Understanding User Economic Behavior in the City Using Large-scale Geotagged and Crowdsourced Data

Yingjie Zhang
Heinz College
Carnegie Mellon University
Pittsburgh, PA 15213, USA
yingjie@cmu.edu

Beibei Li
Heinz College
Carnegie Mellon University
Pittsburgh, PA 15213, USA
beibeili@andrew.cmu.edu

Jason Hong
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
jasonh@cs.cmu.edu

ABSTRACT

The pervasiveness of mobile technologies today have facilitated the creation of massive 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 infer various patterns of human behavior, which help us understand the interactions between humans and cities. In this study, we focus on understanding users economic behavior in the city by examining the economic value from crowdsourced and geotaggged data. Specifically, we extract multiple traffic and human mobility features from publicly available data sources using *NLP* and geo-mapping techniques, and examine the effects of both static and dynamic features on economic outcome of local businesses. 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 more street closure 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 of how 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.

Keywords

Geotagged Social Media, Crowdsourced User Behavior, Econometrics, Location-Based Service, Economic Analysis, City Demand, Mobility Analytics, NLP

1. INTRODUCTION

Rapid urbanization is imposing various urban challenges, especially increased demand on the city infrastructures and on the quality of services. These challenges call for a specific focus on urban systems and their interaction 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 [3]. A decrease in foot traffic can have significantly negative impact on store sales. 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 the 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 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, as it includes many factors, such as user walkability, street connectivity, traffic conditions, and other urban amenities. These multi-dimensional characteristics make it 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. This dynamic nature makes it highly unpredictable with regard to its economic impact on human behaviors. In this research, we extract this information by applying NLP and geo-mapping techniques on large-scale data from Twitter and Foursquare. Using geotagged user-generated data created via mobile and location-based services and crowdsourcing channels, we are able to extract the 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. Little work has been done to examine from a social and economic perspective of such data to infer relationship between humans and cities. This is the main focus of our paper. 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 economic analyses, our research goals are two-fold:

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883066.

- Extract spatial and socioeconomic features of cities from geotagged and crowdsourced data at large scale;
- Apply econometric models to quantify the causal effects of different features on the economic outcome of human behavior towards 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 neighborhood 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 the digitized and crowdsourced user behavior are informative in inferring local demand. The results show a significant positive impact of human foot traffic on local businesses, and significant negative effects due to traffic. Specifically, 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%. We also find that, on average, one 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 quantify the economic impact of not only static features but also dynamic features of users' digitized and crowdsourced behavior on local businesses. 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 analytic from social media. Our results can also help facilitate better policy decision-making about proactive city planning and improve the sustainability of urban neighborhoods. Finally, our work also offers an opportunity for incorporating an economic lens into location-based services and geo-mapping services, which could help improve our understanding of local areas, as well as local search and local advertising.

2. RELATED WORK

We initiate our research focus on two questions: (i) How can we efficiently extract features that would potentially affect local demand from various available resources, including social media channels and official sources? (ii) How can we use these characteristics to evaluate the value of information from an economic perspective? To examine these questions, our paper draws from multiple streams of work.

Geotagged and Crowdsourced Data Analysis. With the growing volume of geographic datasets, more and more studies are attracted by the location-based services [16]. Previous studies used various methods to explore this emerging phenomenon from different perspectives, including usage patterns of location-sharing applications [5]; relationship between people [8]; and detection of real-time events [19]. These studies put various methods forward to evaluate the human mobility patterns. But most of those studies are exploratory analyses, answering what happen and how users behave in the real world. They didn't link their study to the economic values while such further-step analysis can benefit the economic development, or even the entire society.

Economic Values of Users' Behavior. Understanding the economic and social values is the main focus of researchers in marketing or economic related fields [7]. Due to the lack of data, they limit their studies in the online world. However, the microeconomics, especially the performance of small businesses, are largely affected by various location-specific factors. Merely relying on online sources is hard to gain a holistic picture to understand the business mechanism at micro level. Here, we utilize geotagged and crowdsourced data to study their economic values for small businesses.

Economics of Location and Urban System. In addition, our study is also closely related to the economics of location and urban system. This stream of research can be traced back to the 1970s [15]. Different studies used various indicators to detect the market price [2], best location [10], etc. However, the indicators they used to evaluate the 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 market [11].

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: 6pm, 6:30pm, 7pm, 7:30pm, 8pm and 8:30pm (peak dining hours). In total, we have 312,326 data points.

3.1.2 Geotagged and Crowdsourced Data

The local demand is largely affected by the social and economic factors in their neighborhoods. To extract those factors, we collected crowdsourced and geotagged data based on three publicly available sources:

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 [16]. We have approximately 380,000 mobile user check-ins generated within a 30 miles radius from the center of Manhattan. We used geo-coding tools to extract the geographical location (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 geo-coding 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. 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, etc. 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 above datasets, we geotagged all data using 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 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 walkable distance [4].

Restaurant Characteristics Data. Previous studies show that online word-of-mouth does affect restaurants sales becaue restaurants' quality and popularity can be inferred from such crowdsourced information [18]. Besides, restaurants' inherent characteristics also affect customers' choices and the restaurants' profits. To capture those factors, we obtained the restaurants' 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 geo-tags.

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 zipcode level from the US Census website (factfinder2.census.gov/), and recorded the average temperature and daily precipitation during the same time period 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 [14], we evaluate it as a vector with four values: location density, population density, heterogeneity and competitiveness. Formally, a restaurant i has its static spatial features:

$$\text{STATIC\_SPA}_i = \begin{Bmatrix} \text{LOC\_DENSITY}_i, \text{HETEROGENEITY}_i, \\ \text{POP\_DENSITY}_i, \text{COMPETITIVENESS}_i \end{Bmatrix}. \quad (1)$$

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 :

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

Heterogeneity: Similar to the ideas in [14] we use the entropy measurement to assess the level of spatial heterogeneity of an area. Entropy is defined as the expected amount of the information from certain events. We apply it into 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,

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

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

Competitiveness: Given the 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 [9]. Walking is the most common leisure-time physical activity in the US and has been found to have various economic benefits, including urban neighborhood accessibility, increased efficiency of land use and improved urban livability [17]. In this study, we use Foursquare check-in data to measure this human mobility feature (NEIGH_WALK) by tracking both spatial and

Table 1: De nition 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
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.

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-ins 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 that restaurant i's human mobility features as follows:

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

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-ins 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:

$$\text{SOC\_STABLITY}_i = \sum_{u \in U} \left| \left\{ \begin{matrix} (p_m,t_m,p_n,t_n) \in C_u | p_m, \\ p_n \in d(i,l), t_m, t_n \ \ in T \end{matrix} \right\} \right| \quad (7)$$

Incoming Mobility: One way to show the popularity of a neighborhood is that it attracts people from other neighborhoods can be attracted 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:

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

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, etc. Such street closure lowers the accessibility of the neighborhood. In our model, we use user-generated content from Twitter to extract the dynamic traffic conditions.

3.2.4 Street Closure (Event, Construction) Features

In addition to traffic emergencies as described above, 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 and the time length of these closures.

3.2.5 Restaurant-Specific Features

In addition to the above factors, restaurant-level heterogeneity has non-ignorable effects on the 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:

$$\text{REST\_SPE}_i = \{\text{PRICE}_i; \text{RATING}_i; \text{NUMOFREVIEW}_i\}; \quad (9)$$

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 above 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 1.

4. ECONOMETRIC MODELING

Econometrics is a well-established statistics technique to test hypotheses and to predict future changes using economic model. In this paper, our econometric model aims to quantify the causal effects of different features on the economic outcome of human behavior towards local businesses.

4.1 Panel Data Analysis

Based on our time-series dataset, we use a fixed-effect panel model to estimate the impact of different factors in an urban neighborhood on the restaurant bookings. Our main model can be formalized in the following equation:

$$\begin{aligned}\Pr(\text{FULL})_{it} = {} & {}_i + \text{STATIC\_SPA}_i \cdot T_t \cdot {}_1 + \text{HUMAN\_MOB}_{it} \cdot {}_2 \\ & + \text{TRA\_EFF}_{it} \cdot {}_3 + \text{STREET\_CLO}_{it} \cdot {}_4 + \text{REST\_SPE}_{it} \cdot {}_5 \\ & + \text{Controls}_{it} \cdot {} + T_t + {}_{it};\end{aligned} \quad (10)$$

where $\Pr(\text{FULL})_{it}$ 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 before: static spatial feature (STATIC_SPA_i), human mobility feature (HUMAN_MOB_{it}), traffic efficiency feature (TRA_EFF_{it}), street closure feature (STREET_CLO_{it}) and restaurant- specific feature (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 (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 $\text{STATIC\_SPA}_i \cdot T_t$ varies in different time periods, and then the effects of static features in different T can be estimated.

The variable Controls_{it} indicates all possible controls: an interesting thing to note is that our dataset covers the 2013 Christmas and New Year holidays. Furthermore, 2013 winter was much colder than usual along in the northeast coast of the US. 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. Moreover, a restaurant's bookings can be affected by its local advertising and marketing efforts. To account for these, we collected additional data on restaurants' 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.

4.2 Identification

However, to establish a causal relationship between local demand and all those above features of interest, we need to rule out reverse causal explanations and unobservable variables that can cause both the performance outcome and features. In economics, this critical issue is called *endogeneity*. This section discusses three potential types of endogeneity: (i) price endogeneity; (ii) potential endogeneity in traffic and human mobility characteristics; and (iii) selection bias in street events and constructions.

4.2.1 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 this, we apply Instrumental Variable (IV) method. The basic intuition of IV methods is to find alternative variables to substitute for the endogeneous variable in the model, where the IVs should be only correlated with the endogeneous variable (i.e., price) but uncorrelated with the unobserved error term in the model. Here we apply two commonly-used IV methods: Villas-Boas-Winer-style IVs [20] and Hausman-style IVs [13].

Villas-Bios-Winer-style IVs: Following [12], we use lagged prices as IVs with Google Trend data, which records the number of searches for each restaurant's name at monthly

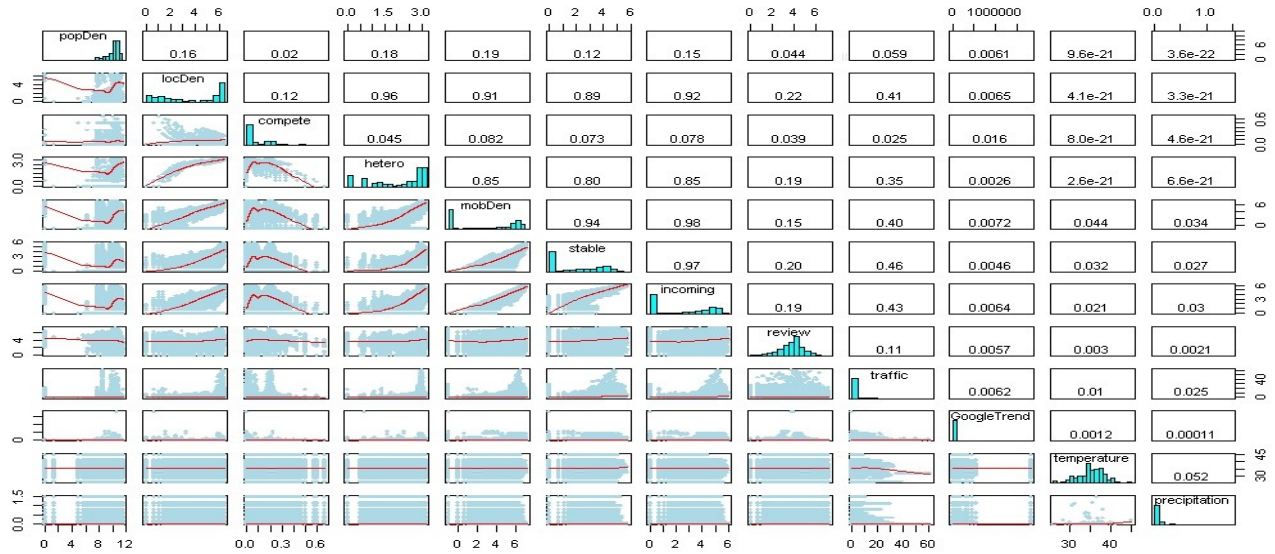

Figure 1: Data Correlograms. Diagonal: Histograms for the continuous variables in the dataset (population, location density, competitiveness, heterogeneity, mobility density, social stability, incoming mobility, review, tra c e ciency, temperature, precipitation). Upper-right: correlations of variable pairs. Bottom-left: scatterplots for joint distributions of variable pairs.

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 the market demand in the short run. Therefore, we can use the lagged 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. 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 the current demand. However, common demand shock 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 [12], we control for restaurant-specific time trend using Google Trend data to alleviate such concerns.

Hausman-style IVs: As discussed in [12], the idea is to use the average price of other similar restaurants (i.e., with the same star ratings or same cuisine type) in the other markets (i.e., neighborhoods). The intuition is that the prices of similar restaurants are correlated with respect to the 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 unobservable factors.

4.2.2 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 above 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 *IV*s 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 the 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 the 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 at consumers with similar tastes, demographics and consumption levels, which, to a large extent, indicate the local development condition of a neighborhood.

4.2.3 Selection Bias in Street Closures Features.

The potential selection bias in street events and street construction is another challenge in studying the economic outcome of human behavior. Specifically, in the context of street closure, the selection bias can be caused by unobserved factors. For example, the reason that the city plan-

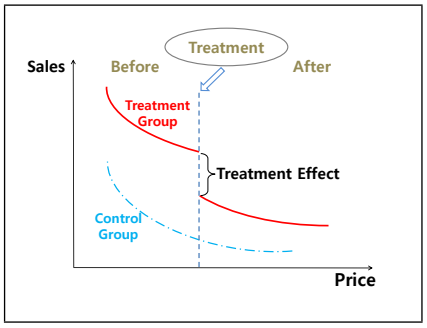

Figure 2: Framework of exploring causal treatment e ects using Di erence-in-Di erence method

ner chooses a particular street to closefor a local event or for construction may be due to some unobserved functional inability of that street (e.g., poor street condition, locational 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) and Difference-in-Difference (DID) methods to examine the causal effect of street closure. We illustrate the basic intuition of our analysis design in Figure 2.

First, we consider a four-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[1]. 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 four-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 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

$$logit_{it} = \ _i + \text{STATIC\_SPA}_i \cdot T_t \cdot \ _1 + \text{HUMAN\_MOB}_{it} \cdot \ _2 + \text{TRA\_EFF}_{it} \cdot \ _3 + \text{STREET\_CLO}_{it} \cdot \ _4 + \text{REST\_SPE}_{it} \cdot \ _5 + \ _{it}: \tag{12}$$

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 (1) we have a large number of sample observations, and (2) 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.

Finally, based on the matched samples, we use the Difference-in-Difference (DID) method to test the causality. 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 unobservables), we apply DID to exploit the exogenous variance in street closure across restaurants and time as the basis for identifying causal effects on local restaurant sales. Our model is as follows,

$$Pr(FULL)_{it} = \ _i + \ _1 Test_t + \ _2 Test_t \times Treat_i + Controls_{it} \cdot \ + T_t + \ _{it}: \tag{13}$$

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, etc. The coefficient of interest is β_2, which captures the effects of street closure in the test period.

5. RESULTS

5.1 Panel Data Model Results

We first start with our main estimation model (Equation 7), 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)[2], January (m_2), February (m_3) and March (m_4). To avoid collinearity, we use only the first three indicators in the regression.

With regard to restaurant-specific features, consistent with theories[3], 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

[1]We selected the time period with the largest number of treated samples: from Dec 24, 2013 to Jan 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.

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

[3]Rating effect is not statistically significant. 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 non-significant coefficient. But we do observe that this effect is positive.

Table 2: Main Estimation Results

Variable	Coef.	Variable	Coef.
Mobility density	0.004**	Heterogeneity	0.008
	(0.002)	$\times m_1$	(0.008)
Social stability	-0.001	Competitive	-0.001
	(0.002)	$\times m_1$	(0.004)
Incoming	0.003	Location	-0.001
mobility	(0.002)	density$\times m_2$	(0.004)
Traffic efficiency	-0.005***	Population	0.005***
	(0.001)	$\times m_2$	(0.001)

Streetomisu(xent228261)1(0)1(0.001)(0)1a4)1(*)1(*)-27.655 fgEd)1(V0150)(001)1(0)3(4)1(*)1(*)-2-12 7ometma1(i)128(e)-1888(7(-1(.0)(0)1(5)

results. By using the Hausman test, we find that our fixed effects model performs better.

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 sub variables: ACCIDENT, DISABLED, DELAYS, HEAVYTRAFFIC, WEATHER and EVENTS (the detailed definitions are provided in Table 1).We find very similar trends for all factors. 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 the neighborhood size matters a lot in our estimation, we consider neighborhoods of different sizes. The result, shown in Figure 3, is 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.

5.4 Interaction Effects Results

In the previous process, we considered the 3,187 restaurants 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.

Interaction Model I: Interaction effects with price level indicator: First, to explore how effects of 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 eq. (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 4 illustrates the coefficients of each feature within each group.

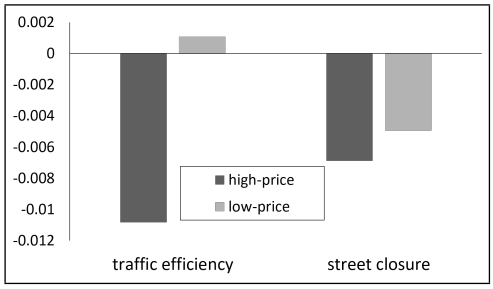

Figure 4: Comparison of interaction effects between traffic-related features and price levels

Interaction Model II: Interaction effects with chain or independent restaurants indicator: Next, in order to examine whether the 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 illustrates such differences.

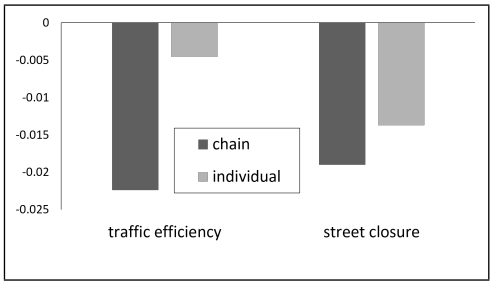

Figure 5: Comparison of interaction effects between traffic-related features and chain/independent indicator with 2,440 restaurants

Furthermore, we apply the above 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:

$$\begin{aligned}\Pr(\text{FULL})_{it} = {} & {}_i + {}_1\text{Test}_t + {}_2\text{Test}_t \times \text{Treat}_i \\ & + {}_3\text{Test}_t \times \text{Treat}_i \times \text{chain}_i + \text{Controls}_{it} \cdot + \boldsymbol{T}_t + {}_{it},\end{aligned} \tag{15}$$

where $chain_i$ is a dummy variable. Again, the lower-order interaction term $chain_i$ is excluded because it is collinear with the fixed effects. The results show that both β_2 ($= -0.053$) and β_3 ($= -0.4214$) are significant and negative, 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 for 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. DISCUSSION AND FUTURE WORK

In this paper, we explore to the economic values in the urban system based on geotagged and crowdsourced data from various large-scale social media sites and publicly available data sources. Using geo-mapping and geo-social-tagging techniques, together with natural language processing, we identify four feature dimensions to describe 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 paper 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 of how we can best make use of the 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 of its neighborhood city 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 the 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. Also, in order to better predict the 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, healthcare, etc. Such research would help us draw a comprehensive picture of the overall urban system and to study the economic dynamics and social interactions more precisely.

7. REFERENCES

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

[2] G. C. Blomquist, M. C. Berger, and J. P. Hoehn. New estimates of quality of life in urban areas. *The American Economic Review*, pages 89–107, 1988.

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

[4] P. Calthorpe. *The next American metropolis: Ecology, community, and the American dream*. Princeton Architectural Press, 1993.

[5] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. *ICWSM*, 2011:81–88, 2011.

[6] P. C. Cheshire and S. Magrini. Population growth in european cities: weather matters–but only nationally. *Regional studies*, 40(1):23–37, 2006.

[7] J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354, 2006.

[8] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD*, pages 1082–1090. ACM, 2011.

[9] R. Ewing and S. Handy. Measuring the unmeasurable: urban design qualities related to walkability. *Journal of Urban design*, 14(1):65–84, 2009.

[10] R. Florida. The economic geography of talent. *Annals of the Association of American geographers*, 92(4):743–755, 2002.

[11] C. Forman, A. Goldfarb, and S. Greenstein. How did location affect adoption of the commercial internet? global village vs. urban leadership. *Journal of urban Economics*, 58(3):389–420, 2005.

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

[13] J. A. Hausman. Valuation of new goods under perfect and imperfect competition. In *The economics of new goods*, pages 207–248. 1996.

[14] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD*, pages 793–801. ACM, 2013.

[15] D. Lambiri, B. Biagi, and V. Royuela. Quality of life in the economic and urban economic literature. *Social Indicators Research*, 84(1):1–25, 2007.

[16] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. 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*, pages 2409–2418. ACM, 2011.

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

[18] X. Lu, S. Ba, L. Huang, and Y. Feng. Promotional marketing or word-of-mouth? evidence from online restaurant reviews. *Information Systems Research*, 24(3):596–612, 2013.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[20] J. M. Villas-Boas and R. S. Winer. Endogeneity in brand choice models. *Management Science*, 45(10):1324–1338, 1999.