The Journal of Transport and Land Use http://jtlu.org
Vol. 12 No. 1 [2019] pp. 723–740

# Using location-based social network data for activity intensity analysis: A case study of New York City

**Haluk Laman**
University of Central Florida
haluklaman@knights.ucf.edu

**Shamsunnahar Yasmin**
Queensland University of Technology
and University of Central Florida
shams.yasmin@qut.edu.au

**Naveen Eluru**
University of Central Florida
naveen.eluru@ucf.edu

**Abstract:** Location-based social networks (LBSN) are social media sites where users check-in at venues and share content linked to their geo-locations. LBSN, considered to be a novel data source, contain valuable information for urban planners and researchers. While earlier research efforts focused either on disaggregate patterns or aggregate analysis of social and temporal attributes, no attempt has been made to relate the data to transportation planning outcomes. To that extent, the current study employs LBSN service-based data for an aggregate-level transportation planning exercise by developing land-use planning models. Specifically, we employ check-in data aggregated at the census tract level to develop a quantitative model for activity intensity as a function of land use and built-environment attributes for the New York City (NYC) region. A statistical exercise based on clustering of census tracts and negative binomial regression analyses are adopted to analyze the aggregated data. We demonstrate the implications of the estimated models by presenting the spatial aggregation profiling based on the model estimates. The findings provide insights on relative differences of activity engagements across the urban region. The proposed approach thus provides a complementary analysis tool to traditional transportation planning exercises.

**Keywords:** Location-based social networks, land-use profiles, check-in, activity intensity, urban planning, clustering, negative binomial regression

**Article history:**
Received: October 3, 2018
Received in revised form:
June 11, 2019
Accepted: September 17, 2019
Available online: October 9, 2019

## 1 Introduction

Smartphone ownership among Americans has rapidly risen to 77% in 2018 from 35% in 2011 (Pew Research Center, 2017). The ubiquity of smartphones with an embedded Global Positioning System (GPS) allows for obtaining precise individual level location information. In fact, according to a recent report by the Pew Research Center (2017), 90% of smartphone users obtain directions, recommendations, and other location specific information from their phone. Several social networking sites (such as Twitter, Foursquare, Gowalla, and Facebook) allow users to share content on their websites with

http://dx.doi.org/10.5198/jtlu.2019.1470
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

geo-coded information often referred to as location based social networks (LBSN). These location-based services allow users to "check-in" at a venue (such as a restaurant or public park) based on their GPS coordinates providing them with location specific status update. While privacy concerns among users have ensured that the usage of location-based services is not universal, a large share of the population still adopts these services. For instance, 28% of American adults use a mobile or LBSN service. Furthermore, 12% of smartphone owners use their phone to check-in locations using the LBSN service. More interestingly, 7% of all adults allow the social media service they are using to automatically share their locations when they update their status. As is expected, the usage is higher among younger individuals - ages between 18-29 (16%), 30-49 (11%), 50-64 (9%) and 65+ (11%). These usage rates for LBSN clearly highlight the small share of adoption. However, given the large number of smartphone users, the data from these services would be larger than the data collected from traditional transportation data collection approaches (such as household surveys). Thus, it is not surprising that in recent years, several studies have explored the use of such LBSN based datasets acquired from websites for data mining, land-use planning, urban mobility analysis and transportation analysis (see Gordon & de Souza e Silva, 2011).

To be sure, the data available from LBSN services is not without limitations (as identified in Hasan & Ukkusuri, 2014). First, the data does not provide detailed information (on gender, age, education) at the individual level. Second, even avid LBSN service users are unlikely to "check-in" every event in their day (particularly the routine activities). The activity start and end times are also unlikely to be available. Finally, the sample of individuals providing the information represents a sample of the population that is unlikely to be representative of the broader population. In fact, recently, a study by Rzeszewski (2008) illustrated that the data obtained from are inherently different based on the user behavior and the social media platform adopted. The authors cautioned analysts considering merging data from multiple platforms. Given these inherent biases, the data obtained from such services are prone to bias at a disaggregate level. On the other hand, employing the data for aggregate level analysis might provide a more representative population behavior. For example, rather than focusing on an individual's activity locations, based on the LBSN data identifying the number of "check-ins" at a spatial unit such as census tract might offer relative differences of activity engagement across the urban region. More importantly, the traditional data collection methods (such as household travel surveys) provide sparse information on such activity engagement information. Thus, employing LBSN data check-ins to identify activity centers (based on attractiveness) across the urban regions will provide a complementary analysis to traditional land-use transportation planning exercises (Frias-Martinez, Soto, Hohwald, & Frias-Martinez, 2012).

Given the large number of LBSN users, the data available provide us with large-scale datasets for activity analysis. The LBSN users provide analysts with detailed spatio-temporal data that can be utilized for planning applications. The main objective of our study effort is to employ LBSN service-based data for aggregate level planning exercise by developing land-use transportation planning models. To elaborate, using activity check-ins within a spatial aggregation as a surrogate measure of attractiveness, the study provides a quantitative relationship between attractiveness and various socio-demographic, points of interest, transportation infrastructure, and land-use attributes. The established relationship will allow transportation and land-use planners to identify what factors affect zonal/destination attractiveness and pro-actively plan for potential new demand with changing socio-demographic, land-use, and/or transportation infrastructure patterns. Example of changes that can be analyzed include the development of mixed-use developments in dense neighborhoods or the addition of public transport infrastructure.

In our research effort, data from LBSN provider Foursquare that allows users to check-in at indoor and outdoor venues (such as café, restaurant or public spaces) via smartphones is utilized. The geocoded data is aggregated using Geographical Information System (GIS) techniques to obtain check-ins at a census tract level also referred to as "activity intensity" for the New York City (NYC) region. The relationship of the computed activity intensity variable with socio-demographics, land-use variables, transportation and infrastructure variables, and points of interests at the census tract level is analyzed to offer insights on the interconnectedness of activity intensity and other attributes.

A statistical exercise based on clustering and negative binomial regression analysis is adopted to analyze the aggregated data. The cluster analysis is performed in order to categorize the census tracts in the NYC region as a function of various exogenous variables. The clustering approach, rather than considering the entire city as homogenous allows us to distinguish across different clusters. Subsequently, cluster specific regression analysis is employed to identify the factors that affect the "check-ins" in the cluster. As the "check-ins" are non-negative integer values, negative binomial regression models were adopted for cluster specific regression models. The models estimated are employed to illustrate the impact of various parameters on check-ins using a hot spot analysis. Hence, we illustrate how the spatial distribution of activity patterns derived by LBSN data can be utilized to reveal urban activity patterns.

The remainder of the paper is organized as follows: Section 2 provides a review of earlier research and positions the current work in context. In Section 3, data source and description are provided. The research methods employed, model results, validation statistics, and hot spot analysis are presented in Section 4. Finally, Section 5 concludes the paper.

## 2 Earlier research and current study in context

The traditional research efforts examine activity travel patterns (and related choices) based on traditional household travel surveys. The literature in this context is quite vast and it is beyond the scope of the paper to document (see Pinjari & Bhat, 2011; Miller, 2014, for a detailed summary of earlier work). With the increasing adoption of mobile devices, there is growing research employing innovative data sources for transportation planning analysis. In this context, we present the review of earlier studies along two streams: (1) research employing social media data for non-transportation research context and (2) research employing social media data for transportation research contexts.

The first stream of studies has mainly originated in the fields of social sciences and computer science. The emphasis of these research efforts is to extract behavioral insights on online activity and offline interactions (Cranshaw, Hong, & Sadeh, 2012). Cheng, Caverlee, Lee, and Sui (2011) derived an algorithm to understand the mobility patterns of LBSN users. By studying several different metropolitan areas, user displacement, radius of gyration, and returning probability of individuals were determined. Their findings can be summarized as; LBSN users follow simple, reproducible patterns which refer to Levy Flight type patterns, social status is coupled to mobility, and content analysis can reveal hidden context between people and locations. More recently, Ahas et al. (2015) and Cao et al. (2015) employed mobile and/or location-based social network data to study temporal and spatial differences in urban regions from multiple countries. A hierarchical statistical approach - the nested Chinese Restaurant Franchise (nCRF) - based on tweet contents of LBSN data of the US was proposed by Ahmad, Hong, and Smola (2013) to infer a latent distribution of user locations. Kling and Pozdnoukhou (2012) also used topic modeling for investigating space-time dynamics of time-stamped and geo-located check-in information. Topic modeling was also employed to process urban activity patterns and classify them for LBSN data of NYC (Hasan & Ukkusuri, 2014).

The second stream of research has explored the viability of social media data for transportation planning purposes. Frias-Martinez et al. (2012) used an unsupervised Neural Networks technique named Self Organizing Maps (SOM) to Manhattan area of NYC. The study findings indicate that LBSN data can serve as a complementary source of information for urban planning development. Cranshaw et al. (2011) employed the fine spatial resolution based on geo-located tweets by clustering nearby locations with similar activities and revealing social-spatial divisions in Pittsburgh. Wakamiya, Lee, and

Sumiya (2011) used LBSN data of three cities in Japan (Osaka, Nagoya, Tokyo) to study the crowd and individual movements across geography by aggregate and dispersion models as well as the semantics of the tweet contents. They combined temporal analysis with k-means clustering based on the spatial check-ins and urban types by tracking common patterns in different regions. Their findings confirmed that crowd activities determined via Twitter could characterize living spaces in cities. Noulas, Scellato, Lambiotte, Pontil, and Mascolo (2012) used a rank based movement model by ranking transitions by distance in order to capture urban mobility pattern variations. A similar study using rank-based models was conducted aiming to determine how a large-scale geolocation data set can be analyzed to classify and refer to individual activity patterns (Cao et al., 2015). As a result of aggregate and disaggregate level analysis in New York, Chicago, and Los Angeles areas, the study concluded that people choose their destinations mainly based on the popularity of these places. Bawa-Cavia (2011) conducted inter urban analysis of Foursquare data in three metropolitan cities (NYC, London, and Paris) to understand the difference in spatial structures across these cities. Zhan, Ukkusuri, and Zhu (2014) deployed supervised (random forest algorithm) and unsupervised (k-means clustering) approaches to infer land use of NYC based on LBSN data. The findings confirm that LBSN data can be used as a complementary data source in land-use planning.

While a number of research studies have been conducted to analyze mobile or location-based social network data, the research is still in its infancy. The analysis has been focused either on disaggregate patterns or aggregate analysis of social and temporal attributes. While these efforts provide useful insights, linkages to transportation planning outcomes such as socio-demographics, land-use variables, transportation and infrastructure variables, and points of interests are poorly understood. The main objective of our proposed effort is to employ check-in data aggregated at the census tract level to develop a quantitative model for activity intensity as a function of socio-demographics, transportation infrastructure, land use, and built environment attributes. The study also recognizes that developing a single model for NYC would be restrictive and of limited use. Hence, prior to modeling, we classify the census tracts in NYC into four clusters as a function of land-use variables. Subsequently, for each cluster, a Negative Binomial Regression model is developed to study activity intensity across the city. The results from these models are employed to conduct a hot spot analysis highlighting the impact of independent variables across the urban region. The hot spot analysis illustrated how the data from LBSN users could assist planners in making informed decisions on mobility and infrastructure needs.

## 3 Data source and descriptive statistics

The original check-in dataset used in a previous study by Cheng et al. (2011) was employed in this research. The data consisted of 220,000 unique users checked-in at 1,200 venues from December 2011 to April 2012.[1] The data was obtained from Location Sharing Services (LSS) applications such as Foursquare, Twitter, TweetDeck, Gowalla. Up to 2,000 most recent geo-labeled tweets for each user were saved (for more details on the dataset format, see Cheng et al., 2011). Using GIS analysis procedures, the check-in data in the NYC region were selected. NYC's population in 2011 was 8.273 million with 2,166 census tract zones based on the zoning system of the US Census Bureau. The aggregated check-in counts were augmented with census tract characteristics, including socio-demographics, land-use characteristics, and points of interests. After the data processing, 624,595 geo-coded check-ins were considered for analysis. The check-ins in the census tract range from 0 through 11,159 with an average of about 288.

In our analysis, we generated a host of variables from four broad categories including: (1) land-use characteristics (such as one and two family buildings, multi-family walk-up buildings, multi-family

[1] The proposed prediction framework of activity check-ins can be employed by using LBSN data from other regions or for any other year if the data is available.

elevator buildings, residential buildings, commercial and office buildings, industrial and manufacturing, transportation and utility, public facilities and institutions, open space and outdoor recreation, and parking facilities), (2) socio-demographics (such as population density, population by age, gender, race, and household characteristics), (3) points of interest (locations such as leisure, tourism, recreational, library, airport, sidewalk café, health places), and (4) built environment (such as bus line, bus stops, subway stops, train stops, ferry landing, park and ride stations, bike route, street centerline, school counts, building footprint, and green area). We have extracted the abovementioned variables at the Census Tract level for the year 2010 to reflect the available LBSN data. A descriptive summary of the characteristics generated for our analysis is presented in Table 1.

**Table 1.** Descriptive statistics of NYC census tracts

| **Variables Name** | **Definition** | **Zonal** | | |
|---|---|---|---|---|
| | | **Minimum** | **Maximum** | **Average** |
| **Dependent variable** | | | | |
| Check-in Counts per CT* | Total number of check-ins per CT | 0 | 11159 | 288.41 |
| **Socio-Demographic Characteristics** | | | | |
| Median Age | Median CT Age / 10 | 0 | 8.45 | 3.57 |
| Caucasian Proportion | Caucasian Population of CT / Total Population of CT | 0 | 1.00 | 0.43 |
| African – American Proportion | African-American population of CT / Total population of CT | 0 | 1.00 | 0.27 |
| Hispanic Proportion | Hispanic population of CT / Total population of CT | 0 | 1.00 | 0.26 |
| Asian Proportion | Asian population of CT / Total population of CT | 0 | 1.00 | 0.12 |
| Children in HH× | Total number of children of CT / Total number of HH of CT | 0 | 0.64 | 0.29 |
| Family HH | Total number of family HH of CT / Total number HH of CT | 0 | 1.00 | 0.64 |
| Average Family Size | Average family size of CT | 0 | 6.09 | 3.27 |
| Rental Vacancy Rate (%) | Rental vacancy units*100 / Total number of units within CT | 0 | 61.20 | 4.75 |
| CT Area | CT area in acres | 0.0134 | 4502.98 | 89.34 |
| Total Population | Total population per CT | 0 | 26588 | 3778.70 |
| **Points of Interest Characteristics** | | | | |
| Automotive | Number of Automotive Related Places per CT | 0 | 7 | 0.04 |
| Government | Number of Governmental Places per CT | 0 | 75 | 4.50 |
| Leisure | Number of Leisure Points per CT | 0 | 13 | 0.72 |
| Tourism | Number of Touristic Places per CT | 0 | 16 | 0.07 |
| Health | Number of Health-Related Places per CT | 0 | 8 | 0.12 |
| Library | Number of Libraries per CT | 0 | 2 | 0.10 |
| Nursing | Number of Nursing Places per CT / 10 | 0 | 12 | 0.56 |
| Senior | Number of Senior places per CT | 0 | 2 | 0.12 |
| Airport | Number of Airports per CT | 0 | 1 | 0.00 |
| Ferry Landing | Number of Ferry Landing per CT | 0 | 4 | 0.02 |
| Beach, Garden, Natural state parks | Number of beaches, gardens, natural state parks per CT /102 | 0 | 422 | 2.59 |

| | | | | |
|---|---|---|---|---|
| Subway Yard | Number of subway yards per CT | 0 | 2 | 0.02 |
| Food Places | Number of food aid related places per CT | 0 | 10 | 0.52 |
| Other Transportation Fac. | Number of other transportation facilities per CT / 10 | 0 | 17 | 0.09 |
| Waste Management Fac. | Number of waste management facilities per CT / 10 | 0 | 47 | 0.24 |
| Children Daycare | Number of children daycare points per CT / 10 | 0 | 11 | 1.19 |
| Bus Depot | Number of bus depots per CT | 0 | 1 | 0.01 |
| Recreation, Plaza, Mall | Number of recreation areas, plazas, malls per CT / 10 | 0 | 12 | 0.51 |
| Sidewalk Café | Number of sidewalk café per CT / 103 | 0 | 13.60 | 0.93 |
| **Transportation Infrastructure Characteristics** | | | | |
| Street Centerline | Total length of street center lines in ft in CT / 104 | 0 | 90.01 | 5.38 |
| Bus line | Total length of bus lines in ft in CT / 103 | 0 | 14.47 | 1.07 |
| Building Footprint | Total area of building footprints in CT / 107 | 0 | 0.87 | 0.29 |
| Building Elevation | Total number of building floors in CT / 103 | 0 | 6.08 | 1.19 |
| Green Area Density | Total green area by CT area in sq-ft / 103 | 0 | 6597.24 | 3.60 |
| Railroad | Total length of railroads in ft in CT / 105 | 0 | 22.92 | 0.36 |
| Bike Route | Total length of bike routes in ft in CT / 104 | 0 | 6.09 | 0.20 |
| Number of Buildings | Total number of buildings in CT / 103 | 0 | 3.25 | 0.49 |
| Bus Stop | Total number of bus stops in CT / 10 | 0 | 6 | 0.61 |
| Subway Entrances | Total number of subway entrances in CT / 10 | 0 | 3.60 | 0.08 |
| Subway Stops | Total number of Subway stops in CT | 0 | 6 | 0.22 |
| **Land-Use Characteristics** | | | | |
| One and Two Family Buildings Density | Total one and two family building lands by CT area in sq-ft / 10 | 0 | 6.79 | 1.78 |
| Multi-Family Walk-Up Buildings Density | Total multi-family walk-up buildings lands by CT area in sq-ft / 10 | 0 | 5.42 | 0.96 |
| Multi-Family Elevator Buildings Density | Total multi-family elevator buildings lands by CT area in sq-ft / 10 | 0 | 11.86 | 0.72 |
| Residential and Comm. Buildings Density | Total residential and commercial lands by CT area in sq-ft / 10 | 0 | 8.64 | 0.51 |
| Commercial and Office Buildings Density | Total commercial and office buildings lands by CT area in sq-ft / 10 | 0 | 6.12 | 0.39 |
| Industrial and Manufacturing Density | Total industrial and manufacturing lands by CT area in sq-ft / 10 | 0 | 5.90 | 0.19 |
| Transportation and Utility Density | Total transportation and utility lands by CT area in sq-ft / 10 | 0 | 8.92 | 0.16 |
| Public Facilities and Institutions Density | Total public facilities and institution lands by CT area in sq-ft / 10 | 0 | 9.10 | 0.57 |
| Open Space and Outdoor Recreation Density | Total open space and outdoor recreation lands by CT area in sq-ft / 10 | 0 | 27.37 | 0.39 |
| Parking Facilities Density | Total parking facility lands by CT area in sq-ft / 10 | 0 | 1.58 | 0.10 |

**CT = Census Tract*

*$^{\times}$HH = Household*

# 4 Empirical analysis

## 4.1 Cluster analysis

Clustering is a widely used statistical analysis tool to categorize items together based on their similarities or dissimilarities (Anderberg, 2014). The aim of the clustering algorithm is to classify the population into k categories based on a multivariate set of exogenous variables. The clusters generated should be internally homogenous while being heterogeneous relative to other clusters (Karlaftis & Tarko, 1998). k-means clustering is a common and straightforward model that uses minimum Euclidean distance between observations (see Pinjari, Eluru, Bhat, Pendyala, & Spissu, 2008, for similar examples).

Based on the host of exogenous variables, a cluster analysis is conducted to categorize the census tracts. Specifically, eight land-use characteristics were employed to undertake the clustering exercise. These characteristics are: one- and two-family buildings, multi-family walk-up buildings, multi-family elevator buildings, mixed residential and commercial buildings, commercial and office buildings, industrial and manufacturing, transportation and utility, public facilities and institutions. The k-means clustering algorithm provided a good fit for a 4-cluster classification. Also, the Bonferroni post-hoc test was used to validate the results of k-means clustering and multiple pairwise comparisons obtained with over 75% of each cluster was found to be statistically significant. The characteristics of the final clusters obtained are presented in Table 2. The spatial distribution of census tracts identified with cluster analysis results is illustrated in Figure 1. Cluster 1 consists of census tracts with single family and/or townhouses. Cluster 2 is represented by a mix of public facilities, commercial buildings, and offices, transportation as well as elevated/high rise buildings. Cluster 3 is composed of low rise residential and commercial buildings. Finally, as can be seen from the figure, a large portion of Cluster 4 are census tracts surrounding central park in Manhattan area. This cluster is predominantly covered by high rise buildings in addition to public facilities and institutions. Based on the characteristics, the clusters are labeled as follows: Cluster 1 – Low-rise residential, Cluster 2 – Public facilities and Commercial, Cluster 3 – Low-rise residential and commercial, Cluster 4 – High-rise residential and public facilities (see Figure 1).

**Table 2.** Final cluster centers

| **Land-Use Variables** | **Cluster 1 (Low Rise Residential)** | **Cluster 2 (Public Facilities and Commercial)** | **Cluster 3 (Low Rise Residential and Commercial)** | **Cluster 4 (High Rise Residential and Public Facilities)** |
|---|---|---|---|---|
| One and Two Family Buildings | 4.01 | 0.50 | 1.65 | 0.43 |
| Multi-Family Walk-Up Buildings | 0.47 | 0.54 | 1.95 | 0.68 |
| Multi-Family Elevator Buildings | 0.13 | 0.56 | 0.38 | 3.30 |
| Mixed Residential and Commercial Buildings | 0.17 | 0.72 | 0.53 | 0.72 |
| Commercial and Office Buildings | 0.21 | 0.70 | 0.29 | 0.31 |
| Industrial and Manufacturing | 0.08 | 0.44 | 0.11 | 0.06 |
| Transportation and Utility | 0.06 | 0.37 | 0.09 | 0.08 |
| Public Facilities and Institutions | 0.30 | 0.90 | 0.50 | 0.57 |
| Number of Census Tract Zones | 587 | 664 | 650 | 265 |

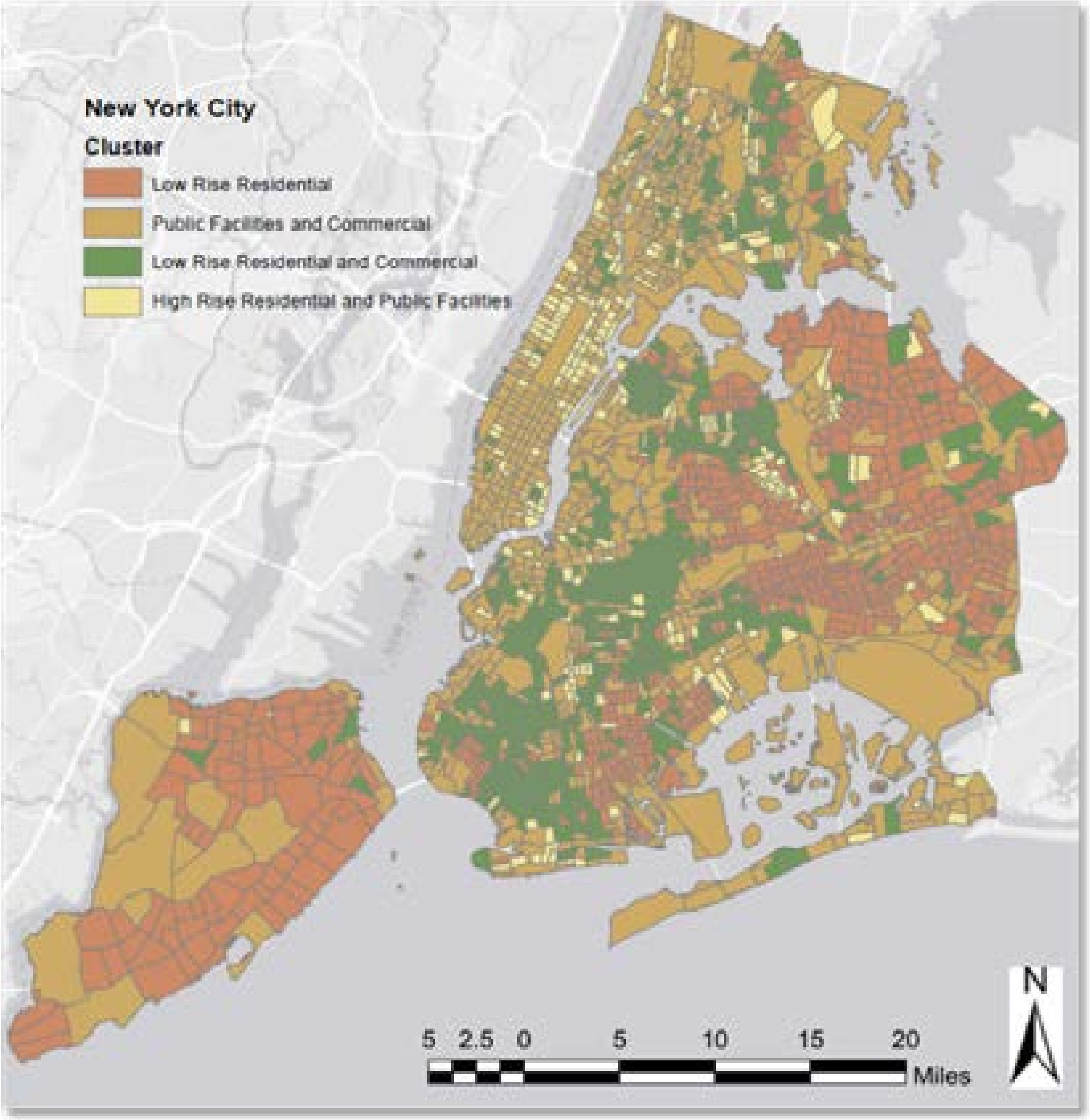

New York City
Cluster
Low Rise Residential
Public Facilities and Commercial
Low Rise Residential and Commercial
High Rise Residential and Public Facilities
5 2.5 0 5 10 15 20
Miles
N

**Figure 1.** Spatial distribution of clusters

### 4.2 Negative binomial (NB) regression model results

Given that activity intensity is represented based on non-negative integers, Negative Binomial (NB) regression approach is employed for our analysis. For the sake of brevity, details on the model formulation are not provided (see Winkelmann, 2008, for more details). For model estimation, two sets of models were estimated. First, a single NB model for New York City CT's (census tracts) was developed (pooled model). Second, NB models specific to each cluster (obtained above) were estimated (cluster-base model).

Prior to discussing the estimation results, we compare the performance of the pooled model and cluster-based models. The model performance was tested based on the computation of Bayesian Information Criterion (BIC) that penalizes the model with a large number of parameters. The BIC for a given empirical model is equal to: BIC= - 2LL + K ln(Q); where LL is the log-likelihood value at convergence, K is the number of parameters, and Q is the number of observations. The model with the lower BIC is the preferred model. The corresponding BIC values for pooled and cluster models are: 23376.2 and 23304.3, respectively. The comparison clearly illustrates the improved fit offered by the cluster specific NB models. For the sake of brevity, we restrict ourselves to the discussion of cluster-based NB models. The model estimation results for the cluster-based NB models are presented in Table 3.

### 4.2.1 Socio-demographic characteristics

Several sociodemographic characteristics influence the activity intensity at the census tract level including: median age by gender, the proportion of population by ethnicity, the average number of children at the household level, average family size, the proportion of family households within census tracts and percentage of rental vacancy rate.

Across the four clusters, the increase in median age has an overall negative effect. While median age coefficients by gender are positive (male median age for cluster 2 and 3 or female median age for cluster 4), the other median age coefficient is negative and slightly larger in magnitude. The result confirms the finding of Sloan, Morgan, Burnap, and Williams (2015) that increasing median age in the census tract reduces activity intensity. The proportion of population by ethnicity has varying trends across clusters. In cluster 1, a higher proportion of Caucasian and African-American increases activity intensity. On the other hand, for cluster 2, a higher proportion of African-American and Hispanic ethnicities are likely to reduce check-in activity. In cluster 3, the Hispanic proportion has a positive influence, while Caucasian proportion has a positive influence in cluster 4. On the other hand, according to the Pew Research Center (2017), overall statistics indicate that Hispanic and African-American social media users proportion are slightly higher than Caucasian users proportion. The findings provide evidence that the same variable can affect census tracts across the region differently. These trends could not have been captured using a pooled model. The increase in the average number of children at the household level, as expected, reduces the activity intensity across clusters with varying magnitudes. The presence of a higher proportion of family households reduces activity in Cluster 3. Correspondingly, a study effort by Do and Gatica-Perez (2013) observed that family households have lower average weekly visited places compared to other household types. The increase in average family size has a positive influence on activity intensity for Cluster 2. Finally, the rental vacancy rate has a negative influence on Cluster 3, probably because the increase in rental vacancy represents a lower occupancy rate resulting in a lower number of activities.

**Table 3.** Negative binomial regression results

| **Variable Names*** | **Cluster 1** | | **Cluster 2** | | **Cluster 3** | | **Cluster 4** | |
|---|---|---|---|---|---|---|---|---|
| | **Estimate** | **t-stat** | **Estimate** | **t-stat** | **Estimate** | **t-stat** | **Estimate** | **t-stat** |
| Constant | 6.32 | 5.65 | 6.12 | 21.17 | 8.20 | 13.10 | 8.97 | 12.13 |
| **Socio-Demographic Characteristics** | | | | | | | | |
| Median Age | -0.67 | -3.77 | -0.27 | -5.78 | -0.38 | -3.13 | -0.74 | -5.79 |
| Caucasian Proportion | 2.26 | 4.50 | - | - | - | - | 0.790 | 2.21 |
| African – American Proportion | 1.48 | 2.98 | -0.61 | -3.99 | - | - | - | - |
| Hispanic Proportion | - | - | -0.73 | -3.97 | 0.55 | 3.04 | - | - |
| Asian Proportion | 2.65 | 4.34 | - | - | - | - | - | - |
| Children in HH+ | -12.91 | -7.62 | -8.36 | -13.01 | -5.90 | -5.41 | -9.30 | -8.15 |
| Family HH | - | - | - | - | -4.01 | -6.29 | - | - |
| Avg. Family Size | - | - | 0.39 | 5.94 | - | - | - | - |
| Rental Vacancy Rate (%) | - | - | - | - | -0.05 | -3.63 | - | - |
| Points of Interest Characteristics | | | | | | | | |
| Automotive | - | - | - | - | - | - | 1.22 | 2.68 |
| Government | - | - | 0.03 | 5.43 | - | - | - | - |
| Leisure | - | - | 0.06 | 2.23 | - | - | 0.24 | 3.71 |
| Tourism | - | - | - | - | - | - | 0.87 | 3.20 |
| Health | - | - | - | - | -0.25 | -2.91 | - | - |
| Library | - | - | -0.27 | -2.49 | - | - | - | - |
| Nursing | - | - | - | - | 0.12 | 2.48 | - | - |
| Senior | 0.75 | 3.43 | - | - | - | - | - | - |
| Airport | - | - | 0.31 | 4.49 | - | - | - | - |
| Ferry Landing | - | - | - | - | - | - | 0.92 | 2.20 |
| State Parks, National and Cultural Inst. | - | - | - | - | - | - | -0.43 | -1.99 |
| Food Places | 0.47 | 4.92 | - | - | -0.09 | -2.64 | - | - |
| Other Transportation Fac. | - | - | - | - | 0.32 | 2.52 | -1.55 | -2.19 |
| Waste Management Fac. | - | - | - | - | -0.29 | -3.35 | 0.35 | 2.72 |
| Children Daycare | - | - | - | - | 0.04 | 1.69 | 0.08 | 2.03 |
| Bus Depot | - | - | - | - | - | - | -2.84 | -2.23 |
| Recreation, Plaza, Mall | 0.16 | 2.56 | 0.05 | 2.08 | - | - | - | - |
| Sidewalk Café | 0.18 | 1.82 | 0.08 | 2.00 | - | - | - | - |
| **Transportation Infrastructure Characteristics** | | | | | | | | |
| Street Centerline | 0.07 | 3.48 | 0.03 | 4.55 | 0.08 | 2.80 | - | - |
| Bus line | 0.08 | 2.14 | - | - | - | - | - | - |
| Railroad | - | - | 0.05 | 2.24 | - | - | 0.18 | 1.74 |
| Bike Route | - | - | 0.39 | 5.65 | - | - | - | - |
| Bus Stop | - | - | 0.12 | 1.86 | 0.18 | 1.69 | - | - |
| Subway Entrances | -0.99 | -1.76 | - | - | 1.45 | 4.72 | 1.44 | 3.21 |
| **Land-Use Characteristics** | | | | | | | | |
| One and Two Family Buildings Density | 0.12 | 2.58 | -0.61 | -7.84 | - | - | -0.61 | -6.28 |
| Multi-Family Walk-Up Build. Density | 0.242 | 2.05 | - | - | - | - | -0.17 | -1.81 |
| Residential and Comm. Build. Density | - | - | 0.129 | 2.74 | 0.516 | 4.52 | - | - |
| Commercial and Office Build. Density | 1.06 | 5.38 | 0.28 | 5.88 | 0.43 | 3.78 | - | - |
| Industrial and Manufacturing Density | 0.36 | 1.74 | -0.13 | -2. 60 | - | - | - | - |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Transportation and Utility Density | - | - | - | - | 0.67 | 3.73 | - | - |
| Building Footprint | - | - | 1.35 | 3.05 | - | - | - | - |
| Building Elevation | 0.18 | 2.29 | 0.17 | 2.35 | 1.31 | 4.55 | 0.93 | 6.06 |
| Green Area Density | - | - | - | - | - | - | -0.06 | -3.06 |
| Number of Buildings | - | - | - | - | -1.70 | -2.81 | - | - |
| Parking Facilities Density | - | - | 0.36 | 2.25 | -0.06 | -1.78 | - | - |
| Summary Statistics | | | | | | | | |
| Number of Census Tracts | 587 | | 664 | | 650 | | 265 | |
| Log-Likelihood | -2551.57 | | -4080.90 | | -3245.30 | | -1486.77 | |
| LR chi square (Number of Predictors) | 395.12 (17) | | 1151.98 (22) | | 793.80 (20) | | 323.38 (18) | |
| Pseudo R2 | 0.072 | | 0.12 | | 0.11 | | 0.09 | |

** Variable definitions are presented in Table 1*

*+ HH = Household*

### 4.2.2 Points of interest characteristics

Several points of interest characteristics influence the activity intensity observed. Cluster 1 is positively influenced by senior centers, food places, recreation plaza and malls, and sidewalk cafes. According to the venue cloud for check-ins generated by Cheng et al. (2011), it is clear that the largest clouds are cafés (i.e., coffee shop), food places, and centers (i.e., shopping malls). In Cluster 2, government-related, leisure related, airport recreation plaza and mall, and sidewalk cafes have a positive influence, while libraries have a negative influence. Li, Goodchild, and Xu (2013) indicated that a large proportion of users are likely to check-in at particular places such as airports. For cluster 3, the variable affecting activity intensity positively include nursing-related, other transportation facilities, children day-care, and sidewalk facilities. On the other hand, variables affecting negatively include health-related, food places, and waste management facilities. Finally, for Cluster 4, automotive related, leisure related, tourism-related, ferry landing, beach, garden, and natural facilities, waste management, children day-care show positive influence. Other transportation facilities and bus depot affect activity negatively in Cluster 4. Overall, the results capture the variation across the various clusters based on the points of interest. The results are hard to compare to earlier work because detailed information on this resolution has rarely been employed in transportation planning applications.

### 4.2.3 Transportation infrastructure

The impact of transportation infrastructure offers significant differences across the clusters. The street centerline length has a positive association with activity intensity in clusters 1 through 3. The bus line length in the census tract has a positive effect on cluster 1 activity intensity. In contrast, Sengstock and Gertz (2012) and Frias-Martinez and Frias-Martinez (2014) imply that national parks are highly associated with social media check-ins. The length of the railroad has a positive impact on activity intensity for cluster 2 and 4. The bicycle route length variable affects the intensity positively in cluster 2 only. The number of building variable in cluster 3 has a negative impact on activity intensity. The number of bus stops has a positive influence on cluster 3 ridership. Finally, the number of subway entrances has a negative influence on cluster 1 activity intensity while positively influencing activity intensity in clusters 3 and 4. According to the tweet content models developed for NYC, Kling and Pozdnoukhov (2012) indicated that transportation facilities are highly mentioned in the morning period while in Manhattan and East Village they were highly references in the evening period.

#### 4.2.4 Land-use characteristics

Land-use characteristics in the census tracts exhibit significant influence on activity intensity. A higher density of one and two-family buildings has a positive effect on activity intensity in cluster 1 while reducing activity intensity in clusters 2 and 4. The increase in density of multi-family walkup units has a positive effect on cluster 1 activity intensity. The residential and commercial density variable has a positive influence on activity intensity for cluster 2 and 3. Commercial and office building density is associated with positive influence for clusters 1 through 3. Similarly, Hu et al. (2015) indicate that commercial zones attract people's attention. Industrial and manufacturing density has a positive influence on activity intensity for cluster 1 and a negative influence on activity intensity for cluster 2. This finding might be affected by the fact highlighted by Frias-Martinez and Frias-Martinez (2014), that industrial land use is at a minimum in most regions of NYC (i.e., less than 8% in Manhattan). Transportation and utility density are positively associated with cluster 3 activity intensity. Building footprint significantly increases activity intensity for clusters 2. Building elevation increase is associated with higher activity intensity for all clusters (except 2). Interestingly, green area density is negatively associated with activity intensity in cluster 4. The building density of the area affects the type of businesses and accordingly affects the behavior of people visiting these areas (Cranshaw et al., 2012). Finally, parking facility density has a positive influence on check-in activity for cluster 2.

### 4.3 Model validation

To validate the model performance, we spatially represent (a) observed check-ins per unit area, (b) check-ins per unit area based on pooled model and (c) check-ins per unit area based on cluster-based models. The patterns of activity check-ins are presented in Figure 2. The categories considered for the three figures are: 0 - 0.05, 0.05 – 0.3, 0.3 – 0.5, 0.5 – 1, 1 – 3, 3 – 5, 5 – 10, 10 – 25, 25 – 50, 50 – 100, 100 – 200, 200 – 250, 250 and higher. From the visual comparison, across the three patterns, it is evident that the activity check-in patterns for cluster-based models are closer to the observed patterns. For instance, the pooled model over-predicts activity around central park and John F. Kennedy airport while the cluster-based models are closer to the observed patterns. To be sure, the cluster-based model also produces slightly different estimates for some census tracts. But overall, it offers more close resemblance to observed patterns.

### 4.4 Hot spot analysis

In this section, to illustrate the influence of exogenous variables, we undertake a unique hot spot analysis. The hot spot analysis is based on the value of the contribution of the individual parameter to the count propensity ($\beta^*x_n$). The contribution to count propensity is plotted by implementing Optimized Kernel Density tool of GIS ArcMap. The tool automatically aggregates the predicted check-in frequency, identifies an appropriate scale of analysis, and corrects for both multiple testing and spatial dependence by calculating the mean center of the input points using a radius search (bandwidth) algorithm. This tool allows us to identify statistically significant spatial groups of high values (hot spots) and low values (cold spots). Statistically significant hot and cold spots indicate that rather than a random pattern, the corresponding explanatory variable prediction exhibit statistically significant spatial dispersion. The variables chosen for the hot spot analysis are: Children by HH, One and Two-Family Buildings, Sidewalk Café, Median Age, Street Centerline, and Building Elevation. The spatial representations are presented in Figure 3. Light green background color indicates that both hot and cold spots exist on the figure, whereas blue background indicates that the heat map includes only hot or cold spots along with neutral areas. The results clearly illustrate the spatial regions that are significantly affected by these variables across the NYC region.

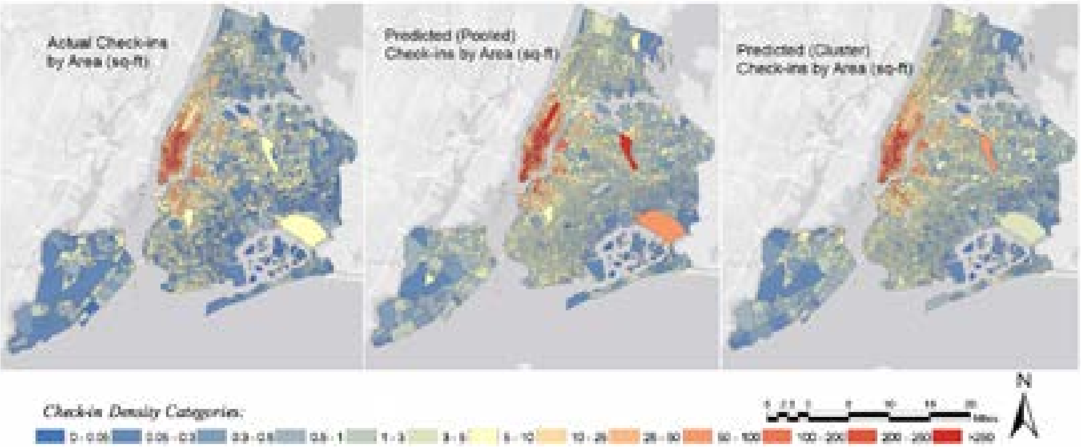
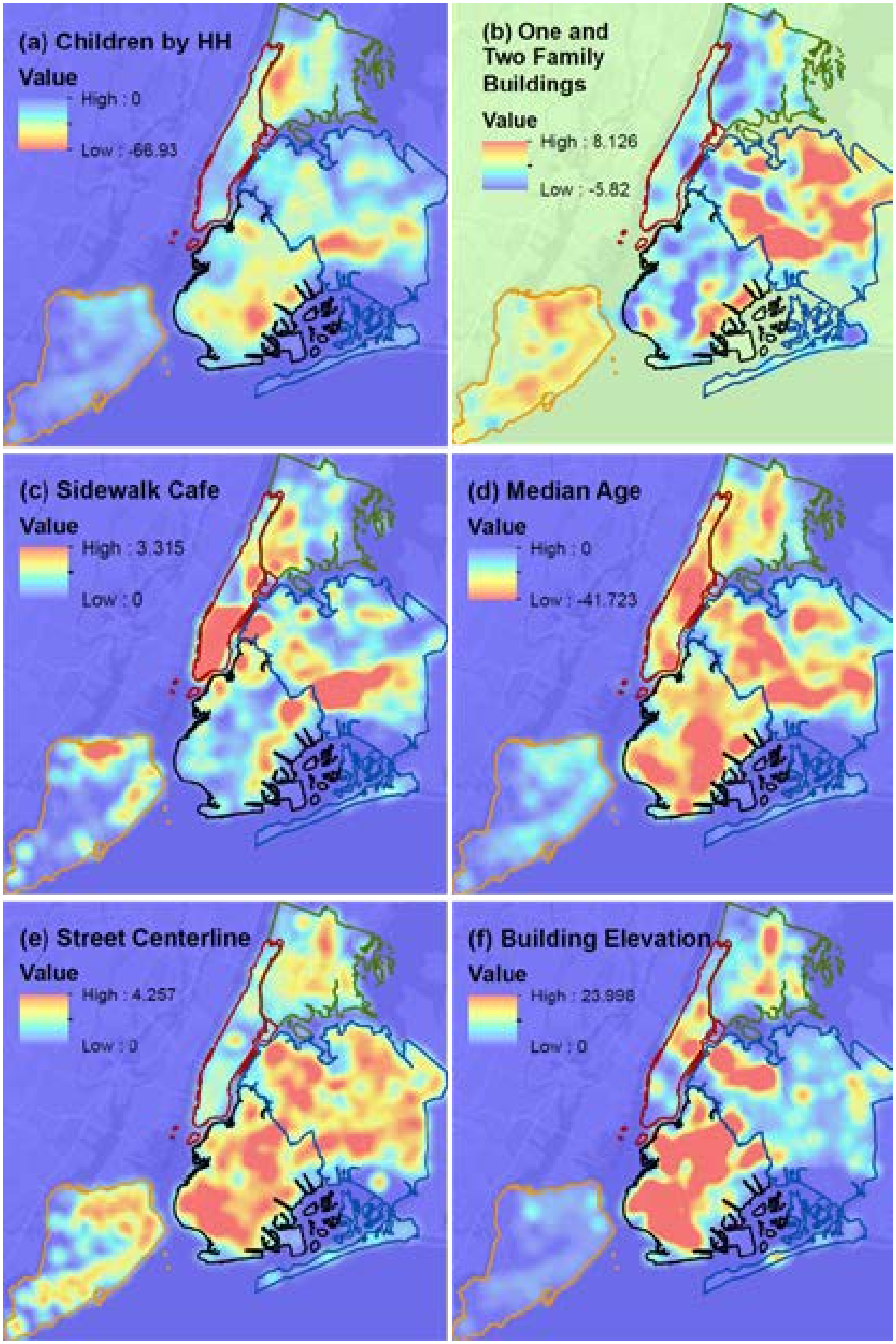

Actual Check-ins by Area (sq-ft)
Predicted (Pooled) Check-ins by Area (sq-ft)
Predicted (Cluster) Check-ins by Area (sq-ft)
Check-in Density Categories:
0 - 0.05
0.05 - 0.3
0.3 - 0.5
0.5 - 1
1 - 3
3 - 5
5 - 10
10 - 25
25 - 50
50 - 100
100 - 200
200 - 250
>250
0 1 2 4 8 12 16 20
Miles
N

**Figure 2.** New York City check-in density predictions


(a) Children by HH
Value
High : 0
Low : -68.93
(b) One and Two Family Buildings
Value
High : 8.126
Low : -5.82
(c) Sidewalk Cafe
Value
High : 3.315
Low : 0
(d) Median Age
Value
High : 0
Low : -41.723
(e) Street Centerline
Value
High : 4.257
Low : 0
(f) Building Elevation
Value
High : 23.998
Low : 0

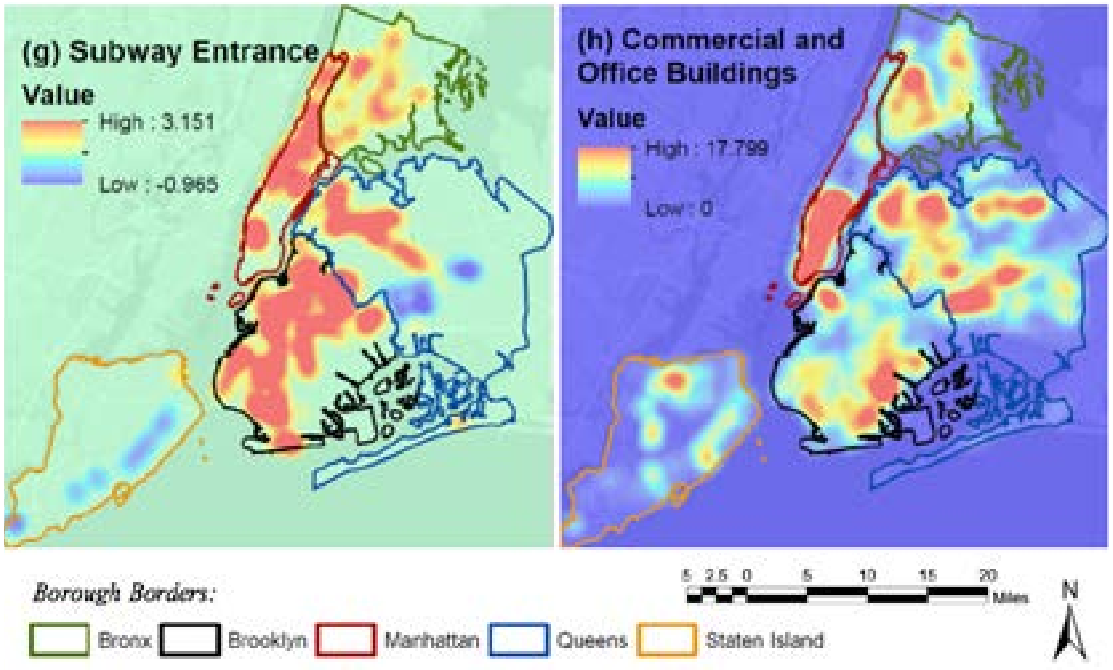


**Figure 3.** Kernel density estimate heat maps for selected exogenous variables; (a) children by HH, (b) one- and two-family buildings, (c) sidewalk café, (d) median age, (e) street centerline, (f) building elevation, (g) subway entrance, (h) commercial and office buildings

## 5 Conclusion

The current study employed a location-based social network (LBSN) service-based data for aggregate level transportation planning exercise by developing land-use planning models. Specifically, we employed check-in data aggregated at the census tract level to develop a quantitative model for activity intensity as a function of land-use and built environments attributes for the New York City (NYC) region. The detailed exogenous variables considered were socio-demographics, land-use variables, transportation variables, and points of interests at the census tract level. The study also recognized that developing a single model for NYC would be restrictive and of limited use. Hence, prior to modeling, we classified the census tracts in NYC into four groups as a function of eight different land-use variables. The clusters identified were labeled as follows: Cluster 1 – Low-rise residential, Cluster 2 – Public facilities and Commercial, Cluster 3 – Low-rise residential and commercial, and Cluster 4 – High-rise residential and public facilities. The clustering approach, rather than considering the entire city as homogenous, allowed us to distinguish across different clusters. Subsequently, for each cluster as well as for the whole region, Negative Binomial (NB) Regression models were developed to study activity intensity patterns across the city. We compared the performance of the pooled model and cluster models by using the Bayesian Information Criterion. The comparison clearly illustrated the improved fit offered by the cluster specific NB models.

From the estimation results, we found that there are variations across the different clusters based on different exogenous variables. Moreover, the effects of the variables found to be different for some clusters and the pooled model supporting our hypothesis that activity intensity profile is not the same across the entire region. To further validate the model performance, we spatially represented the observed check-ins and predicted check-ins based on pooled and cluster-based models. From the visual comparison, across the three patterns, it was evident that the activity check-ins pattern for cluster-based models is closer to the observed patterns. We also illustrated the impact of various parameters on check-ins using a hot spot analysis. This tool enabled us to identify statistically significant spatial groups of high values (hot spots) and low values (cold spots). The results clearly illustrated the spatial regions that are significantly affected by different variables across the NYC region. The findings from our study provided insights on relative differences of activity engagements across the urban region. The proposed approach thus provides a complementary analysis tool to traditional transportation planning exercises.

The paper is not without limitations. The dataset employed in our analysis is from December 2011 through April 2012. Ideally, the consideration of a more recent time would be beneficial. The reader would note that the methodology developed could be applied to analyze newer versions of data that are freely available or purchased at a cost for an urban region of interest. The data used in our analysis is for a part of the year. Hence, accommodating for seasonality effects was not possible. Towards accommodating for these effects, it would be useful to consider obtaining data for a full year and generating the Check-in measures across different seasons. The dependent variable thus generated can be analyzed using the proposed model to identify seasonality differences. For our analysis, we did not consider the spatial correlations across different neighboring census tracts. In the future, it might be beneficial to examine the influence of spatial correlation in the count models.

## Data availability

The check-in data used in the study is sourced from http://infolab.tamu.edu/data/. The independent variables generated at the census tract level for this study cannot be shared because of the data confidentiality issues.

## References

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, X., Liu, Y., … Zook, M. (2015). Everyday space–time geographies: Using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *International Journal of Geographical Information Science, 29*(11), 2017–2039.

Ahmed, A., Hong, L., & Smola, A. J. (2013). Hierarchical geographical modeling of user locations from social media posts. *Proceedings of the 22nd international conference on World Wide Web*, Rio de Janeiro, May 13-17.

Anderberg, M. R. (2014). *Cluster analysis for applications: Probability and mathematical statistics: A series of monographs and textbooks, vol. 19*. Cambridge, MA: Academic Press.

Bawa-Cavia, A. (2011). Sensing the urban: Using location-based social network data in urban analysis. *Pervasive PURBA Workshop*, San Francisco, June 12-15.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data *Computers, Environment and Urban Systems, 51*, 70–82.

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. *ICWSM, 2010*, 81–88.

Cranshaw, J., Hong, J. I., & Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. *The 6th International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, June 4–7.

Do, T. M. T, & Gatica-Perez, D. (2013). The places of our lives: Visiting patterns and automatic labeling from longitudinal smartphone data. *IEEE Transportation Mobility Computations*, 1, 1.

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. *Proceedings 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust & 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, Washington DC, September 3–5.

Frias-Martinez, V., & Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence, 35*, 237–245.

Gordon, E., & de Souza e Silva, A. (2011). *Net locality: Why location matters in a networked world.* Hoboken, NJ: John Wiley & Sons.

Hasan, S., & Ukkusuri, S. V. (2014). Urban activity pattern classification using topic models from online geo-location data, *Transportation Research Part C: Emerging Technologies, 44*, 363–381.

Hu., Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems, 54*, 240–254.

Karlaftis, M. G., & Tarko, A. P. (1998). Heterogeneity considerations in accident modeling. *Accident Analysis & Prevention, 30*(4), 425–433.

Klin, F. & Pozdnoukhov, A. (2012). When a city tells a story: Urban topic analysis. *Proceedings of the 20th International Conference on Advances in Geographic Information Systems-SIGSPATIAL '12*, 482-485. New York: ACM. doi.org/10.1145/2424321.2424395

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr, *Cartography and Geographic Information Science, 40*(2), 61–77.

Miller, H. J. (2014). Activity-based analysis. *Handbook of regional science* (pp. 705724). New York: Springer.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: Universal patterns in human urban mobility. *PLoS One, 7*(5): e37027. doi.org/10.1371/journal.pone.0037027

Pew Research Center. (2017). Social media fact sheet, *Pew Res. Cent. Internet, Sci. Tech.* Retrieved from https://www.pewinternet.org/fact-sheet/social-media/

Pinjari, A., Eluru, N., Bhat, C., Pendyala, R., & Spissu, E. (2008). Joint model of choice of residential neighborhood and bicycle ownership: Accounting for self-selection and unobserved heterogeneity. *Transportation Research Record, 2082*, 17–26.

Pinjari, A. R., & Bhat, C. R. (2011). Activity-based travel demand analysis. In *A handbook of transport Economics*. Cheltenham, UK: Edward Elgar Publishing.

Rzeszewski, M. (2018). Geosocial capta in geographical research–a critical analysis, *Cartography and Geographic Information Science, 45*(1), 18–30.

Sengstock, C., & Gertz, M. (2012). Latent geographic feature extraction from social media. *Proceedings of the 20th International Conference on Advances in Geographic Information Systems-SIGSPATIAL '12*, 149–158. New York: ACM. doi.org/10.1145/2424321.2424342

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS One, 10*(3), e0115545.

Wakamiya, S., Lee, R., & Sumiya, K. (2011). Urban area characterization based on semantics of crowd activities in Twitter (Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). *Lecture Notes in Computer Science, 6631*, 108–123. doi.org/10.1007/978-3-642-20630-6_7

Winkelmann, R. (2008). *Econometric analysis of count data.* Berlin, Germany: Springer Science & Business Media.

Zhan, X., Ukkusuri, S. V., & Zhu, F. (2014). Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics, 14*(3–4), 647–667.