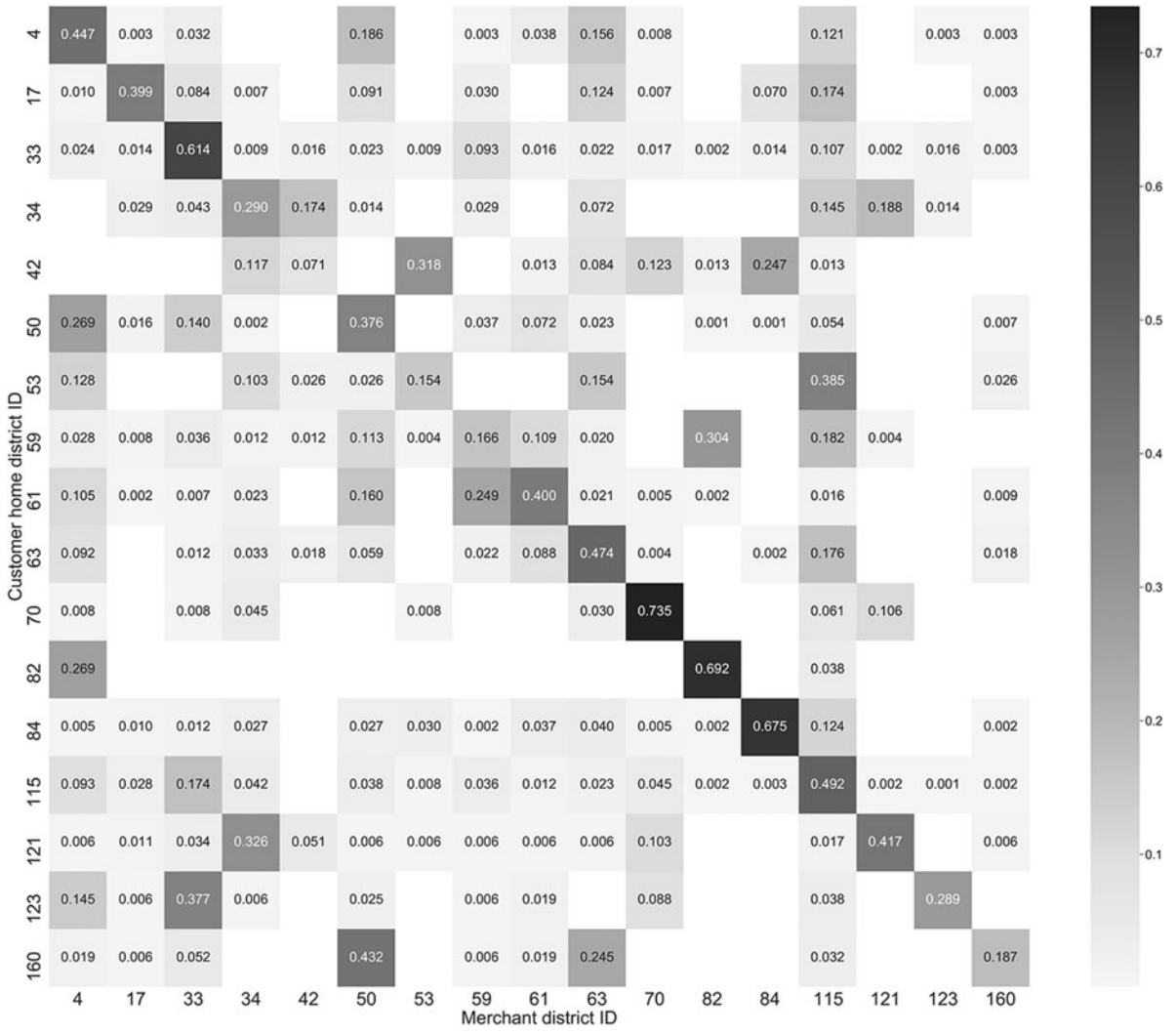

FIG. 5. Mobility patterns of customers for the Restaurant category. Customers do not necessarily visit restaurants in the same districts as their home locations, showing a different trend from the grocery category.

Table 5 shows the β coefficient output of the regression models for (1) Grocery and (2) 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 Tables A1–A6 since 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 whereas 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, such as mobility, merchant, and income diversity. Originally, we hypothesized that the mobility

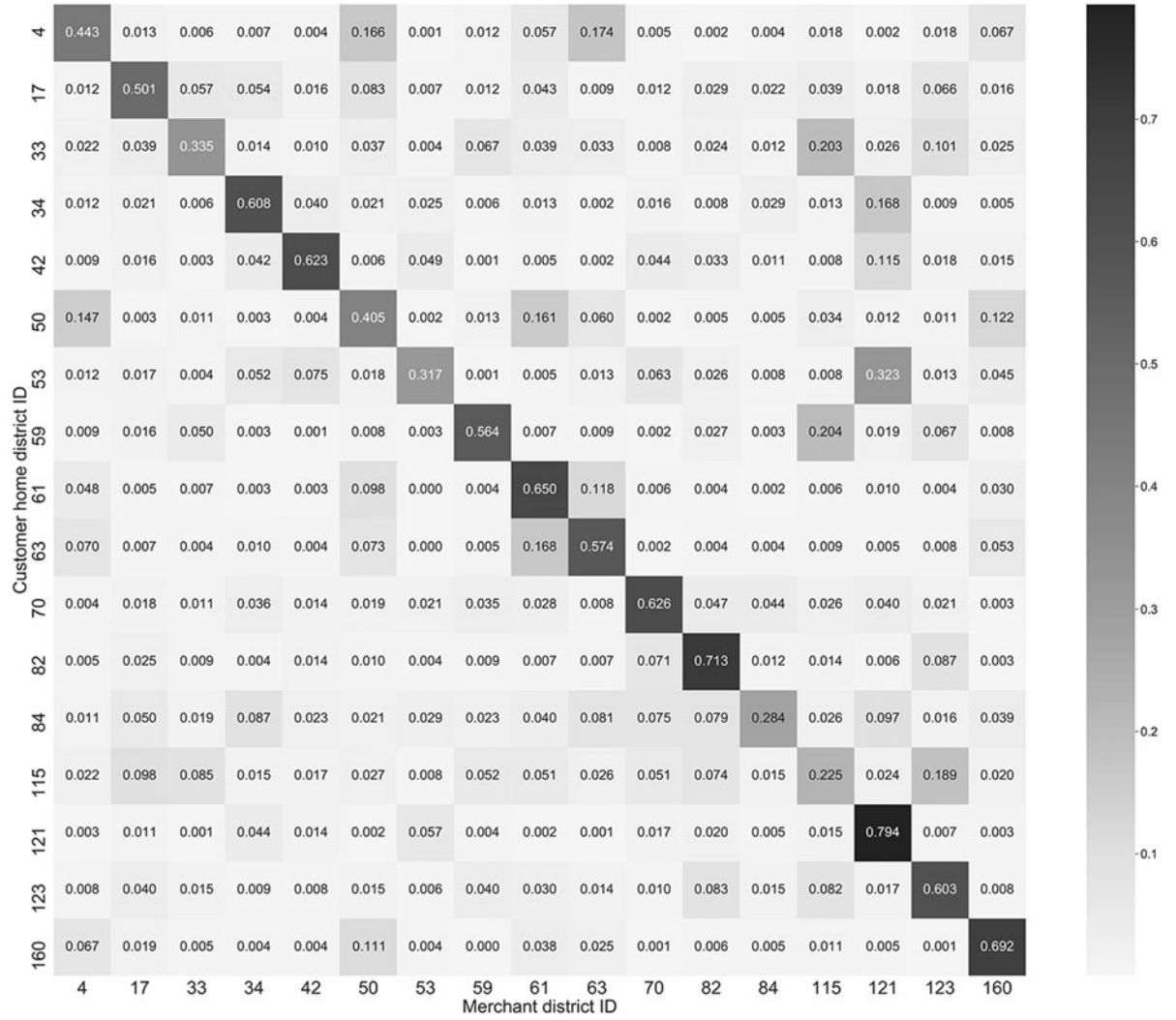

FIG. 6. Mobility patterns of customers for the GS category.

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.

We have also conducted another type of mobility analysis to understand the differences regarding customers' shopping behavior for each merchant category. Figures 4, 5, 6 and 7 show 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, Figure 5 shows that merchants in the Restaurant category have more biased distributed with respect to the mobility patterns.

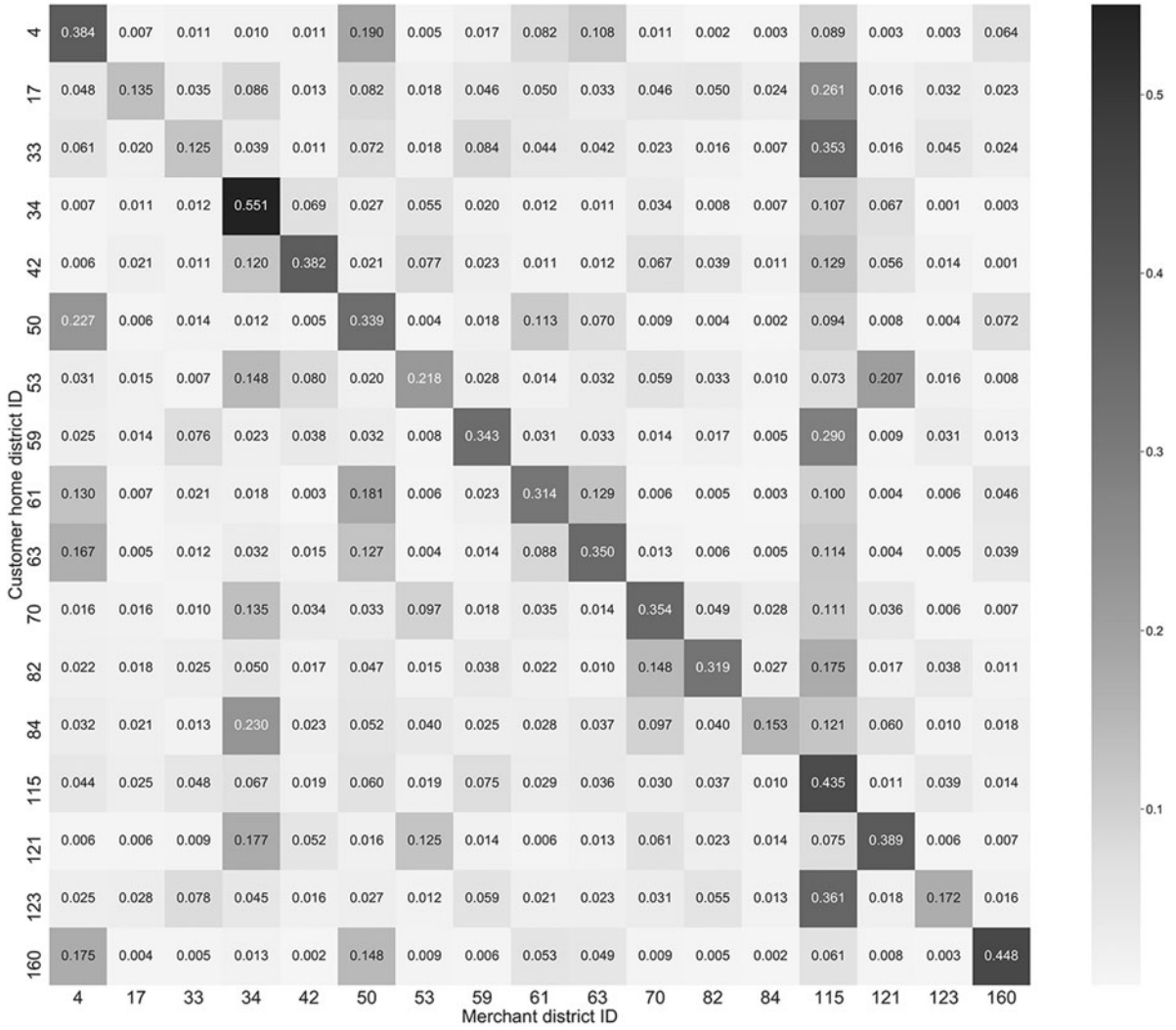

FIG. 7. Mobility patterns of customers for the Clothing stores category.

Summary and Conclusion

In this article, 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 to estimate market shares in competitive facility location problems, our study provides key insights for reliable use of Huff or other gravity-based models in competitive facility location problems. Compared with the conventional survey-based approaches, the major advantages of our transaction-based approach include: (1) no requirement for surveys where data collection cost and data quality might be an issue, (2) the ability to directly compare different categories of shopping, and (3) 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 by using a gravity-based approach. However, we believe that our approach provides various benefits that 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.

Authors' Contributions

Y.S. and M.B. were involved in idea generation, design and implementation of experiments, drafting the article, and writing. B.B. and A.S.P. were involved in idea generation, results, discussion, feedback, and final revision and drafting of the article.

Acknowledgments

The authors are grateful to the financial institution that provided the credit card transaction data for this research.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. UK Small Business Statistics. Available online at https://www.fsb.org.uk/uk-small-business-statistics.html (last accessed December 30, 2020).
2. Ghosh A. Parameter nonstationarity in retail choice models. J Bus Res. 1984;12:425–436.
3. Geisel MS, Narasimhan C, Sen SK. Quantifying the competitive impact of a new entrant. J Bus Res. 1993;26:263–277.
4. Drezner Z. Facility location: A survey of applications and methods. New York: Springer, 1995.
5. Plastria F. Static competitive facility location—An overview of optimisation approaches. Eur J Oper Res. 2001;129:461–470.
6. Gonzalez-Benito O. Spatial competitive interaction of retail store formats: Modeling proposal and empirical results. J Bus Res. 2005;58:457–466.
7. Berman O, Drezner T, Drezner Z, et al. (Eds). Modeling competitive facility location problems: New approaches and results. In: Decision Technologies and Applications. Catonsville, MD: INFORMS, 2009, pp. 156–181.
8. Merino M, Ramirez-Nafarrate A. Estimation of retail sales under competitive location in Mexico. J Bus Res. 2016;69:445–451.
9. Drezner T. Gravity models in competitive facility location. In: Eiselt HA, Marianov V (Eds). Contributions to Location Analysis. Cham, Switzerland: Springer, 2019, pp. 253–275.
10. Jensen MC. Value maximization, stakeholder theory, and the corporate objective function. J Appl Corp Finance. 2001;14:8–21.
11. Wernerfelt B. The relation between market share and profitability. J Bus Strategy. 1986;6:67–74.
12. Drezner T. Derived attractiveness of shopping malls. IMA J Manag Math. 2006;17:349–358.
13. Isaacman S, Becker R, Caceres R, et al. Identifying important places in people's lives from cellular network data. In: Proceedings of the International Conference on Pervasive computing. Springer, 2011, pp. 133–151.
14. Blondel VD, Esch M, Chan C, et al. Data for development: The D4D challenge on mobile phone data. arXiv preprint arXiv:1210.0137, 2012.
15. Bogomolov A, Lepri B, Staiano J, et al. Once upon a crime: Towards crime prediction from demographics and mobile data. In: Proceedings of the 16th International Conference on Multimodal Interaction. ACM, 2014.
16. Singh VK, Bozkaya B, Pentland A. Money walks: Implicit mobility behavior and financial well-being. PLoS One. 2015;10:e0136628-17.
17. Hasan S, Schneider CM, Ukkusuri SV, et al. Spatiotemporal patterns of urban human mobility. J Stat Phys. 2012;151:304–318.
18. Zheng Y, Liu F, Hsieh H-P. U-Air: When urban air quality inference meets big data. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data mining. ACM, 2013, pp. 1436–1444.
19. Drezner T. A review of competitive facility location in the plane. Logist Res. 2014;7:114–12.
20. Huff D, McCallum BM. Calibrating the Huff model using ArcGIS business analyst. ESRI White Paper, 2008;1–33.
21. Eiselt HA, Laporte G, Thisse J-F. Competitive location models—A framework and bibliography. Transp Sci. 1993;27:44–54.

22. Hotelling H. Stability in competition. Econ J. 1929;39:41.
23. Drezner T. Locating a single new facility among existing, unequally attractive facilities. J Reg Sci. 1994;34:237–252.
24. Leonardi G, Tadei R. Random utility demand models and service location. Reg Sci Urban Econ. 1984;14:399–431.
25. Drezner T, Drezner Z. Competitive facilities: Market share and location with random utility. J Reg Sci. 1996;36:1–15.
26. Drezner T, Drezner Z, Kalczynski P. A cover-based competitive location model. J Oper Res Soc. 2010;62:100–113.
27. Huff DL. Defining and estimating a trading area. J Mark. 1964;28:34.
28. Reilly WJ. The law of retail gravitation. New York: Knickerbocker Press, 1931.
29. Huff DL. A programmed solution for approximating an optimum retail location. Land Econ. 1996;42:293–303.
30. Bozkaya B, Yanik S, Balcisoy S. A GIS-based optimization framework for competitive multi-facility location-routing problem. Netw Spat Econ. 2010;10:297–320.
31. Berman O, Krass D. Locating multiple competitive facilities—Spatial interaction models with variable expenditures. Ann Oper Res. 2002;111: 197–225.
32. Aboolian R, Berman O, Krass D. Competitive facility location and design problem. Eur J Oper Res. 2007;182:40–62.
33. Hodgson MJ. A location—Allocation model maximizing consumers' welfare. Reg Stud. 2007;15:493–506.
34. Drezner T. A procedure for estimating the attractiveness of shopping malls. In: Proceedings of 29th Annual DSI Meeting, Las Vegas, NV, November 1998, pp. 1090–1092.
35. Gautsch DA. Specification of patronage models for retail center choice. J Mark Res. 1981;18:162–174.
36. Singla V, Rai H. Investigating the effects of retail agglomeration choice behavior on store attractiveness. J Mark Anal. 2016;4:108–124.
37. Suarez-Vega R, Gutierrez-Acuna JL, Rodrıguez-Dıaz M. Locating a supermarket using a locally calibrated huff model. Int J Geogr Inf Sci. 2015;29: 217–233.
38. Wilson AG. Retailers' profits and consumers' welfare in a spatial interaction shopping model. In: Theory and Practice in Regional Science. London: Pion, 1976, pp. 42–59.
39. Hodgson MJ. A hierarchical location-allocation model for primary health care delivery in a developing area. Soc Sci Med. 1988;26:153–161.
40. Handy SL, Niemeier DA. Measuring accessibility: An exploration of issues and alternatives. Environ Plan A. 1997;29:1175–1194.
41. Wang F. Measurement, optimization, and impact of health care accessibility: A methodological review. Ann Assoc Am Geogr. 2012;102:1104–1112.
42. Delamater PL, Messina JP, Grady SC, et al. Do more hospital beds lead to higher hospitalization rates? A spatial examination of Roemer's law. PLoS One. 2013;8:e54900.
43. Sevtsuk A, Kalvo R. Patronage of urban commercial clusters: A network-based extension of the huff model for balancing location and size. Environ Plan B Urban Anal City Sci. 2018;45:508–528.
44. Busu M, Vargas MV, Gherasim IA. An analysis of the economic performances of the retail companies in Romania. Manage Mark Chall Knowl Soc. 2020;15:125–133.
45. Ashbrook D, Starner T. Using GPS to learn significant locations and predict movement across multiple users. Pers Ubiquitous Comput. 2003;7: 275–286.
46. Shaw B, Shea J, Sinha S, et al. Learning to rank for spatiotemporal search. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM'13, New York, NY: Association for Computing Machinery, 2013, pp. 717–726.
47. Suzuki J, Suhara Y, Toda H, et al. Personalized visited-POI assignment to individual raw GPS trajectories. ACM Trans Spat Algorithms Syst. 2019; 5.3:1–28.
48. Zhang D, Zhang D, Xiong H, et al. NextCell: Predicting location using social interplay from cell phone traces. IEEE Trans Comput. 2015;64: 452–463.
49. Leng Y, Koutsopoulos HN, Zhao J. Profiling presence patterns and segmenting user locations from cell phone data. CoRR, abs/1805.12208, 2018.
50. Wang L, Fan H, Wang Y. Site selection of retail shops based on spatial accessibility and hybrid BP neural network. ISPRS Int J Geoinf. 2018;7:202.
51. Wang Y, Li S, Zhang X, et al. Site selection of digital signage in Beijing: A combination of machine learning and an empirical approach. ISPRS Int J Geoinf. 2020;9:217.
52. Ouyang J, Fan H, Wang L, et al. Site selection improvement of retailers based on spatial competition strategy and a double-channel convolutional neural network. ISPRS Int J Geoinf. 2020;9:357.
53. Anderson JE. The gravity model. Natl Bur Econ Res. 2010;93:170–192.
54. Kenndy J, Eberhart RC. Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks. IEEE, 1995, pp. 1942–1948.
55. Escalante HJ, Montes M, Sucar LE. Particle swarm model selection. J Mach Learn Res. 2009;10:405–440.
56. Ma H, Guan X, Wang L. A single-facility competitive location problem in the plane based on customer choice rules. J Data Inf Manag. 2020;2: 323–336.
57. Nakanishi M, Cooper LG. Parameter estimation for a multiplicative competitive interaction model: Least squares approach. J Mark Res. 1974;11:303.

Cite this article as: Suhara Y, Bahrami M, Bozkaya B, Pentland AS (2021) Validating gravity-based market share models using large-scale transactional data. *Big Data* 9:3, 188–202, DOI: 10.1089/big.2020.0161.

Abbreviations Used

FK = foreign key
GS = gas stations
OLS = ordinary least squares
PK = primary key
PSO = particle swarm optimization

(*Appendix follows* →)

Appendix

Appendix Table A1. Grocery stores

District ID	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

Appendix Table A2. Gas stations

District ID	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

Appendix Table A3. Clothing stores

District ID	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

(*Appendix continues* →)

Appendix Table A4. Restaurants

District ID	*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

Appendix Table A5. Ordinary least-squares regression model between Huff model performance and indicators

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

Gas station (5541).

Appendix Table A6. Ordinary least-squares regression model between Huff model performance and indicators

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

Restaurants (5812).