Using Online Geotagged and Crowdsourced Data to Understand Human Offline Behavior in the City: An Economic Perspective

YINGJIE ZHANG, BEIBEI LI, and JASON HONG, Carnegie Mellon University

32

The pervasiveness of mobile technologies today has facilitated the creation of massive online crowdsourced and geotagged data from individual users at different locations in a city. Such ubiquitous user-generated data allow us to study the social and behavioral trajectories of individuals across both digital and physical environments. This information, combined with traditional economic and behavioral indicators in the city (e.g., store purchases, restaurant visits, parking), can help us better understand human behavior and interactions with cities. In this study, we take an economic perspective and focus on understanding human economic behavior in the city by examining the performance of local businesses based on the values learned from crowsourced and geotagged data. Specifically, we extract multiple traffic and human mobility features from publicly available data source geomapping and geo-social-tagging techniques and examine the effects of both static and dynamic features on booking volume of local restaurants. Our study is instantiated on a unique dataset of restaurant bookings from *OpenTable* for 3,187 restaurants in New York City from November 2013 to March 2014. Our results suggest that foot traffic can increase local popularity and business performance, while mobility and traffic from automobiles may hurt local businesses, especially the well-established chains and high-end restaurants. We also find that, on average, one or more street closure (caused by events or construction projects) nearby leads to a 4.7% decrease in the probability of a restaurant being fully booked during the dinner peak. Our study demonstrates the potential to best make use of the large volumes and diverse sources of crowdsourced and geotagged user-generated data to create matrices to predict local economic demand in a manner that is fast, cheap, accurate, and meaningful.

CCS Concepts: • **Information systems** → *Spatial-temporal systems*; *Location based services*; *Users and interactive retrieval*;

Additional Key Words and Phrases: Geotagged social media, crowdsourced user behavior, econometrics, location-based service, econometric analysis, city demand, mobility analytic

ACM Reference format:
Yingjie Zhang, Beibei Li, and Jason Hong. 2017. Using Online Geotagged and Crowdsourced Data to Understand Human Offline Behavior in the City: An Economic Perspective. *ACM Trans. Intell. Syst. Technol.* 9, 3, Article 32 (December 2017), 24 pages.
https://doi.org/10.1145/3078851

Authors' addresses: Y. Zhang and B. Li, H. John Heinz III College, Carnegie Mellon University, Pittsburgh, PA, USA; emails: yingjie@cmu.edu, beibeili@andrew.cmu.edu; J. Hong, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA; email: jasonh@cs.cmu.edu.

https://doi.org/10.1145/3078851

1 INTRODUCTION

Rapid urbanization is imposing various challenges on urban environments, in particular increasing demand on city infrastructures and on the quality of services. These challenges call for a specific focus on urban systems and their interactions with humans and businesses. In particular, properties of a city, such as transportation, street facilities, and neighborhood walkability, and their impacts on human behavior are at the core of sustainability and local economy. For example, when major streets in Boston were locked down during the Marathon Bombing in April 2013, the estimated costs to local businesses ranged from $250 to $333 million a day ([7]). A decrease in foot traffic can have significantly negative impact on store sales (e.g., [39]). These kinds of economic losses can lead to a negative effect on the local economy and can impose a long-term effect on the future sustainability of an urban neighborhood and quality of life. Therefore, understanding the patterns of human behavior in the city, especially how humans respond to city infrastructures and services (i.e., street closures, traffic conditions, etc.) from an economic perspective is critical in helping policy makers proactively improve city planning for better social welfare.

One major challenge here is in quantifying and measuring the quality of city infrastructures and services (i.e., street closures, traffic conditions, etc.) because it includes many factors, such as user walkability in an urban area, street connectivity (e.g., temporary closure of street facilities for events or construction), transportation and traffic conditions, and other urban amenities. These multidimensional characteristics make it very difficult to quantify and measure the service quality in an urban system. Furthermore, it reflects a combination of not only the static spatial and social elements in an urban environment, but also the dynamic characteristics of an urban system (e.g., traffic, events, and human mobility). This dynamic nature makes it highly unpredictable with regard to its economic impact on human behaviors. Recently, the pervasiveness of mobile technologies has facilitated the creation of massive online crowdsourced and geotagged data from individual users in real time and at different locations in the city. Such ubiquitous user-generated data allow us to study the social and behavioral trajectories of individuals across both digital and physical environments. This information, combined with traditional human economic and behavioral indicators in the city (e.g., store purchases, restaurant visits, parking), can help us better understand human behavior and interactions with the city, as well as improve the quality of life of human beings. In this research, we extract multilevel features of city infrastructures and services by applying geomapping and geo-social-tagging techniques on large-scale publicly available data from Twitter and Foursquare. In particular, using geotagged user-generated data created via mobile and location-based services and crowdsourcing channels, we are able to extract fine-grained information on various real-time traffic conditions, street events, and human movements that would otherwise be impossible to measure.

Another major challenge in this research lies in measuring the economic impacts of city infrastructures and services on human behavior. Previous studies have shown the advantages of using such ubiquitous user-generated data created through mobile and crowdsourced channels to explore various patterns of human behavior ([9, 12, 35, 41]). However, little work has been done to examine from a social and economic perspective how such data can be used to study human behavior in the city to infer relationship between humans and cities. Fu, Ge et al. and Fu, Xiong et al. [19, 20] are two studies that explore the values of individual check-ins, smart card transactions, and other mobility features. But they focused on the rankings of residential real estates, while we are interested in the short-term dynamics of small business in a urban city. In particular, using methods devised from economics, we focus on understanding the economic behavior of users in the city by examining the economic value from such large-scale and fine-grained information extracted from geotagged and crowdsourced channels.

Combining spatial, traffic, and human mobility analytics with econometric analyses, our major research goals are twofold:

- Extract both spatial and socioeconomic features of cities from online geotagged and crowdsourced data at large scale;
- Apply econometric models to quantify the causal effects of different features on the economic outcome of offline human behavior toward local businesses.

We instantiate our study in the context of local restaurants' booking performance by using a unique dataset of restaurant reservations from OpenTable, a major U.S. restaurant booking website. The dataset contains complete information from November 2013 to March 2014 for 3,187 restaurants in New York City. In addition, we use information on neighborhoods from four main sources across various social media channels and location-based services: (i) social and geographical information about local neighborhoods, (ii) street events and construction information collected from NYC's online map portal, (iii) human mobility information from approximately 380,000 Foursquare user mobile check-ins, and (iv) traffic-related information extracted from 18,900 individual geotagged tweets from Twitter.

Our final results show that features extracted from digitized and crowdsourced user behavior are informative in inferring local demand. Specifically, we find a significant positive impact of human foot traffic on local businesses and significant negative effects due to traffic, such as bus delays and disabled vehicles. In particular, a 10% increase in the density of human foot traffic increases the probability of a restaurant being fully booked during dinner peak hour by 4%, whereas a 10% increase in real-time transportation traffic density can decrease this probability by 5%. Moreover, we find that, on average, one or more street event or construction project nearby can decrease the probability of a restaurant being fully booked during the peak dinner hour by 4.7%. Our econometric methods alleviate the potential concerns of endogeneity from different factors in an urban system and support our findings from a causal perspective.

Our key contributions can be summarized as follows. (i) We propose a fast and effective way to leverage large-scale data from geotagged and crowdsourced social media to learn user economic behavior and local demand in the city. (ii) To the best of our knowledge, ours is the first study to conduct a causal analysis to quantify the economic impact of both static and dynamic features of users' digitized and crowdsourced behavior on small businesses in an urban setting. Our findings can help local businesses to understand the social and economic development of different urban areas and to improve marketing strategies by leveraging large-scale spatial, traffic, and human mobility analytics from social media. Our results can also help facilitate better policy decision-making about proactive city planning and improve the sustainability of urban neighborhoods. For example, our model can help urban planners conduct an ex ante analysis on the opportunity cost of a construction project before starting it. (iii) Our work also offers an opportunity for incorporating an economic lens into location-based services and geomapping services, which could help improve our understanding of local areas, as well as local search and local advertising.

The remainder of the article is organized as follows. We review some previous relevant studies in the Section 2. In Section 3 and Section 4, we discuss our data and our empirical econometric models. Section 5 discusses how we validate our empirical results. We furthermore test several robustness checks in Section 6. Finally, we conclude in Section 7 with a summary of potential policy implications and future directions.

2 RELATED WORK

Our study draws from and builds on the following streams of literature.

2.1 Geotagged and Crowdsourced Data Analysis

With the growing volume of geographic datasets, especially of geotagged datasets, more and more researchers are attracted by location-based services ([32, 46, 53, 55]). Previous studies used various methods to explore this emerging phenomenon from different perspectives, including usage patterns of location-sharing applications ([9, 35]), relationships between people ([12, 21, 31]), and detection of real-time events ([45, 50]). These studies put various methods forward to evaluate human mobility patterns. Noulas et al. [37] evaluated mobility features via selected historical visits, categorical preferences and social filtering; Karamshuk et al. [26] measured them by popularity, incoming flow, etc. However, most of those studies are exploratory analyses, answering what happens and how users behave in the real world. They didn't link their studies to economic values, and such further-step analysis can benefit economic development or even an entire society.

2.2 Consumer Social and Economic Behavior

Understanding consumers' social and economic behavior in the city is the main focus of researchers in marketing or economics-related fields ([11, 28, 48]). Due to the lack of data, prior literature tends to limit its focus on the online world. However, microeconomics, especially the performance of small businesses, is largely affected by various location-specific factors, such as neighborhood design, human mobility features, and location popularity. Merely relying on online sources (i.e., online word-of-mouth) makes it hard to gain a holistic picture to understand the urban economy at micro level. In this study, we utilize geotagged and crowdsourced data to study consumers' social and economic behavior in the city and to understand the associated impacts on local small businesses.

2.3 Economics of Location and Urban System

In addition, our study is also closely related to the economics of location and urban systems. This stream of research can be traced back to the 1970s ([29]). Different studies used various indicators to detect market price ([5, 42]), the best location ([16, 47]), and more. Zheng et al. [54] also summarizes potential applications in terms of urban computing for economy. However, the indicators they used to evaluate economic values were based on historical records or census data, such as demographics, crime rates, and climate records. One of the disadvantages is that such indicators cannot precisely capture the real-time performance of an urban system and its impacts. This can potentially present more implications for understanding the relationship between an urban system and the local economy. More recently, studies from information systems and urban economics looked at the interactions between new technology and local markets. For example, Forman et al. [18] found that the adoption of commercial Internet is more likely in rural areas than in urban areas. Forman et al. [17] and Langer et al. [30] focused on how the interaction of online and offline retailers affects consumer choice of channels. They found substitution effects between online and offline channels ([17]) and that channel usage is both heterogeneous and dynamic across buyers ([30]).

2.4 Causal Analysis on Panel Data

Estimating causal effects is a central goal in quantitative empirical research, especially with observational panel data. Literature has shown the effectiveness and applications of different econometric methods, including Propensity Score Matching ([4, 28, 38]), Instrument Variables ([3, 22]), Difference-in-Difference Analysis ([14, 48]), and the like. These methods can help us eliminate potential endogeneity issues when measuring causal effects, especially when data have some limitations. In this article, we applied several of the preceding methods in our econometric analysis, as well as in the robustness checks, to guarantee the findings on causality.

Fig. 1. Geographical distribution of restaurants in NYC.

3 DATA

Our dataset consists of observations of 3,187 Manhattan (NYC) restaurants from November 29, 2013 to March 6, 2014. The data were collected from multiple sources.

3.1 Data Source Description

3.1.1 Restaurant Reservation Data. We have approximately three months of restaurant reservation data from OpenTable from November 29, 2013 to March 6, 2014. This website offers an online network system to connect reservations between restaurants and consumers. Specifically, the website lists real-time reservation availability information given different requested time slots. Our dataset contains information about reservation availability for a party of two for six different time slots: 6 pm, 6:30 pm, 7 pm, 7:30 pm, 8 pm, and 8:30 pm (peak dining hours). In total, we have 312,326 data points. We visualize the geographical distribution of the restaurants in Figure 1.

3.1.2 Geotagged and Crowdsourced Data. Local demand is largely affected by social and economic factors in neighborhoods. To extract those factors, we collected crowdsourced and geotagged data based on three publicly available sources (the time window is the same as that in our restaurant reservation data):

(a) NYC street closure data. We collected street closure data from the official map portal (gis.nyc.gov/streetclosure/). Every day, it publishes information about street closures caused by street or intersection construction projects or special events in Manhattan. After removing duplicate projects, we obtained a total of 3,700 construction projects. Most of the projects, which were captured at a granular level, cover only one to two blocks. This information allowed us to pin down the effects of street closures on nearby restaurants.

(b) Foursquare check-ins data. We crawled Foursquare mobile check-ins publicly visible on Twitter. Previous research has shown the potential of approximating user footprints with mobile check-ins ([27, 32]). We have approximately 380,000 mobile user check-ins generated within a 30-mile radius from the center of Manhattan. We used geocoding tools to extract the geographical locations (i.e., latitude and longitude information) of the check-ins.

(c) Traffic-related tweets data. We extracted tweets related to traffic from Twitter using NLP and geocoding techniques. We conducted this step using two approaches. First, we considered the entire Twitter dataset over the three-month period and extracted traffic-related keywords. This approach has been widely used in recent work (see, for example, Hua et al. [25]). In addition, we identified and extracted information from influential users on Twitter who tweeted primarily about traffic. Specifically, we used all the tweets post by “511 NYC Area (@511NYC)”, whose information is provided by the New York State Department of Transportation. The tweets include different types of real-time traffic conditions, such as accidents, heavy traffic, special events, bus delays, and the like. We extracted 18,000 traffic tweets that cover our data period (i.e., 100 days). Again, we were able to extract the geo-coordinates associated with all these tweets to infer the exact location of each traffic incident.

To link all of the preceding datasets, we geotagged all data using the Google Map API. Because neither OpenTable data nor street closure data contain geographical coordinates, we first translated street addresses into geo-coordinates. Then, we computed the direct distance[1] between each of the pairs: restaurant and restaurant, restaurant and street closure, restaurant and check-ins, and restaurant and traffic tweets. Here, we consider "neighborhood" as a 0.5-mile-radius area, which we assume is a walk-able distance ([8, 43, 51]).

(d) Restaurant Characteristics Data. Previous studies show that online word-of-mouth does affect restaurant sales because restaurants' quality and popularity can be inferred from such crowdsourced information ([34, 52]). In addition, a restaurant's inherent characteristics also affect customers' choices and the restaurant's profits. To capture those factors, we obtained restaurant characteristics from both *OpenTable* and *Yelp*. From *OpenTable*, we have detailed information on price level (ranging from 1 to 5), number of reviews, star rating (ranging from 1 to 5), and cuisine type. We also collected information about whether the restaurants offer promotion points for consumers to redeem OpenTable Dining Cheque. To obtain more complete promotion information for each restaurant, we crawled restaurants' promotion data from *Yelp* and matched the *Yelp* and *OpenTable* restaurants based on their names, street addresses, and geotags.

3.1.3 Local Census and Weather Data. To better examine the socio-demographics of neighborhoods and control other possible factors, we collected local population information at zip-code level and recorded the average temperature and daily precipitation during the same time period. Population data were obtained from the US Census website (factfinder2.census.gov/) and weather data were crawled from Weatherbase (www.weatherbase.com/).

3.2 Feature Extraction

We created five different sets of features to measure the characteristics of each restaurant, including four location-related categories and one restaurant quality-related feature.

3.2.1 Static Spatial Features. This set of features models a restaurant's static spatial characteristics (STATIC_SPA). Similar to Karamshuk et al. [26], we evaluate it as a vector with four values: location density, population density, heterogeneity, and competitiveness. Formally, the static spatial tuple of restaurant i is:

$$\text{STATIC\_SPA}_i = \left\{ \text{LOC\_DENSITY}_i, \text{HETEROGENEITY}_i, \text{POP\_DENSITY}_i, \text{COMPETITIVENESS}_i \right\}. \tag{1}$$

Density: For each restaurant i, we measure its popularity using the number of nearby restaurants (LOC_DENSITY_i) and population size (POP_DENSITY_i). Formally, with the nearby restaurant $j \in d(i, l)$ (a disk of radius l around restaurant i), the location density is defined as:

$$\text{LOC\_DENSITY}_i = |j | j \in d(i, l)|. \tag{2}$$

[1]In addition to direct distances, we also used Google Map API to compute the distances with Google-Map-based recommended route. The correlation between two types of distances is 0.99. Hence, it is valid to use direct distances as a proxy since the computation of Google-Map-based distances is time-consuming.

Heterogeneity: Similar to the ideas in Karamshuk et al. [26], we use the entropy measurement to assess the level of spatial heterogeneity of an area. Entropy is defined as the expected amount of information from certain events ([13]). We apply it to the frequency of restaurant types in the area. For example, an area with only Chinese restaurants has low heterogeneity, whereas a neighborhood with all kinds of Asian restaurants enjoys a higher heterogeneity. Each restaurant i has its own cuisine type χ_i. We denote $N_\chi(i,l)$ as the number of nearby restaurants with cuisine type χ in disk $d(i,l)$, and $\chi \in \Gamma$, where Γ is a set of all cuisine types. We denote $N(i,l)$ as the total number of restaurants in this area. Formally,

$$\text{HETEROGENEITY}_i = -\sum_{\chi \in \Gamma} \frac{N_\chi(i,l)}{N(i,l)} \times log\left(\frac{N_\chi(i,l)}{N(i,l)}\right). \quad (3)$$

The negative sign indicates that a higher level of diversity in terms of cuisine types has a higher heterogeneity value.

Competitiveness: Given a restaurant i with given cuisine type χ_i, we measure the proportion of nearby restaurants of the same cuisine type χ_i with the total number of restaurants within this area. Intuitively, an area with only Chinese restaurants would have a relatively high level of competitiveness because all the restaurants sell similar products. The restaurant in the most competitive area has the value closest to 1 (which indicates that all the restaurants in that area offer the same cuisine style):

$$\text{COMPETITIVENESS}_i = \frac{N_{\chi_i}(i,l)}{N(i,l)}. \quad (4)$$

3.2.2 Human Mobility Features. As is well known, walkability is an import concept in the design of a community ([15, 44]). Walking is the most common leisure-time physical activity in the United States and has been found to have various economic benefits, including urban neighborhood accessibility, increased efficiency of land use, and improved urban livability ([33]). In this study, we use Foursquare check-in data to measure this human mobility feature (NEIGH_WALK) ([26, 37]) by tracking both spatial and temporal characteristics of users' check-ins. Here, we use $(p,t) \in C$ to denote a check-in recorded in place p and at time t, where C is the set of the Foursquare check-in dataset. Specifically, we measure the *mobility density*, *social stability,* and *incoming mobility* of the area. This feature vector is based on the data that are collected within a certain period (i.e., one day). Mathematically, we define restaurant i's human mobility features as follows:

$$\text{NEIGH\_WALK}_i = \left\{\text{MOB\_DENSITY}_i, \text{SOC\_STABILITY}_i, \text{IN\_MOBILITY}_i\right\}. \quad (5)$$

Mobile Density: To assess the general popularity of an area, we measure the total number of check-ins collected among the neighborhood of restaurant i, within time period T:

$$\text{MOB\_DENSITY}_i = |(p,t)|p \in d(i,l), t \in T|. \quad (6)$$

Social Stability: The popularity of an area can be reflected in two ways: whether it can maintain current consumers for a long period of time and whether it can attract consumers from its neighborhoods. Social stability measures the first scenario, while incoming mobility evaluates the second. We use consumers' consecutive check-in behaviors to assess the stability of current consumers staying in the same place. Here, we define $C_u \subset C$ as the check-in subsets of user $u \in U$, where U represents the set of all users in our data. Formally, by denoting a tuple (p_m, t_m, p_n, t_n)

and two consecutive check-ins (p_m, t_m), (p_n, t_n), we have:

$$\text{SOC\_STABLITY}_i = \sum_{u \in U} \left| \left\{ \begin{array}{c} (p_m, t_m, p_n, t_n) \in C_u | p_m, \\ p_n \in d(i,l), t_m, t_n \; inT \end{array} \right\} \right| \tag{7}$$

Incoming Mobility: One way to show the popularity of a neighborhood is that it attracts people from other neighborhoods for shopping and visiting. Thus, not only the ability to maintain consumers, but also the attraction of potential consumers from other areas can reflect the popularity of an area. To capture this factor, we use consecutive check-in transitions to measure this flow:

$$\text{IN\_MOBILITY}_i = \sum_{u \in U} \left| \left\{ \begin{array}{c} (p_m, t_m, p_n, t_n) \in C_u | p_m \notin d(i,l), \\ p_n \in d(i,l), t_m, t_n \in T \end{array} \right\} \right| \tag{8}$$

3.2.3 Dynamic Traffic Efficiency Features. Traffic efficiency features (denoted as TRA_EFF) measure the dynamic neighborhood accessibility. Every day, there are various emergencies leading to the (partial) closure of certain streets, such as traffic accidents, traffic jams, bus delays, and the like. Such street closure lowers the accessibility of the neighborhood. In our model, we use user-generated content from Twitter to extract dynamic traffic conditions.

3.2.4 Street Closure (Event, Construction) Features. In addition to traffic emergencies, some street closures are longer term, such as road construction or special city events. We use a street closure feature (denoted as STREET_CLO) to measure the average level of street accessibility within a given neighborhood by capturing whether there are any locked-down streets in this neighborhood. This dummy variable indicates whether there are events or street construction projects within a given restaurant's neighborhood. Furthermore, rather than using a simple binary variable, we count the exact number of closed streets using another variable, NUMPROJ.

3.2.5 Restaurant-Specific Features. In addition to the preceding factors, restaurant-level heterogeneity has non-negligible effects on business performance. In order to control for such effects and to determine a causal effect of urban neighborhood accessibility, we build a restaurant-specific feature vector (REST_SPE) with three commonly used elements. We use price level (divided into five degrees), star rating level, and number of reviews to assess the restaurant's popularity and quality. Specifically, restaurant i's restaurant-specific features are denoted:

$$\text{REST\_SPE}_i = \{\text{PRICE}_i, \text{RATING}_i, \text{NUMOFREVIEW}_i\}. \tag{9}$$

Price level: PRICE_i denotes the level of the average price of the restaurant. Based on the data we obtained from OpenTable, we divide price into five levels, with a higher level indicating a higher average price.

Rating: RATING_i represents the quality of the restaurant from OpenTable. In our dataset, we collected the star level of each restaurant, as labeled by thousands of consumers.

Comment reviews: NUMOFREVIEW_i is the aggregated number of reviews about restaurant i on the OpenTable website, which, to some extent, indicates its popularity.

For a better understanding of variables in our setting, we present the definitions and statistics summary of all variables (including the preceding feature variables, as well as outcome variables and controls in the following model section) in Table 1 and display the statistics summary of the important continuous variables in Figure 2.

4 ECONOMETRIC MODELING

As an accepted technique for testing hypotheses and predicting future changes, econometric modeling has the advantage of allowing us to study the effects of interesting variables from a causal perspective. In this article, our econometric model aims to quantify the causal effects of different

Table 1. Definition and Statistics Summary of Variables

Variable	Definition	Mean	Std.Err	Min	Max
Pr(FULL)	Probability of being full	0.2	0.39	0	1
LOC_DENSITY	Number of restaurants	38.86	2.38	0	620
POP_DENSITY	Population size	22,697.27	1.29	144	110,194
COMPETITIVENESS	Proportion of same-type restaurants	0.091	0.12	0	0.67
HETEROGENEITY	Entropy of restaurant types	2.03	1.11	0	3.17
MOB_DENSITY	Total number of mobile check-ins	21.12	3.31	0	1,465
SOC_STABILITY	Consecutive check-ins in the same area	15.8	2.55	0	772
IN_MOBILITY	Incoming flows of mobile check-ins	19.69	2.6	0	608
TRA_EFF	Number of traffic-related tweets	1.67	1.55	0	78
ACCIDENT	Number of accident-related tweets	0.1	0.38	0	5
DISABLED	Number of disabled-vehicles-related tweets	0.1	0.38	0	5
DELAYS	Number of bus delays-related tweets	0.14	0.48	0	8
HEAVYTRAFFIC	Number of heavy traffic-related tweets	0.04	0.26	0	4
WEATHER	Number of weather-related tweets	0.04	0.32	0	9
EVENTS	Number of events-related tweets	0.09	0.55	0	9
STREET_CLO	Whether the area has street closures	0.088	0.28	0	1
NUMPROJ	Number of street closure projects	0.12	0.59	0	19
PRICE	Price dollar level (OpenTable)	2.53	0.62	2	4
RATING	Numerical star rating (OpenTable)	4.02	0.39	1	5
NUMOFREVIEW	Total number of reviews (OpenTable)	40.45	1.24	0	1,451
DEALS	Whether restaurant has deals on Yelp	0.01	0.11	0	1
PROMOTION	Whether restaurant in promotion list (OpenTable)	0.15	0.36	0	1
GOOGLE_TREND	Google search volume of each query	4,428.47	37,854.12	0	1,830,000
TEMPERATURE	Whether temperature is above zero degree.	0.84	0.37	0	1
PRECIPITATION	Whether precipitation is above zero.	0.58	0.49	0	1
HOLIDAY	Whether in the holiday season	0.17	0.38	0	1
Number of Observations: 312,326		Time Periods: 11/29/2013-3/8/2014			

Data source: New York City, with 0.5-mile-range neighborhoods. Variables are computed at daily level.

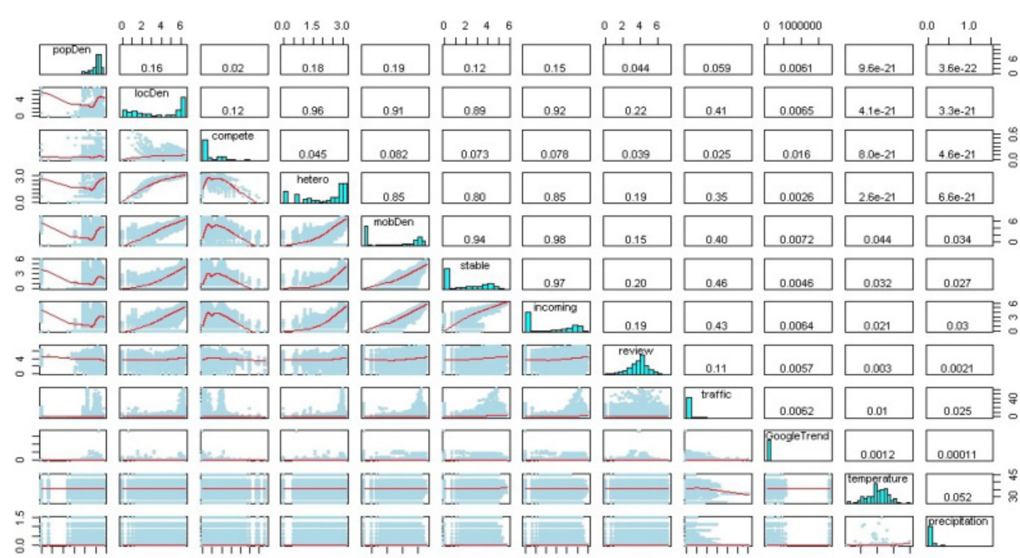


Fig. 2. Data Correlograms. Diagonal: Histograms for the continuous variables in the dataset (population, location density, competitiveness, heterogeneity, mobility density, social stability, incoming mobility, review, traffic efficiency, temperature, precipitation). Upper-right: Correlations of variable pairs. Bottom-left: Scatter plots for joint distributions of variable pairs.

features on the economic outcome of human behavior toward local businesses. In this section, we will discuss in detail how we apply econometric modeling approaches to empirically quantify the different causal impacts of various location-based features.

4.1 Panel Data Analysis

Our panel data are cross-sectional time-series data, which include a collection of observations for multiple restaurants at multiple time series. Therefore, a panel data analysis can better help us address the causal relationship because it considers both the cross-sectional variation across restaurants as well as the temporal variation within each restaurant over time. Specifically, we use a fixed-effect panel model to estimate the impact of different factors in an urban neighborhood on restaurant bookings. Our main model can be formalized in the following equation:

$$\begin{aligned}\mathrm{Pr(FULL)}_{it} &= \alpha_i + \mathrm{STATIC\_SPA}_i \cdot T_t \cdot \delta_1 + \mathrm{HUMAN\_MOB}_{it} \cdot \delta_2 + \mathrm{TRA\_EFF}_{it} \cdot \delta_3 \\ &\quad + \mathrm{STREET\_CLO}_{it} \cdot \delta_4 + \mathrm{REST\_SPE}_{it} \cdot \delta_5 + \mathrm{Controls}_{it} \cdot \phi + T_t + \epsilon_{it},\end{aligned} \tag{10}$$

where $\mathrm{Pr(FULL)}_{it}$ is the probability that a restaurant i is full (i.e., no available reservation slots) at day t. The dependent variable captures the restaurant's booking performance (similar to [1]). We assume that a higher probability of being full potentially indicates a better sales performance of the restaurant. The model includes all features defined earlier: static spatial feature ($\mathrm{STATIC\_SPA}_i$), human mobility feature ($\mathrm{HUMAN\_MOB}_{it}$), traffic efficiency feature ($\mathrm{TRA\_EFF}_{it}$), street closure feature ($\mathrm{STREET\_CLO}_{it}$) and restaurant- specific feature ($\mathrm{REST\_SPE}_{it}$). The coefficients δ_1, δ_2, δ_3, δ_4 and δ_5 capture the impacts of different factors.

The above equation represents both entity fixed effects and time fixed effects: (a) α_i is the restaurant's fixed factor. It is irrelevant to any time period and captures the potential restaurant-level unobserved characteristics that are unlikely to vary over time (e.g., unobserved restaurant quantities such as kitchen size or number of seats). (b) T_t captures the time fixed effect, which controls for the time trend that is common across all the restaurants (e.g., weekend effect). In our study, we consider week dummies, month dummies, and weekday dummies in T_t. Notice that the spatial features ($\mathrm{STATIC\_SPA}_i$) are time-invariant, and, therefore, we drop them from the fixed effect estimation process because α_i includes all time-invariant factors. To capture any potential effects from the spatial features over time, we include an interaction term between the static spatial features and the time trend. In this way, the interaction term $\mathrm{STATIC\_SPA}_i \cdot T_t$ varies in different time periods, and then the effects of static features in different T can be estimated.

The variable $\mathrm{Controls}_{it}$ indicates all possible controls: An interesting thing to note is that our dataset covers the 2013 Christmas and New Year holidays. Furthermore, the 2013 winter was much colder than usual along the northeastern US coast. To account for these potential factors, we consider two additional controls in our model: HOLIDAY (i.e., whether it is during Christmas/New Year holiday) and weather (TEMPERATURE, whether the daily temperature is above zero degrees centigrade; PRECIPITATION, whether the daily precipitation is greater than zero[2]). Moreover, a restaurant's bookings can be affected by its local advertising and marketing efforts. To account for these, we collected additional data on restaurant marketing efforts. For each restaurant, we collected its promotion information (e.g., valid time period of deals) in Yelp (i.e., DEALS) and from OpenTable (i.e., PROMOTION, whether the restaurant is on OpenTable's promotion list). Finally, ϵ_{it} is an independent and identically distributed random error term.

[2]We considered using average daily temperature and precipitation instead of the dummies. The findings are similar.

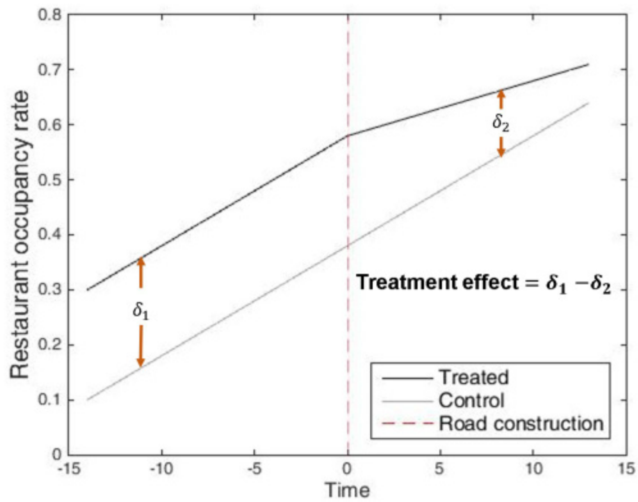


Fig. 3. Framework of exploring causal treatment effects using Difference-in-Difference method. δ_1 is the pretreatment difference in the outcome (i.e., restaurant occupancy rate) between treated and control groups, and δ_2 is the posttreatment difference. The change between δ_1 and δ_2 is the causal effect driven by treatment.

4.2 Causal Effects of Street Closures

The potential selection bias in street events and street construction is one challenge in studying the economic outcome of human behavior. Specifically, in the context of street closure, selection bias can be caused by unobserved factors. For example, the reason that the city planner chooses a particular street to close for a local event or for construction may be due to some unobserved functional inability of that street (e.g., poor street condition, focal inconvenience). Such unobserved factors may cause both the decision of street closure and the decrease in sales for local stores, regardless of the street closure. To account for such an endogeneity issue and to identify the impact from a causal perspective, we conduct an additional analysis by combining Propensity Score Matching (*PSM*) [38] and Difference-in-Difference (*DID*) methods to examine the causal effect of street closure. The basic idea of the DID method is to compare the average change over time in the outcome variable between the treated and control groups. The difference in change suggests causal treatment effect. We illustrate the basic intuition of our analysis design in Figure 3.

First, we consider a 4-week time window as the experiment period and divide it into two time periods: The first 14 days are the baseline period, while the latter 14 days are the test period. In the baseline period, no street closure (i.e., events or construction) occurs within a 0.5-mile range of all the restaurants. In the test period, some restaurants experience street closure within the same area.[3] Second, we divide restaurants into two groups: a Treatment group in which the restaurants have at least one nearby street closure in the test period, and a Control group in which the restaurants remain unaffected in the overall 4-week time window. Third, to address the issue of selection bias in street closure, we use Propensity Score Matching (*PSM*) for the counterfactual analysis. The idea of *PSM* is to match restaurants in the Treatment group with those in the Control group based

[3]We selected the time period with the largest number of treated samples: from December 24, 2013 to January 20, 2014. We filtered the whole sample to make the resulting samples satisfy the requirements of period division. To account for the potential bias introduced by the time period selection, we tested different starting times or different lengths of time window. The results stay highly consistent.

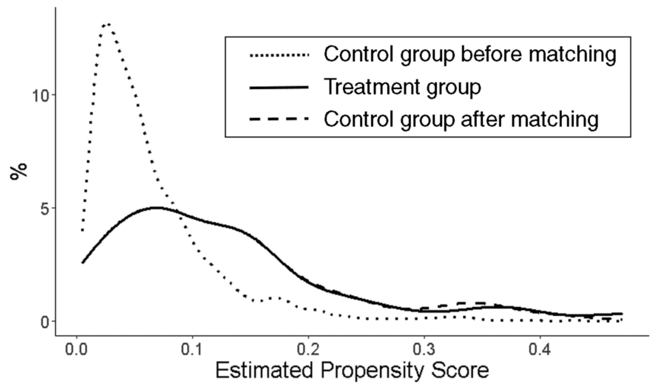


Fig. 4. Distribution of Propensity Scores for Treatment Group and Control Groups (both matched and unmatched). This figure indicates that the matched control restaurants have a propensity score distribution more similar to the treated ones than the unmatched control restaurants.

on their likelihood (i.e., propensity score) of being treated. The matching process would help eliminate the concern that some other observed restaurant characteristics would potentially lead to both the treatment decision and the observed outcome. Specifically, a logit regression is used to estimate the propensity score for each restaurant:

$$P(D_{it} = 1|V_{it}) = \frac{1}{1 + exp((-logit_{it})}, \tag{11}$$

where

$$\begin{aligned} logit_{it} = \alpha_i &+ \text{STATIC\_SPA}_i \cdot T_t \cdot \delta_1 + \text{HUMAN\_MOB}_{it} \cdot \delta_2 \\ &+ \text{TRA\_EFF}_{it} \cdot \delta_3 + \text{STREET\_CLO}_{it} \cdot \delta_4 + \text{REST\_SPE}_{it} \cdot \delta_5 + \epsilon_{it}. \end{aligned} \tag{12}$$

In the Logit regression function, the propensity score $P(D_{it} = 1|V_{it})$ indicates the likelihood of the restaurant being selected in the treatment group. V_{it} represents the observable feature vectors (i.e., static special features, human mobility features, traffic efficiency features, street closure features, and restaurant-specific features) of restaurant i at time t. In the matching process, we use the K-nearest neighbor algorithm. Specifically, the optimal matched pairs of treated and control observations are those that produce the minimum distance in their propensity scores. Therefore, the restaurants in a matched pair share a similar possibility of being selected for treatment (i.e., street closure). However, the only difference between a matched pair is that one is being treated and the other is not, which nicely simulates a randomized control experimental setting. Note that *PSM* is particularly appropriate in our case because (i) we have a large number of sample observations, and (ii) we are able to incorporate a large variety of observed time-varying and time-invariant restaurant-level characteristics into the matching process. Both advantages allow us to identify pairs of restaurants with high similarity. Figure 4 shows the performance of our propensity score matching, which indicates that the matched control restaurants have a propensity score distribution more similar to the treated ones than the unmatched control restaurants.

Finally, based on the matched samples, we use the *DID* method to test the causality. In particular, to ensure that there are no unobserved differences related to the treatment (i.e., the quality may differ even within the two matched samples due to unobserved features), we apply *DID* to exploit the exogenous variance in street closure across restaurants and time as the basis for identifying

Table 2. Main Estimation Results

Category	Variable	Coef^{M}		Coef^{I}		Coef^{II}	
Human Mobility	MOB_DENSITY$^{(L)}$	0.004**	(0.002)	0.115***	(0.032)	0.010**	(0.003)
	SOC_STABILITY$^{(L)}$	−0.001	(0.002)	−0.084**	(0.028)	0.001	(0.003)
	INC_MOBILITY$^{(L)}$	0.003	(0.002)	0.016	(0.037)	−0.001***	(0.002)
Traffic Efficient	TRA_EFF$^{(L)}$	−0.005***	(0.001)	−0.099***	(0.013)	−0.008***	(0.001)
Street Closure	STREET_CLO	−0.014***	(0.004)	−0.192***	(0.047)	−0.014***	(0.004)
Interaction term between Static Spatial Features and Monthly Indicators	LOC_DENSITY$^{(L)}$ × m_1	0.002	(0.004)	0.064	(0.048)	0.003	(0.004)
	POC_DENSITY$^{(L)}$ × m_1	0.011***	(0.001)	0.104***	(0.011)	0.011***	(0.001)
	HETEROGENEITY × m_1	0.008	(0.008)	0.042	(0.102)	0.006	(0.008)
	COMPETITIVE × m_1	−0.001	(0.004)	−0.373	(0.248)	−0.023	(0.021)
	LOC_DENSITY$^{(L)}$ × m_2	−0.001	(0.004)	0.021	(0.049)	−0.001	(0.004)
	HETEROGENEITY × m_2	0.008	(0.008)	0.038	(0.103)	0.007	(0.004)
	COMPETITIVE × m_2	−0.053*	(0.021)	−0.651**	(0.242)	−0.053*	(0.021)
	LOC_DENSITY$^{(L)}$ × m_3	0.000	(0.003)	0.008	(0.009)	−0.001	(0.004)
	POC_DENSITY$^{(L)}$ × m_3	0.001	(0.001)	0.004	(0.009)	0.001	(0.001)
	HETEROGENEITY × m_3	0.009	(0.008)	0.104	(0.103)	0.010	(0.008)
	COMPETITIVE × m_3	−0.031	(0.021)	−0.435	(0.251)	−0.029	(0.021)
Restaurant Specific Features	PRICE	−0.034*	(0.011)	−0.219*	(0.109)	−0.034**	(0.011)
	RATING	0.002	(0.004)	0.025	(0.043)	0.003	(0.004)
	NUMREVIEW$^{(L)}$	0.013***	(0.002)	0.135***	(0.022)	0.013***	(0.002)
Controls	Promotion	Yes		Yes		Yes	
	Weather	Yes		Yes		Yes	
	Time	Yes		Yes		Yes	
	Google Trend	Yes		Yes		Yes	
Observations		258,090		258,090		Y258,090	
M: Main estimation results.		I: Robustness test I (Logit model).					
II: Robustness test IV (1 mile)		(L): Logarithm of the variable.					

Controls: promotion, temperature, precipitation, and Google trends; Methods: entity and time fixed effects; Data: 0.5-mile neighborhoods in NYC.
*p-value <0.05 **p-value<0.01 ***p-value<0.0001.
Standard errors are shown in parentheses.

causal effects on local restaurant sales. Following previous studies [48], our model is as follows,

$$Pr(\texttt{FULL})_{it} = \alpha_i + \beta_1 \texttt{Test}_t + \beta_2 \texttt{Test}_t \times \texttt{Treat}_i + \texttt{Controls}_{it} \cdot \phi + T_t + \epsilon_{it}, \tag{13}$$

where α_i is restaurant-level fixed effect, $Test_t$ indicates the test ($t = 1$) or baseline ($t = 0$) period, and $Treat_i$ indicates whether restaurant i is in the treatment group. Note that, similar to the main estimation, we add additional control variables, such as weather, holiday indicator, and the like. The coefficient of interest is β_2, which captures the effects of street closure in the test period. The control variables are the same as those used in Equation (10).

5 RESULTS

5.1 Panel Data Model Results

We first start with our main estimation model (Equation (10)), the main coefficients of which are shown in Table 2. We allow interactions between static spatial features and time trend indicators to capture the impacts of static features over time. Specifically, we define four monthly indicators: November and December jointly (m_1),[4] January (m_2), February (m_3), and March (m_4). To avoid collinearity, we use only the first three indicators in the regression.

[4]Our data contain two days from November 2013, so we merge them into the December month dummy.

Our estimation yields some interesting findings. First, among the three elements of human mobility features, only mobile density shows significant effect. Specifically, the coefficient of mobile density indicates that a 1% increase in the unit of mobile density will lead to a 0.004 increase in the probability of being full. Although the magnitude of this estimate is small, it would turn into a significant increase with other outcome measures, such as the restaurant's revenues. To the contrary, the effects of social stability and incoming mobility are not significant. This quantifies the business potential of a popular place with accessible human walkability (e.g., shopping mall, tourist attractions, etc.). Second, the two significant negative estimates of traffic efficiency feature and street closure feature present their impacts on urban small business performance. In particular, the marginal effect of the street closure feature is -0.005, indicating that, compared to a restaurant whose neighborhood has no street closure project, a restaurant that is near a street closure project (due to either road construction or city events) would have a 0.014 decrease in its probability of being full. This impact is much higher than most of the other estimates, meaning that street closure has higher negative impact on business performance than the others. Hence, one crucial implication from this finding is that when choosing the proper location, a new restaurant needs to avoid an area that has long-term street construction. With regard to restaurant-specific features, consistent with theories,[5] we find that price has a negative effect on restaurant bookings and that the effect of price is significantly larger than that of the other features. The number of reviews presents a significant and positive effect. In addition, our results also show that warm, sunny weather has a significant and positive effect on local restaurants. This is consistent with previous studies that use weather or climate as one measure of an urban system [10, 40]. However, our finding makes a further step to quantify the economic value of this factor. Regarding the interactions between static spatial features and time trends (i.e., location density, population density, heterogeneity, and competitiveness with month indicators m_1, m_2, m_3), we find that most of them do not have significant impacts, suggesting that most effects from the static spatial features are time-invariant and absorbed by the fixed effect. The preceding results are based on lag-term instrument variables. We also use our alternative instrument variables and obtain similar results. The results in Figure 7 illustrate effects from all time-varying variables: mobility density, social stability, incoming mobility, traffic efficiency, street closure, price comment reviews, and ratings.

5.2 PSM and DID Model Results

To deal with potential selection bias in street closure, we combine the *PSM* and *DID* methods to explore causal effects. Column (i) in Table 3 shows the coefficients from our causal estimation. The coefficient of "Test" is positive, indicating that, on average, the baseline booking trend is increasing during this test time period. This is reasonable because it is the holiday season when more consumption is likely to occur. Interestingly, we find a significant and negative sign of the interaction term "Test×Treat," suggesting a negative causal effect of street closure on bookings.

One might argue that the time period we cover is special because it might cover some unobserved feature related to the holiday. To better assess our model and results, we conduct robustness tests on several alternative periods before and after this holiday season. We find that the interaction term still shows a significant negative sign, whereas the baseline time trend is not significant. Column (ii) in Table 3 shows the results from one alternative period. Furthermore, to measure the treatment effect at different levels of street closure, we add another interaction term,

[5]Rating effect is not statistically significant in this context. We notice that more than 75% of restaurants have a star rating higher than or equal to 3.9, showing a relative small variance. Due to the potential inflation of the numerical ratings, it may result in the nonsignificant coefficient. But we do observe that this effect is positive, which is consistent with previous findings [3, 15].

Table 3. PSM and DID Model Results on Causal Impact of Street Closure

Variables	(i)	(ii)	(iii)	(iv)
Test × Treat × NUMPROJ	–	–	-0.047*	–
			(0.024)	
Test × Treat	−0.074***	−0.018*	−0.058***	−0.053**
	(0.018)	(0.007)	(0.019)	(0.020)
Test	0.066***	0.024	0.067***	0.057**
	(0.018)	(0.018)	(0.011)	(0.002)
Test × Treat × Chain	–	–	–	−0.421***
				(0.102)
Promotion control	Yes	Yes	Yes	Yes
Weather control	Yes	Yes	Yes	Yes
Time control	Yes	Yes	Yes	Yes
Observations	11,424	11,144	11,424	8.400
(i)&(iii)&(iv): 12/24/2013-1/20/2014; (ii): 11/29/2013-12/26/201				

The estimated coefficient of "Test×Treat" indicates a statistically significant and negative treatment effect of street closure on restaurant bookings.
*p-value <0.05 **p-value<0.01 ***p-value<0.0001.
Standard errors are shown in parentheses.

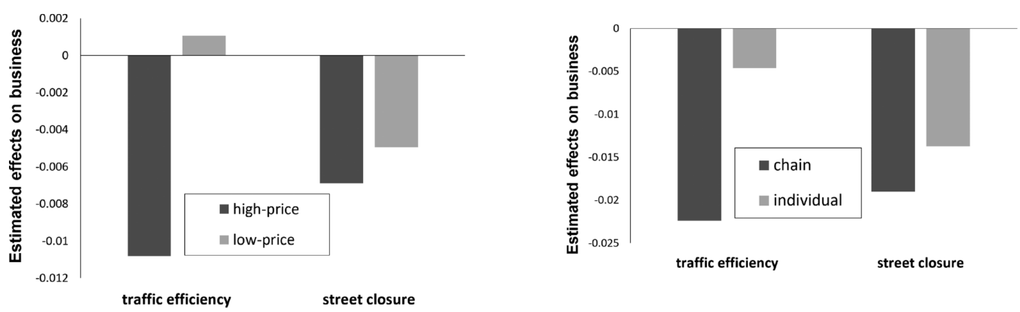
Test × Treat × NUMPROJT (the number of nearby street events/constructions), which is similar to that in Tucker and Zhang [48]. The corresponding model is described in Equation (14). The result is shown in Column (iii), Table 3. We find results consistent with our main model. Moreover, coefficient δ_6 is negative and significant, suggesting that one or more street event nearby leads to a 4.7% decrease in the probability of a restaurant being fully booked.

$$\begin{aligned}\Pr(\text{Full})_{it} = {} & \alpha_i + \beta_1 \text{Test}_t + \beta_2 \text{NUMPROJ}_{it} + \beta_3 \text{Test}_t \times \text{Treat}_i + \beta_4 \text{Test}_t \times \text{NUMPROJ}_{it} \\ & + \beta_5 \text{Treat}_i \times \text{NUMPROJ}_{it} + \beta_6 \text{Test}_t \times \text{Treat}_i \times \text{NUMPROJ}_{it} \\ & + \text{Controls}_{it} \cdot \phi + T_t + \epsilon_{it}\end{aligned} \tag{14}$$

5.3 Interaction Effects Results

In the previous process, we considered the 3,187 restaurants in our sample as a single group, which might lead to some bias because of heterogeneity at the restaurant level. In this subsection, we will look into smaller restaurant groups and examine the interaction effects of those features of interest. The results of the following two interaction models are shown in Figure 5.

Interaction Model I: Interaction effects with price level indicator: First, to explore how effects of the traffic efficiency feature and the street closure feature vary with price level, we divide the restaurants into two groups: expensive restaurants and cheap restaurants. Then we add two interaction terms between price dummies (denoting whether or not the price is high) and the two traffic-related features: traffic efficiency feature and street closure feature. We hold other things constant, as in the main estimation (Equation (10)). The results show that the coefficients of the interaction terms are significantly negative, indicating that higher priced restaurants are more like to be affected by traffic conditions. Figure 5(a) illustrates the coefficients of each feature within each group.



(a) Price Levels (b) Chain/independent Restaurants

Fig. 5. Comparisons of interaction effects of traffic-related features (i.e., traffic efficiency, street closure) on restaurant bookings. (a) Comparison between high-price and low-price restaurants; (b) comparison between chain and independent restaurants.

Interaction Model II: Interaction effects with chain or independent restaurant indicator: Next, in order to examine whether brands have any impacts under this scenario, we divided the 3,187 restaurants into three groups: chain restaurants, independent restaurants, and others. Among them, there are 86 well-established chain restaurants with 15 brands and 2,354 independent restaurants. By using interaction terms combining the chain dummy (denoting whether it is a chain restaurant) with the traffic efficiency feature and street closure feature, we run a fixed-effect regression over the 2,440 restaurants. The coefficients are both positive, while only the coefficient of the interaction term between the chain dummy and traffic efficiency feature is significant. It implies that chain restaurants will be affected more than independent restaurants by unexpected traffic conditions. Figure 5(b) illustrates such differences.

Furthermore, we apply the preceding division to the *PSM* and *DID* estimation procedure to explore whether different restaurants (e.g., chain and individual) would be affected by street conditions differently:

$$\begin{aligned}\Pr(\mathrm{FULL})_{it} &= \alpha_i + \beta_1 \mathrm{Test}_t + \beta_2 \mathrm{Test}_t \times \mathrm{Treat}_i + \beta_3 \mathrm{Test}_t \times \mathrm{Treat}_i \times \mathrm{chain}_i \\ &\quad + \mathrm{Controls}_{it} \cdot \phi + T_t + \epsilon_{it},\end{aligned} \tag{15}$$

where $chain_i$ is a dummy indicator variable. Again, the lower-order interaction term $chain_i$ is excluded because it is collinear with the fixed effects. The results are shown in Column (iv), Table 3. We found that both β_2 and β_3 are significant and negative (i.e., $\beta_3 = -0.4214$ and $\beta_2 = -0.053$), suggesting that chain restaurants tend to be affected more than independent restaurants by road closures.

Interestingly, our findings from this interaction model seem to suggest that chain restaurants are likely to be much more negatively affected by the street closures when compared to independent restaurants. This is reasonable because, for chain restaurants, when one location becomes less accessible customers who really like the food tend to substitute away to an alternative location with easy access to the same chain restaurants. However, for independent restaurants, customers who really like the food do not have an easy alternative for substitution. As a result, they may have a much higher switching cost compared to the case of chain restaurants, which might help keep independent restaurants from losing customers. Our results have potential in helping franchised restaurant chains to better understand the effects of city events and street closures, and to improve their marketing strategies to reduce the potential economic loss.

6 ROBUSTNESS TESTS

In this section, we aim to examine the robustness of our results with a discussion about the identification issue in our main econometric model, several falsifications, and robustness checks, as well as a model comparison.

6.1 Identification

To establish a causal relationship between local demand and all those features of interest, we need to rule out reverse causal explanations and unobserved variables that can cause both the performance outcome and features. This section discusses two potential types of endogeneity: (i) price endogeneity and (ii) potential endogeneity in traffic and human mobility characteristics.

6.1.1 Price Endogeneity. One challenge in estimating price effects on restaurant bookings is that restaurant owners may change their price in response to demand, and consumers change their demand in response to price. This loop of causality is referred to as the Price Endogeneity issue in economics. Without ruling out such endogeneity concerns, we cannot draw a causal conclusion about the quantity of the effects on outcome performance merely from the coefficient of price. To account for price endogeneity, we apply two commonly used instrument variables (IV) methods: Villas-Boas-Winer-style IVs ([49]) and Hausman-style IVs ([24]).

Villas-Bios-Winer-style IVs: Following Archak et al., Ghose et al., and Villas-Boas and Winer [3, 22, 49], we use lagged prices as IVs with Google Trend data. This dataset records the number of searches for each restaurant's name at the monthly level. The intuition for this IV method is that prices in different time periods are correlated with each other because of common costs (e.g., restaurant employee salaries, operational costs, cost for food materials). However, cost is likely to be stable and uncorrelated with market demand in the short run. Therefore, we can use the lagged price (i.e., the last-period price) as an IV to substitute for the current period price in the model.

Note that lagged price is a valid IV only if the unobserved variables are not correlated over time ([2]). One may argue that there might exist some common demand shock over time (e.g., product popularity or trend), which could potentially be correlated not only with current-period price but also with last-period price. If so, the lagged price may not be a valid IV because it will be once again correlated with current demand. However, common demand shock that is correlated over time is essentially a trend. In particular, the search volume of each restaurant's name extracted from Google Trend data can reflect the demand trends of these restaurants. Using a similar approach as in Archak et al. and Ghose et al., [3, 22], we control for restaurant-specific time trends using Google Trend data to alleviate such concerns.

Hausman-style IVs: As discussed in Brynjolfsson et al., Ghose et al., and Nevo [6, 22, 36], the idea is to use the average price of other similar restaurants (i.e., with the same star ratings or same cuisine type) in other markets (i.e., neighborhoods). The intuition is that the prices of similar restaurants are correlated with respect to similar costs, but the demand shocks in different markets are unlikely to be correlated. Hence, the average price at similar restaurants in other markets can be a valid IV for the price of the focal restaurant. In addition, we also use various control variables (i.e., promotions, holidays, weather) to account for the time-varying unobserved factors.

6.1.2 Endogeneity in Traffic and Mobility Features. Traffic and human mobility characteristics also have potential endogeneity issues because both of these mobility characteristics and the restaurant bookings might be correlated with local business popularity or advertising promotions. We consider similar instrumental variable methods as earlier for addressing the price endogeneity issue:

Villas-Bios-Winer-style IVs: Similar to the usage of lagged price, we use lagged (i.e., last time period) traffic/human mobility variables, together with Google Trend data, as the IVs of the traffic and human mobility variables of the current time period. The intuition is that dynamic traffic and human moving patterns are correlated over time because of stable community designs. For example, a shopping mall always enjoys a relatively high popularity and traffic pressures in different time periods. And such stable patterns are less likely to be affected by a short-term demand shock.

Hausman-style IVs: The intuition here is that traffic and human mobility can be highly related to local neighborhood development costs. However, such costs are unlikely to be correlated with market demand changes in the short run. Therefore, we consider the neighborhoods of similar restaurants as an indicator for the urban development condition of neighborhoods of the given restaurant. The "similar" restaurants can be selected using various criteria, including restaurants with the same ratings, same price levels, or same cuisine types. It is a realistic approximation because local restaurants with similar characteristics are likely to target consumers with similar tastes, demographics, and consumption levels, which, to a large extent, indicate the local development condition of a neighborhood.

6.2 Falsification Check

A plausible concern is that the performance of restaurants may lead to the street closure decision. This might occur when the government decides to improve the popularity of an area by improving its traffic conditions. In this case, our identification strategy for the effects of street closures on small businesses becomes questionable. To defend against this threat, we conduct two different tests for our checks: (i) We use lagged performance to predict current street closures and estimate a logistic regression with restaurant fixed effects, and (ii) we test whether there are anticipation effects of street closures by regressing current performance on the lead value of street closure. For further validation, we test whether street closure affects the performance of restaurants that are far from the closed streets. To avoid some potential noisy factors of short-term closures that may be caused by emergencies, we consider only long-term street closures (longer than 1 week) in this test. Since the neighborhood we used earlier is 200 meters in size, we select some other 200-meter areas that are far from the closed streets (e.g., at a direct distance of 2,500 meters to 2,700 meters). We use the main regression but include two street closures dummies indicating whether there are closed streets outside or within its neighborhood.

Empirically, first, using two different models, we check whether reverse causality exists in our settings. The results are shown in Table 4. In both cases, we find insignificant coefficients of lagged performance or lead value of street closure. Thus, these models do not seem to provide any evidence of reverse causality. Second, we test whether the street closure will affect the performance of restaurants that are far from the closed streets. The results show that the coefficient of the outside-neighborhood dummy is insignificant while the within-neighborhood dummy is still significant. The lack of evidence of closed streets' effects on remote restaurants further strengthens our main results.

6.3 Robustness Tests

To assess the robustness of features, model, and results, we conduct four additional robustness tests:

Robustness Test I: Use the same variables on alternative models: The dependent variable is discrete, covering six probability numbers. In this sense, the linear regression model may not fit the data very well. We tried different models, such as logit. In this alternative model, we consider the dummy dependent variable as the indicator of whether the reservation is available at 7 pm, which

Table 4. Falsification Checks Results

	Coef^I	**Coef^II**	**Coef^III**
STREET_CLO (lead value)	−0.0034(0.004)	–	–
Y (lagged value)	–	0.031(0.049)	–
STREET_CLO (within neighborhood)	–	–	-0.014^{***}(0.004)
STREET_CLO (outside neighborhood)	–	–	−0.000(0.003)
Observations	255,439	68,727	258,090

* p-value <0.05 **p-value<0.01 ***p-value<0.0001.
In the falsification check of reverse causality, we first regress current performance on the lead value of street closure (Coef^{I} column); then we use lagged performance to predict current street closures in a logistic form (Coef^{II} column). In the second falsification check (Coef^{III} column), we compare effects of outside- and within-neighborhood street closures.
Both checks include all other variables and controls as shown in Equation (10).

Table 5. Comparison of Fixed Effects Model and Random Effects Model

Category	Variable	Coef^{M}		Coef^{I}	
Human Mobility	MOB_DENSITY$^{(L)}$	0.004^{**}	(0.002)	0.008^{***}	(0.002)
	SOC_STABILITY$^{(L)}$	−0.001	(0.002)	0.002	(0.002)
	INC_MOBILITY$^{(L)}$	0.003	(0.002)	−0.004	(0.002)
Traffic Efficient	TRA_EFF$^{(L)}$	-0.005^{***}	(0.001)	-0.005^{***}	(0.001)
Street Closure	STREET_CLO	-0.014^{***}	(0.004)	-0.014^{***}	(0.004)
Restaurant-Specific Features	PRICE	-0.034^{*}	(0.011)	−0.001	(0.007)
	RATING	0.002	(0.004)	0.001	(0.001)
	NUMREVIEW$^{(L)}$	0.013^{***}	(0.002)	0.001^{***}	(0.002)
Controls	Promotion	Yes		Yes	
	Weather	Yes		Yes	
	Time	Yes		Yes	
	Google Trend	Yes		Yes	
Observations		258,090		258,090	
M: Main estimation results.		I: Random effects results			
(L): Logarithm of the variabl.					

*p-value <0.05 **p-value<0.01 ***p-value<0.0001.
Standard errors are shown in parentheses.

we believe is the most common time that NYC residents go out for dinner. We find similar results with the main estimation.

Robustness Test II: Replace fixed effects with random effects: In our main model, we combine spatial features with time trend to see impacts over time because the entity fixed-effects method omits all time-invariant and individual level features. To test the effectiveness of these static features directly, we test the random effects model (shown in Table 5) and find similar results. By using the Hausman test [23], we find that our fixed effects model performs better.

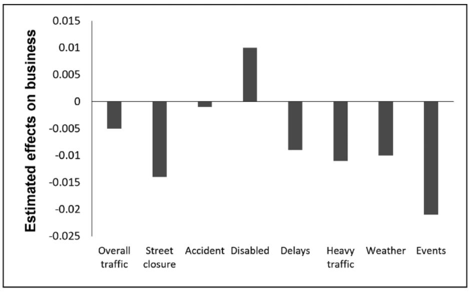


Fig. 6. Effects of individual traffic types.

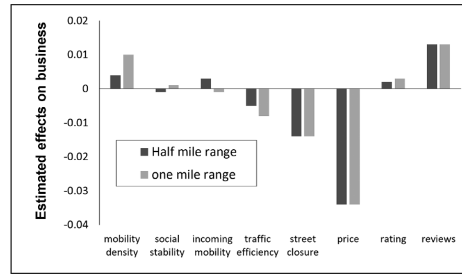


Fig. 7. Comparison of effects between 0.5-mile and 1-mile range neighborhoods.

Robustness Test III: Use detailed traffic information: To extract the detailed dynamic traffic conditions, we apply the keyword-extraction technique to classify tweets into different types based on their keywords: traffic accidents, heavy traffic jams, bus delays, etc. That is, we divide the TRA_EFF into six subvariables: ACCIDENT, DISABLED, DELAYS, HEAVYTRAFFIC, WEATHER, and EVENTS (the detailed definitions are provided in Table 1). We find very similar trends for all factors, and the results are shown in Figure 6. In particular, we find a significant negative effect of bus delays on business performance. One explanation is that our dataset was collected in NYC, where public transportation is a major choice, especially during rush hour (dinner time).

Robustness Test IV: Use alternative range of neighborhood on the same model: To examine whether a 0.5-mile range is a valid definition of neighborhood and whether neighborhood size matters greatly in our estimation, we consider neighborhoods of different sizes. The result, shown in Figure 7, is that the impact of each factor is similar to that of the 0.5-mile range, while the mobile density and dynamic traffic features show larger impacts.

6.4 Model Comparisons

To evaluate the features we proposed in predicting economic values under the urban system, we compare our model with multiple alternative models. Specifically, we started with a single logistic model with human mobility features only and then added, step-by-step, street closure features, dynamic traffic efficiency features, static spatial features, and finally restaurant-specific features. The Receiver Operating Characteristic (ROC) curves are plotted in Figure 8. First, we show that all features have value in predicting the economic values, as the prediction performance increases as more features are added into the regression model. Second, models 1, 2, and 3 track the performance of dynamic features. These show that mobility features have the largest power in prediction. This plot also indicates significant improvement from M3 to M4, and M4 to complete model, where we added spatial features and restaurant-specific features, respectively. This suggests that, in predicting the performance of small businesses in an urban setting, it is important to consider all three factors: the static and dynamic features of the neighborhood and the restaurant-specific characteristics. Last, but not the least, we show that our complete model (i.e., with all proposed features) performs significantly better than alternative models.

7 DISCUSSION AND FUTURE WORK

In this article, we explore economic values in the urban system based on geotagged and crowd-sourced data from various large-scale social media sites and publicly available data sources. Using geomapping and geo-social-tagging techniques, we identify four feature dimensions to describe

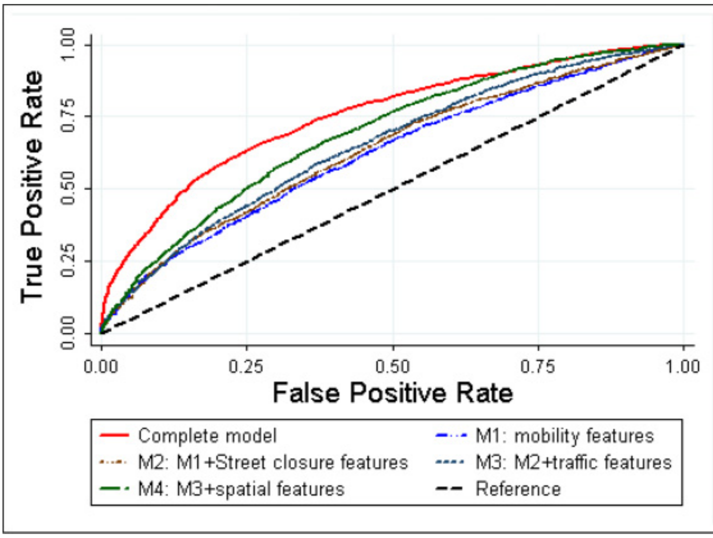


Fig. 8. Model comparisons. We compare our complete model with the following alternative models: M1: model with only human mobility features; M2: model with human mobility features and street closure features; M3: model with human mobility features, street closure features, and traffic efficiency features; M4: model with human mobility features, street closure features, traffic efficiency features, and static spatial features.

the potential social and economic factors of local demand. After evaluating these features while also accounting for the potential endogeneity issues, our econometric model is able to quantify the economic and social value of the extracted features on local demand from a causal perspective.

On a broader note, the objective of this article is to illustrate how multiple and diverse sources of publicly available crowdsourced data can be mined and incorporated into the prediction of local demand to enhance the understanding of users' economic behavior through its interactions with local businesses. Our study demonstrates the potential for best making use of large volumes of user-generated content and geotagged social media data to create matrices that capture multidimensional characteristics in a manner that is fast, cheap, accurate, and meaningful. Local businesses can use this information to proactively design their business strategies (e.g., advertising and promotions) when facing a potential change in city neighborhood services. Furthermore, it can help government decision makers to understand local economic trends. For example, it is useful for urban planners to be able to quantify the opportunity cost and, moreover, the overall expected economic outcome of an urban project or event in a location under various urban and economic conditions. Since our data come from publicly available channels, we can easily apply our methodology to other categories of local businesses in various locations. Such analyses can help small businesses gain insights into their local urban systems and economies, which, in turn, increases their success and the sustainability of urban neighborhoods.

Our research also has implications for location-based services, such as Google Maps, by making it possible to incorporate data into understanding local neighborhoods. Specifically, they can use the model we propose to specify location efficiency scores in predicting the economic potential for a new market. For example, one possibility would be to provide an "economic index" of each neighborhood for new businesses to predict their demand in different locations and thus optimize their location selection.

Our work has several limitations, some of which can serve as fruitful areas for future research. Our analysis is based on a randomly selected subset of Twitter and Foursquare data. It can be improved by leveraging more data from other crowdsourced channels to gain a more comprehensive understanding of traffic and human mobility conditions. Specifically, our traffic-related features are only an approximation of traffic conditions extracted from tweets post by NYC transportation government. This might have some limitations, including the possible time gap between the real-time condition and posting time. It can be improved with the use of sensor tools that record and transmit real-time data so we can extract traffic features. Such data would be more accurate without being more costly in extraction. With the new data, our model can still provide a fast way to predict or evaluate effects. Also, in order to better predict local demand, future work can look into not only the geographic and socioeconomic perspectives of cities, but also other natural and environmental aspects, such as climate and pollution factors, health care, and more. Such research would help us draw a comprehensive picture of the overall urban system and study economic dynamics and social interactions more precisely.

REFERENCES

[1] Michael Anderson and Jeremy Magruder. 2012. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal* 122, 563 (2012), 957–989.

[2] Joshua Angrist and Alan B. Krueger. 2001. *Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments*. Technical Report. National Bureau of Economic Research.

[3] Nikolay Archak, Anindya Ghose, and Panagiotis G. Ipeirotis. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science* 57, 8 (2011), 1485–1509.

[4] Peter C. Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46, 3 (2011), 399–424.

[5] Glenn C. Blomquist, Mark C. Berger, and John P. Hoehn. 1988. New estimates of quality of life in urban areas. *The American Economic Review* (1988), 89–107.

[6] Erik Brynjolfsson, Yu Hu, and Michael D. Smith. 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science* 49, 11 (2003), 1580–1596.

[7] BusinessWeek. 2013. It costs $333 million to shut down Boston for a day. Retrieved from www.businessweek.com/articles/2013-04-19/it-costs-333-million-to-shut-down-boston-for-a-day.

[8] Peter Calthorpe. 1993. *The Next American Metropolis: Ecology, Community, and the American Dream*. Princeton Architectural Press.

[9] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. 2011. Exploring millions of footprints in location sharing services. *ICWSM* 2011 (2011), 81–88.

[10] Paul C. Cheshire and Stefano Magrini. 2006. Population growth in European cities: Weather matters–but only nationally. *Regional Studies* 40, 1 (2006), 23–37.

[11] Judith A. Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* 43, 3 (2006), 345–354.

[12] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD*. ACM, 1082–1090.

[13] Thomas M. Cover and Joy A. Thomas. 2012. *Elements of Information Theory*. John Wiley & Sons.

[14] Brett Danaher and Michael D. Smith. 2014. Gone in 60 seconds: The impact of the Megaupload shutdown on movie sales. *International Journal of Industrial Organization* 33 (2014), 1–8.

[15] Reid Ewing and Susan Handy. 2009. Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design* 14, 1 (2009), 65–84.

[16] Richard Florida. 2002. The economic geography of talent. *Annals of the Association of American Geographers* 92, 4 (2002), 743–755.

[17] Chris Forman, Anindya Ghose, and Avi Goldfarb. 2009. Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management Science* 55, 1 (2009), 47–57.

[18] Chris Forman, Avi Goldfarb, and Shane Greenstein. 2005. How did location affect adoption of the commercial internet? Global village vs. urban leadership. *Journal of Urban Economics* 58, 3 (2005), 389–420.

[19] Yanjie Fu, Yong Ge, Yu Zheng, Zijun Yao, Yanchi Liu, Hui Xiong, and Jing Yuan. 2014. Sparse real estate ranking with online user reviews and offline moving behaviors. In *Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM),* IEEE, 120–129.

[20] Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. 2014. Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1047–1056.

[21] Sarah Jean Fusco, Katina Michael, and M. G. Michael. 2010. Using a social informatics framework to study the effects of location-based social networking on relationships between people: A review of literature. In *Proceedings of the 2010 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 157–171.

[22] Anindya Ghose, Panagiotis G. Ipeirotis, and Beibei Li. 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science* 31, 3 (2012), 493–520.

[23] Jerry A. Hausman. 1978. Specification tests in econometrics. *Econometrica: Journal of the Econometric Society* (1978), 1251–1271.

[24] Jerry A. Hausman. 1996. Valuation of new goods under perfect and imperfect competition. In *The Economics of New Goods*. University of Chicago Press, 207–248.

[25] Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. 2013. STED: Semi-supervised targeted-interest event detectionin in twitter. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1466–1469.

[26] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. 2013. Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD*. ACM, 793–801.

[27] Kevin J. Krizek. 2003. Operationalizing neighborhood accessibility for land use-travel behavior research and regional modeling. *Journal of Planning Education and Research* 22, 3 (2003), 270–287.

[28] Kun Kuang, Meng Jiang, Peng Cui, and Shiqiang Yang. 2016. Steering social media promotions with effective strategies. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

[29] Dionysia Lambiri, Bianca Biagi, and Vicente Royuela. 2007. Quality of life in the economic and urban economic literature. *Social Indicators Research* 84, 1 (2007), 1–25.

[30] Nishtha Langer, Chris Forman, Sunder Kekre, and Baohong Sun. 2012. Ushering buyers into electronic channels: An empirical analysis. *Information Systems Research* 23, 4 (2012), 1212–1231.

[31] Zhenhui Li, Fei Wu, and Margaret C. Crofoot. 2013. Mining following relationships in movement data. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining*. IEEE, 458–467.

[32] Janne Lindqvist, Justin Cranshaw, Jason Wiese, Jason Hong, and John Zimmerman. 2011. I'm the mayor of my house: Examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2409–2418.

[33] Todd Litman. 2003. Economic value of walkability. *Transportation Research Record: Journal of the Transportation Research Board* 1828 (2003), 3–11.

[34] Xianghua Lu, Sulin Ba, Lihua Huang, and Yue Feng. 2013. Promotional marketing or word-of-mouth? Evidence from online restaurant reviews. *Information Systems Research* 24, 3 (2013), 596–612.

[35] Eugene J. McCann. 2004. 'Best places': Interurban competition, quality of life and popular media discourse. *Urban Studies* 41, 10 (2004), 1909–1929.

[36] Aviv Nevo. 2001. Measuring market power in the ready-to-eat cereal industry. *Econometrica* 69, 2 (2001), 307–342.

[37] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. 2012. Mining user mobility features for next place prediction in location-based services. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM)*. IEEE, 1038–1043.

[38] Deborah N. Peikes, Lorenzo Moreno, and Sean Michael Orzol. 2008. Propensity score matching. *The American Statistician* 62, 3 (2008), 222–231.

[39] Olga Perdikaki, Saravanan Kesavan, and Jayashankar M. Swaminathan. 2012. Effect of traffic on sales and conversion rates of retail stores. *Manufacturing & Service Operations Management* 14, 1 (2012), 145–162.

[40] Jennifer Roback. 1982. Wages, rents, and the quality of life. *The Journal of Political Economy* (1982), 1257–1278.

[41] Daniel M. Romero, Chenhao Tan, and Johan Ugander. 2013. On the interplay between social and topical structure. In *Proceedings of ICWSM.*

[42] Sherwin Rosen. 1979. Wage-based indexes of urban quality of life. *Current Issues in Urban Economics* 3 (1979), 324–345.

[43] Andrew Rundle, Kathryn M. Neckerman, Lance Freeman, Gina S. Lovasi, Marnie Purciel, James Quinn, Catherine Richards, Neelanjan Sircar, Christopher Weiss, and others. 2009. Neighborhood food environment and walkability predict obesity in New York City. *Environmental Health Perspectives* 117, 3 (2009), 442–447.

[44] Brian E. Saelens, James F. Sallis, and Lawrence D. Frank. 2003. Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures. *Annals of Behavioral Medicine* 25, 2 (2003), 80–91.

[45] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*. 851–860.

[46] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. 2011. Socio-spatial properties of online location-based social networks. *ICWSM* 11 (2011), 329–336.

[47] Mark Edward Stover and Charles L. Leven. 1992. Methodological issues in the determination of the quality of life in urban areas. *Urban Studies* 29, 5 (1992), 737–754.

[48] Catherine Tucker and Juanjuan Zhang. 2011. How does popularity information affect choices? A field experiment. *Management Science* 57, 5 (2011), 828–842.

[49] J. Miguel Villas-Boas and Russell S. Winer. 1999. Endogeneity in brand choice models. *Management Science* 45, 10 (1999), 1324–1338.

[50] Xun Wang, Feida Zhu, Jing Jiang, and Sujian Li. 2013. Real time event detection in twitter. In *International Conference on Web-Age Information Management*. Springer, 502–513.

[51] Asha Weinstein Agrawal, Marc Schlossberg, and Katja Irvin. 2008. How far, by which route and why? A spatial analysis of pedestrian preference. *Journal of Urban Design* 13, 1 (2008), 81–98.

[52] Inge M. Wetzer, Marcel Zeelenberg, and Rik Pieters. 2007. Never eat in that restaurant, I did! Exploring why people engage in negative word-of-mouth communication. *Psychology & Marketing* 24, 8 (2007), 661–680.

[53] Chao Zhang, Guangyu Zhou, Quan Yuan, Honglei Zhuang, Yu Zheng, Lance Kaplan, Shaowen Wang, and Jiawei Han. 2016. GeoBurst: Real-time local event detection in geo-tagged tweet streams. SIGIR 2016.

[54] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 38.

[55] Yu Zheng, Huichu Zhang, and Yong Yu. 2015. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2.

Received November 2016; revised February 2017; accepted March 2017