Landscape and Urban Planning 180 (2018) 166–178

Contents lists available at ScienceDirect

Landscape and Urban Planning

journal homepage: www.elsevier.com/locate/landurbplan

Research Paper

Quantifying place: Analyzing the drivers of pedestrian activity in dense urban environments

Yuan Lai, Constantine E. Kontokosta*

Dept. of Civil and Urban Engineering & Center for Urban Science and Progress, New York University, United States

ARTICLE INFO

Keywords:
Pedestrian mobility
Walkability
Urban planning
Urban computing
Machine learning
Urban data

ABSTRACT

Understanding pedestrian behavior is critical for many aspects of city planning, design, and management, including transportation, public health, emergency response, and economic development. This study bridges in-situ observations of pedestrian activity and urban computing by integrating high-resolution, large-scale, and heterogeneous urban datasets and analyzing both fixed attributes of the urban landscape (e.g. physical and transit infrastructure) with dynamic environmental and socio-psychological factors, such as weather, air quality, and perceived crime risk. We use local pedestrian count data collected by the New York City (NYC) Department of Transportation (DOT) and an extensive array of open datasets from NYC to test how pedestrian volumes relate to land use, building density, streetscape quality, transportation infrastructure, and other factors typically associated with urban walkability. We quantify, classify, and analyze place dynamics, including contextual and situational factors that influence pedestrian activity at high spatial–temporal resolution. The quantification process measures the urban context by extracting rich, yet initially fragmented and siloed, urban data for individual geolocations. Based on these features, we then construct contextual indicators by selecting and combining features relevant to pedestrian activity, and develop a typology of place to support the generalizability of our analysis. Finally, we use multivariate regression models with panel-corrected standard errors to estimate how specific contextual features and time-varying situational indicators impact pedestrian activity across time of day, day of the week, season, and year. The results provide insights into the key drivers of local pedestrian activity and highlight the importance accounting for the immediate urban environment and socio-spatial dynamics in pedestrian behavior modeling.

1. Introduction

Many factors of the urban built environment influence the number and activity of pedestrians at a given place and time. In practice, urban planners and transportation engineers analyze local pedestrian volumes in order to measure the walkability of a neighborhood, physical activity, traffic safety, impacts of new development, and foot traffic near local retail establishments (Cervero & Duncan, 2003; Day, 2016; Klein, Koeser, Hauer, Hansen, & Escobedo, 2016; Pikora et al., 2002; Southworth, 2005). Increasingly, big data and urban computing are generating renewed interest in urban mobility and locational intelligence to understand – and ultimately predict – pedestrian activity and the effects of physical and environmental factors on behavior. However, analyzing these dynamics at micro- and meso-scales is complicated by the diversity of hyperlocal urban socio-technical systems. Without some quantification of this complex environment, solely counting pedestrians will not provide comparable and generalizable knowledge across locations and time. The absence of contextual and situational information has remained a major barrier to more robust analyses of human activity in urban environments.

There exist non-trivial gaps between observational studies of human mobility behavior and research on the influence of the urban built, natural, and social environments in understanding place. William Whyte, a pioneer in the study of urban street life, was one of the first to combine qualitative observations of human activity with manually-collected characteristics of the surrounding urban environment (Whyte, 1980). Broadly, Whyte investigated how human activity was impacted by urban externalities including space layout, built form, street furniture, and environmental factors such as noise and light. Whyte's work was revolutionary as it was one of the earliest attempts to measure human activity in public space using an objective, data-driven approach. However, manually measuring and observing the physical environment is labor intensive, time-consuming, and difficult to generalize, which, despite the significant practical impact of his work,

* Corresponding author at: 370 Jay Street, 12th Floor, Brooklyn, 11201 New York, United States.
E-mail addresses: yuan.lai@nyu.edu (Y. Lai), ckontokosta@nyu.edu (C.E. Kontokosta).

https://doi.org/10.1016/j.landurbplan.2018.08.018
Received 10 July 2017; Received in revised form 18 December 2017; Accepted 24 August 2018
Available online 13 September 2018
0169-2046/

limited the scope of the research to a relatively few public plazas in New York City (NYC). New data streams and computational methods, however, now enable us to extract a broad array of locational descriptors, measure pedestrian activity, and develop models that help to explain the drivers of urban pedestrian behavior. While this empirical approach cannot replace rigorous qualitative methods to understand social dynamics, it can be used to reveal new insights from data that can inform and advance our collective understanding of place.

This study bridges in-situ observations of pedestrian counts and urban computing by integrating high-resolution, large-scale, and heterogeneous urban datasets and analyzing both fixed attributes of the urban landscape (e.g. physical and transit infrastructure) with dynamic environmental and socio-psychological factors, such as weather, air quality, and perceived crime risk. We use local pedestrian count data collected by the NYC Department of Transportation (DOT) and an extensive array of open datasets from NYC to test how pedestrian volumes relate to land use, building density, streetscape quality, transportation, and other factors typically associated with urban walkability. We quantify, classify, and analyze place dynamics, including contextual and situational factors that influence pedestrian activity at high spatial–temporal resolution. The quantification process measures the urban context by extracting rich, yet initially fragmented and siloed, urban data for individual geolocations. Based on these features, we then construct contextual indicators by selecting and combining features relevant to pedestrian activity, and develop a typology of place to support the generalizability of our analysis. Finally, we use multivariate regression models with panel-corrected standard errors to estimate how specific contextual features and time-varying situational indicators impact pedestrian activity across time of day, day of the week, season, and year.

We begin by introducing the literature on measuring pedestrian activity and quantifying urban place. We present our research hypotheses and their theoretical motivation, and describe our data collection and integration process for pedestrian count data and ancillary datasets in the urban domain. We then discuss our methodology and the results of both the urban typology classification and our model analyzing the drivers of pedestrian activity. Finally, we conclude with a discussion of potential applications of our research, its limitations, and future work.

2. Literature review

The literature on understanding the impacts of place on pedestrian activity spans three specific themes: methods for measuring pedestrian activity, the quantification of urban externalities, and the dynamics between human behavior and the built environment. Location-based pedestrian counts are commonly used as they provide a normalized metric for local population and traffic volume (Sallis, 2009). Such data are typically collected through two methods of in-situ observation: automated monitoring and manual counting. Sensors and computer vision technology have become a more pervasive and cost-effective method for tracking pedestrians in real-time (Dollar, Wojek, Schiele, & Perona, 2012; Hediyeh, Sayed, Zaki, & Mori, 2014; Ma & Chan, 2013; Placemeter, 2017; Traunmueller, Johnson, Malik, & Kontokosta, 2018). However, the reliability and validity of such methods remains a source of concern, as previous studies indicate that infrared counters, for instance, tend to under-count pedestrians (Schneider, Arnold, & Ragland, 2009) and video technologies can raise serious privacy concerns. On the other hand, manual counting has been the traditional approach for collecting observational data at a specific location and time period (Davis, King, & Robertson, 1988; Hajrasouliha & Yin, 2014). City agencies and business organizations often conduct pedestrian counts to monitor local commercial activity and the effect of space programming changes (Garment District, 2016; Times Square, 2016). Although labor intensive and producing relatively small sample sizes, manual counting provides ground-truth data with consistent units, a clearly-defined observation period, and simple counting methods (DOT, 2016; Gehl, 2008; MPO, 2016).

Conventionally, in-situ observation classifies data by geolocation (latitude and longitude) as a specific point that does not contain other defining spatial or place characteristics. This fact constrains the contextual interpretation of pedestrian activity data, due to the lack of associated measurements of urban street life. The concept of 'Five Themes of Geography' (location, place, human-environment interaction, movement, and region) was first adopted by the Association of American Geographers in 1984 and then widely used by other educational institutions (Boehm & Petersen, 1994). This concept defines 'place' as a logical progression that represents physical, environmental, and cultural attributes surrounding a location (Petersen, 1994). A place does not have a strict spatial dimension and it represents a multidimensional phenomenon with identity (name), shape (urban form), use (land use and buildings), connection (transit), cultural meanings, and temporal life (events and activity) (Golledge, 1992; Tuan, 1975). With an increasing concern around walkability in city planning and urban design, several studies and algorithms have been developed to try to define what makes a place conducive to pedestrian activity (Day, Boarnet, Alfonzo, & Forsyth, 2006; Greenwald & Boarnet, 2001; Yin et al., 2013). Most of this work focuses on pedestrian friendliness and design guidelines that promote non-vehicular transportation, public health, or quality-of-life at a district level (Frank et al., 2010; Frank et al., 2006; Leslie et al., 2007; Owen et al., 2007). Contextual urban factors such as urban form, street design, and socioeconomic characteristics have been found to influence the nature and extent of walkable areas (Alexander, 1977; Gehl, 1987; Jacobs, 1961; Lynch, 1960). Data on these urban externalities are either generated through on-site observations and qualitative studies, or extracted from city-level datasets (Clifton & Kreamer-Fults, 2007). The more quantitative of this previous work has used geographic information system (GIS) software to map data and generate spatial statistics by location (Leslie et al., 2007). A general approach uses circular buffers to extract data on land use, road networks, traffic volume, biking facilities, buildings, and parking areas (Brownson, Hoehner, Day, Forsyth, & Sallis, 2009; Croner, Sperling, & Broome, 1996; Neckerman, Purciel-Hill, Quinn, & Rundle, 2013; Schneider et al., 2009; Tresidder, 2005). The selected radii of a given buffer area are defined by empirical walkability measures such as 1/4 mile, 1/8 mile, or 1/16 mile (Ewing, Hajrasouliha, Neckerman, Purciel-Hill, & Greene, 2016; Yin et al., 2013; Villanueva, 2014). Based on buffer-extracted variables, researchers further developed indicators including density, diversity, design, destination, and distance to transit as the precursors for a quantitative summary of a place (Ewing et al., 2016; Frank, Schmid, Sallis, Chapman, & Saelens, 2005; Freeman et al., 2013). However, these efforts are typically conducted with data of low spatial resolution or aggregated geographies, such as census tracts or zip codes, and exclude more dynamic in-situ conditions, including population composition, weather, and anomalous events, due to computational limitations.

Macro-scale studies of urban mobility tend to focus on network analytics or regional patterns, with less attention given to micro-locations and micro-behaviors (Calabrese, Pereira, Di Lorenzo, Liu, & Ratti, 2010; Hillier & Hanson, 1989; Qian et al. 2017; Yuan, Zheng, & Xie, 2012). Many of these studies, typically from the transportation planning literature, emphasize one aspect of urban systems, such as a road network or particular transit mode, without considering localized places as an ensemble of multiple objects and dimensions (Lukermann, 1964). These limitations prevent a deep understanding how and why human activity varies by place and time within and across distinct neighborhoods or communities (Batty et al., 1998). A recent emphasis on urban computing and big data offers new computational methodologies and data sources for measuring and defining place (Griffin et al., 2016). "Smart city" initiatives and ubiquitous sensing technologies generate rich urban data from various sources and domains (Kitchin 2014; Kontokosta 2016). These new data sources provide opportunities to measure a locality at a geo-point through data fusion of multiple cross-domain datasets (Zheng, 2015). Computationally, any geolocation can be joined with various spatial data (land parcel, Census Block, or Community District) by physical proximity and further integrated with non-spatial data by unique identifier (land parcel ID), administrative label (neighborhood or district name), or place name (e.g. 'Times Square'). This quantification process enriches any geolocation with contextual information collectively definable as a 'place', which enables us to better interpret in-situ observational data and infer local activities. However, there has been an absence of scalable, reproducible approaches for quantifying places from multiple disparate datasets and analyzing in-situ observation data with surrounding conditions, since this process requires both long-term observational data at the local level and comprehensive datasets on urban conditions across individual systems.

3. Background and motivation

3.1. The NYC urban data landscape

New York City is one of the most walkable cities in the U.S., with a long legacy of promoting pedestrian activity through innovative policy, design, and technology. The Department of Transportation (DOT) has been conducting a bi-annual pedestrian count since 2009 to establish a longitudinal dataset for analysis of commercial corridors and priority intersections (NYC DOT, 2016). Every year in May and September, DOT conducts manual counting at 114 locations (100 street intersections and 14 bridge locations) across the five boroughs (Fig. 1). Each pedestrian count includes observations during three time periods in the same week: weekday morning (7am–9am), weekday afternoon (4 pm–7pm), and weekend mid-day (12 pm–2 pm). This is the most extensive pedestrian monitoring program by a city agency in NYC. Such observational data have important value for long-term economic development studies, since pedestrian volumes are connected to most urban activities, including commuting, shopping, dining, and tourism.

NYC has also been a pioneer in data-driven city operations and open data, making it a data-rich environment to study urban dynamics. Features relevant to pedestrians and mobility patterns, including land use, buildings, street network, public transit facilities, and sidewalks, have been digitized and disseminated as publicly-available data. However, these data have been collected in silos by individual agencies, creating challenges to the effective integration and analysis of disparate datasets. When properly extracted, cleaned, and joined, such data enable the quantification of a hyperlocal area based on its surrounding urban context, which allow us to characterize and define the physical and environmental dimensions of a place. Our research interests, then, originate from a simple question: can we quantify the elements a place and infer local human activity from its characteristics? If so, such knowledge would be of value to numerous city agencies, urban planners, and designers in order to analyze, predict, and encourage pedestrian activity in the future.

3.2. Research hypotheses

Our goal is to quantify the location-based characteristics of a place and explore their impacts on observed pedestrian activity. Computationally, we extract geo-referenced information through spatial querying of multiple datasets simultaneously, while calculating specific indicators to characterize both the proximate urban context and time-varying situational conditions. We capture variances in pedestrian

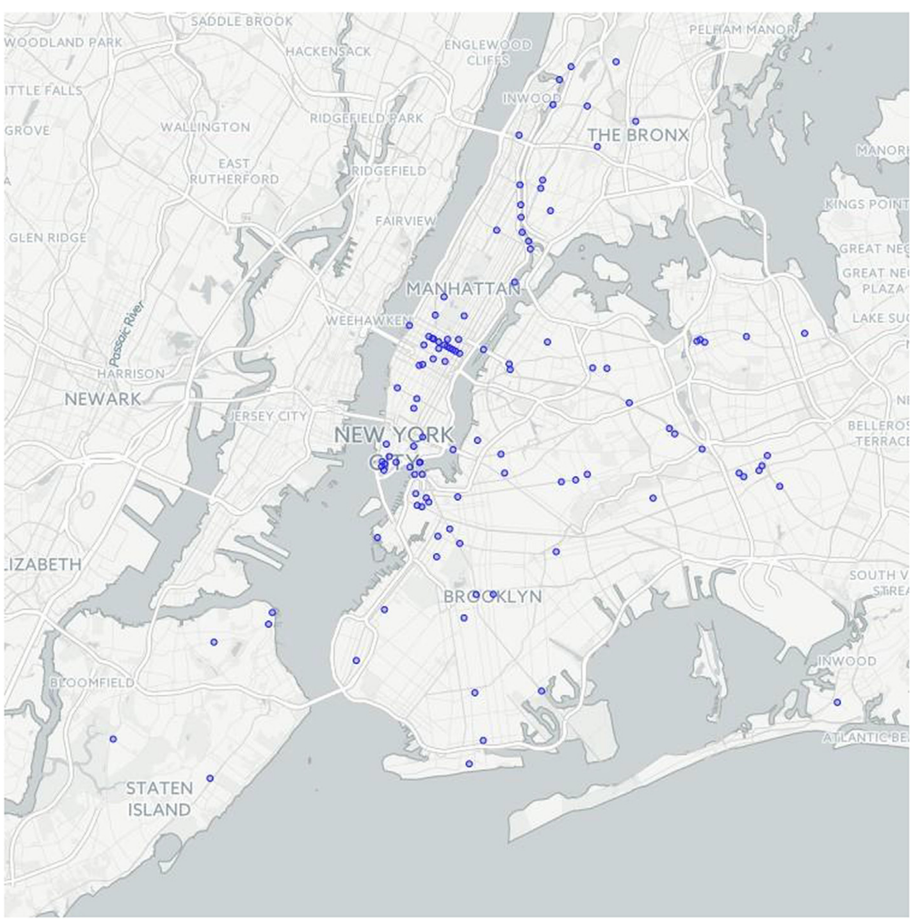

Fig. 1. DOT pedestrian count locations.

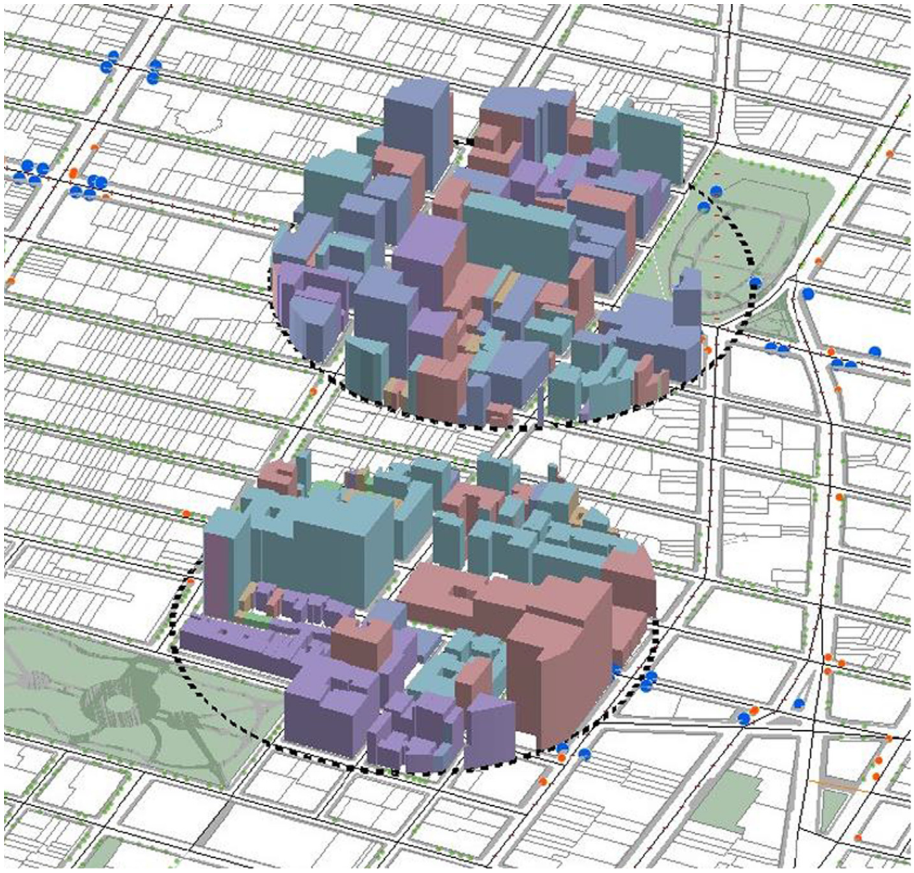

Fig. 2. Each place as an ensemble of features defined through cross-domain urban data fusion.

activity – measured by observed population counts – based on several attributes of the built, natural, and social environment in cities, including infrastructure, population, and socio-economic characteristics. The results are intended to improve our understanding of the drivers of pedestrian activity and to provide empirical evidence of the factors that influence walkability in dense urban environments.

There are two hypotheses that we test in this study. First, we assume that pedestrian behavior is influenced by time invariant features of urban form that are consistent across time and space. To operationalize this, we begin by defining a *place* as an ensemble of features measuring the contextual characteristics of natural, physical, and social systems (Eq. (1)):

$$Place_i = f\,(CT_{1,i},\ CT_{2,i},\cdots,CT_{n,i}) \tag{1}$$

where *CT* are individual *contextual factors* for a given place *i*. Such contextual factors measure (relatively) fixed land use, transportation, population, infrastructure, and socio-economic characteristics of a location (Fig. 2). These indicators can be further classified and categorized into an urban typology based on calculated spatial differentials in the variance of these attributes.

Second, controlling for the fixed contextual factors described above, we expect temporal fluctuations in pedestrian activity will be impacted by dynamic features of the urban environment, defined here as *situational factors*. These time-varying hyperlocal factors include weather conditions, local anomalous events, crime activity, and environmental conditions. Both long-term contextual factors and short-term situational factors drive local pedestrian behavior as follows (Eq. (2)):

$$PedestrianActivity_{i,t} = f\,(CT_{1,i},\ CT_{2,i},\cdots,CT_{n,i},\ SI_{1,i,t},\ SI_{2,i,t},\cdots,SI_{n,i,t}) \tag{2}$$

where pedestrian activity at place *i* during time *t* can be inferred by its urban context (*CT*) and situational conditions (*SI*). Situational factors may help explain variances in pedestrian activity, but will not impact the underlying place typology.

4. Data and methodology

4.1. Data fusion

We first assemble various features from multiple, heterogeneous data sources to measure land use, building density, population density, street network, street design, and public transit facilities, among others (Table 1). The datasets are all publicly-available "open" data from federal, state, and city agencies, including the U.S. Census Bureau, the NYC Department of Transportation (DOT), the Department of City Planning (DCP), the Metropolitan Transportation Authority (MTA), the Department of Parks and Recreation, the New York Police Department (NYPD), NYC311, and the Department of Information Technology and Telecommunications (DoITT) (NYC DOT, 2016; PLUTO, 2016; LEHD, 2016; MTA Subway, 2016; MTA Bus, 2016; LION, 2015; DOITT, 2015; NYC Park, 2015; NYC Trees, 2015; NYPD, 2015; NYC 311, 2015).

In our data preparation process, we first integrate non-spatial data with related geometries using table-based attribute join by unique identifier based on its spatial or areal unit (e.g. tax lot, census block, neighborhood tabulation area, etc.). We then combine different identifiers for varying geographies through a spatial join or attribute join by equivalency table. For instance, we measure local population based on the Longitudinal Employer-Household Dynamics (LEHD) origin–destination employment statistics by the U.S. Census Bureau, which tracks the total number of workers ($\sum workers$) and residents ($\sum residents$) in a census block and provides a linkage between home and work census blocks (LEHD, 2016). We use the MapPLUTO dataset from DCP to extract land use and building characteristics data at the tax lot level (PLUTO, 2016; MapPLUTO, 2016). Relevant variables from this dataset include land use type, building floor area, building space usage, zoning,

Table 1
Data collection.

Data	Geometry	Time	Unit	Source
Bi-annual pedestrian count	Point	2009–2016	Location	Department of Transportation
PLUTO*	Polygon	2010–2015	Tax lot	Department of City Planning
Park area	Polygon	2015	Land parcel	Department of Parks & Recreation
Subway station entry	Point	2015	Location	Metropolitan Transportation Authority
Bus stop location	Point	2016	Location	Metropolitan Transportation Authority
Street geometry	Polyline	2016	Street segment	Department of City Planning LION*
Sidewalk coverage	Polygon	2016	Sidewalk geometry	Department of Information Technology & Telecommunications
Street pavement quality rating	Polyline	2015	Street segment	Department of Transportation
Street tree location	Point	2015	Location	Department of Parks & Recreation
Street bench location	Point	2016	Location	Department of Transportation
Bike rack location	Point	2016	Location	Department of Transportation
Local population	Polygon	2010–2015	Census block	U.S. Census LEHD*
Median household income	Polygon	2015	NTA*	Department of City Planning
Local crime (major felony incidents)	Point	2015	Location	New York City Police Department
311 complaints by category	Point	2010–2015	Location	NYC 311
Hourly temperature, humidity& weather event	–	2010–2015	Citywide	Weather Underground

NOTES: PLUTO: Primary Land Use Tax Lot Output; *LION: NYC Single Line Street Base Map; *LEHD: Longitudinal Employer-Household Dynamics; *NTA: Neighborhood Tabulation Area.

number of units, and land assessed value, extracted with a spatial-computable geometry identified by a unique property ID, known as the Borough-Block-Lot (BBL) number. Each tax lot is located within a 2010 census block, and each block is unique within a census tract. Thus, we first join the geometry of all census blocks with LEHD data by the census block unique identifier, and then assign each tax lot (as BBL) to the census block (as census block ID) in which it is located. Through this approach, we make local population and land use cross-computable at both spatial resolutions.

Our data collection also includes variables with high spatial and temporal variation, such as reported crime incidents from NYPD and NYC 311 citizen complaints (NYPD, 2015; NYC 311, 2015). Each incident is reported as a geo-point with time-stamp capturing the location and time of occurrence. Previous studies suggest that weather conditions will also have an impact on pedestrian volumes, thus we collect historical city-wide hourly weather data on temperature, humidity (dew point), and event (precipitation) for matching dates and hours for each observation period (Weather underground, 2016; Aultman-Hall, Lane, & Lambert, 2009). We define these dynamic indicators as situational factors in our explanatory models.

To measure local pedestrian activity, we use the bi-annual pedestrian count dataset by DOT as introduced in the Background section. This dataset contains pedestrian counts, location ID and geo-coordinates, and the date and time when the observation was conducted (DOT, 2016). To match our urban contextual data with each observation, we transform the original time-series data structure and treat each observation as a unique input by location, observation time period, date, and year. The final panel dataset has 3600 observations over 6 years (2010–2015) at 100 intersections (14 locations on bridges are excluded).

In order to quantify the characteristics of a selected place, we develop a spatial query algorithm in the Python programming environment to extract and integrate our datasets in a computationally efficient manner. Based on each geo-point, the algorithm generates a circular buffer of user-defined radius (e.g., 1/8-mile) to extract associated data and return outputs based on the equations we develop for appropriate indicators. For instance, our space diversity output captures lot level land use diversity based on an entropy index value given by (Eq. (3)):

$$Entropy = \sum_j \frac{[P_j \times ln(P_j)]}{ln(j)} \tag{3}$$

where P_j is the proportion of development in the *jth* land use type. The entropy index is normalized by the natural logarithm of the total number of land use types within the buffer, ranging from 0 (homogeneous) to 1 (diverse) to represent the relative diversity of mixed uses (Brown et al., 2009; Kockelman, 1997; Kontokosta, 2014; Leslie et al., 2007). For data contained within specific administrative boundaries (e.g. population at the census block level), we use a weighted sum based on the total area of each census block by land use captured within the selected buffer. For data without any administrative boundaries, we calculate total length of polyline (streets), total geometry area (sidewalk surface), or total number of points (street trees, bus stops), etc. In total, there are 35 output variables from our initial spatial query process.

Our spatial data query methodology allows us to select any point location in NYC and collect cross-domain information for a selected buffer area and time period. This process can also be re-created for any other city with an open data infrastructure. For example, to quantify an intersection in Lower Manhattan's Financial District, we are able to measure its surrounding context and situational conditions based on the matched location and time of a specific pedestrian count observation (Table 2). During the spatial query process, we found the typical 1/4-mile radius buffer used in previous studies to be too large to measure hyperlocal variances in a dense urban area like much of NYC. Therefore, we conduct the spatial query with three buffer sizes (1/4-mile, 1/8-mile, and 1/16-mile radius) and select the optimal radius (1/8-mile)

Table 2
Query raw output by location and time with one pedestrian count in Lower Manhattan Financial District.

Location Id:35	Year: 2015	Season: Spring	Time: Weekday PM	Hourly pedestrian: 2149	Street quality rating: (0–10): 7
Sidewalk area (Sq.Ft.): 305,407	Street trees: 18	Street bench: 0	Bike rack: 9	Bus stop: 6	Subway station entry: 8
Built area (Sq.Ft.): 13,034,991	Office area percentage: 77%	Residential area percentage: 16%	Retail area percentage: 4%	Other area percentage: 3%	Land use diversity: 0.59
Workers: 103,764	Residents: 7343	Temperature (F): 84	Dew Point (F): 50	Precipitation (mm):0	
Crime (assault): 5	Crime (murder): 0	Crime (rape): 0		Complaint (noise): 275	Complaint (health): 58

based on the resultant covariance matrix and model performance.

4.2. Feature engineering

From the spatial query, each place can be represented as a vector of raw variables:

$$x_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \cdots \\ x_{i,j} \end{bmatrix} \quad (4)$$

where x_{ij} is the *jth* raw variable output for *ith* location. Based on previous theories on the definition place described in our literature review, we create eight contextual variables by transforming or combining individual raw variable outputs. For instance, we measure 'streetscape' by calculating sidewalk coverage weighted by pavement quality and street amenities including street trees, public benches, and bike racks (LION, 2015; NYC Trees, 2015; DOITT, 2015; DOT Bicycle, 2016; DOT Bench, 2016). Previous studies use relative measures to evaluate local walkability and quality of urban design by a creating walkability index (Ewing & Clemente, 2013; Frank et al., 2005; Leslie et al., 2007). In a similar manner, we measure each place using normalized scores (Z-score) to integrate our indicators and identify a typology of place and its effect on local pedestrian activity (Table 3).

4.3. Urban typology

K-means clustering is used to classify our sample locations into three (k = 3) distinct clusters to generate a typology of localized urban form (shown as a radar plot in Fig. 3a and across the study area in Fig. 3b). We select three clusters based on Silhouette analysis of inter-cluster variance and within-cluster similarity (Fig. 4), and denote each as type 0, type 1, and type 2, respectively. Type 0 (in red) are places defined by high building density, significant concentrations of jobs, and accessibility to transit, but with relatively low proportions of housing. Type 1 (in blue) are places with attractive streetscapes and moderate building densities with a diversity of uses. Although not immediately proximate to public transit, these locations are primarily residential and exhibit traits that are consistent with pedestrian-friendly environments. Type 2 (in green) areas have low building densities and a relatively equal number of workers and residents. They have similar pedestrian volumes during weekday afternoon and weekend mid-day, and low volumes during weekday mornings, suggesting that pedestrian activity is driven by local residents rather than commuters.

Due to selection bias in the DOT's choice of pedestrian count locations, our typology classification does not necessarily represent all places in the city. Since K-means clustering groups samples based on inter-cluster variance and within-cluster similarity, the characteristics of each group (type 0, type 1, and type 2) may change when adding in new samples. However, our method and approach can be generalized to apply to any location in any city, provided the appropriate data are available.

4.4. Modeling pedestrian activity

Our typology classification captures physical infrastructure, land use, and other dimensions of a place that can be considered fixed, or those that change relatively infrequently. Of course, there are dynamic aspects of the urban environment and its condition (weather, crime, local complaints or construction events) that vary considerably, and can have a significant impact on walkability and pedestrian activity (defined here as *situational variables*). To understand both the physical and environmental correlates of pedestrian behavior, we construct a multivariate model with both contextual and situational variables, including local crime activity, 311 complaints, and weather conditions (Table 4). Considering pedestrian count y_{it} as a scalar observed at place i and time t, we model the relationship between a place and local pedestrian volume as a log-linear regression model specified as:

$$\ln(y_{it}) = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_d x_{idt} + \varepsilon_{it} = \sum_{d=0}^{d} \beta_d (x_{it}) + \varepsilon_{it} \quad (5)$$

where x is the independent variable, β is a coefficient to be estimated, and d is the total number of explanatory variables. The model includes time-varying indicators (x_{it} varies by year and location, e.g., local population), time-invariant inputs ($x_{it} = x_i$ for all t, e.g., number of bus stops) and space-invariant inputs ($x_{it} = x_t$ for all i, e.g., citywide weather condition).

The model captures annual changes in land use and population, extracted from annual updates to the PLUTO and Census datasets, and concurrent situational conditions based on the specific date and time of a given pedestrian count location. Assuming pedestrian activity will vary by time of day, we subset our data to weekday mornings (AM, n = 1200), weekday afternoons (PM, n = 1100), weekend mid-day (MD, n = 1200) periods, and run a panel Ordinary Least Squares (OLS) multivariate regression model on each subset to identify temporal variations in the drivers of pedestrian activity. Given possible serial correlation in the time-series count data, panel-corrected standard errors (PCSE) are used to provided more robust estimates in the presence of cross-sectional heteroscedascity (Beck & Katz, 1995). We also use the Mean Absolute Percentage Error (MAPE) to measure the goodness of fit of the regression and predictability of pedestrian activity by time and place.

There are several considerations in our choice of models. Linear or log-linear models have been widely adopted in previous studies to test statistical relationship between walkability measures and pedestrian

Table 3
Contextual variable, parameter and quantification.

Contextual Variable	Parameter
Density	$FloorAreaRatio = \frac{\sum BuildingFloorArea}{BufferArea}$
Diversity	$LanduseMixEntropy = \sum_j \frac{[P_j \times \ln(P_j)]}{\ln(j)}$
Streetscape = SidewalkCoverage × PavementQuality × StreetAmenity	$SidewalkCoverage = \frac{\sum SidewalkArea}{BufferArea}$
	Pavement quality = street quality inspection score (0–1)
	$StreetAmenity = \frac{\sum(bench, tree, bikerack)}{BufferArea}$
Transit	$PublicTransitAccess = \sum SubwayStationEntry \times 10 + \sum BusStop$
Jobs	$TotalLocalJobs = \sum LocalEmployees$
Home	$TotalLocalResidents = \sum LocalResidents$
Retail	$RetailAreaPercentage = \frac{\sum RetailFloorArea}{\sum BuildingFloorArea}$
Income	*Median household income*

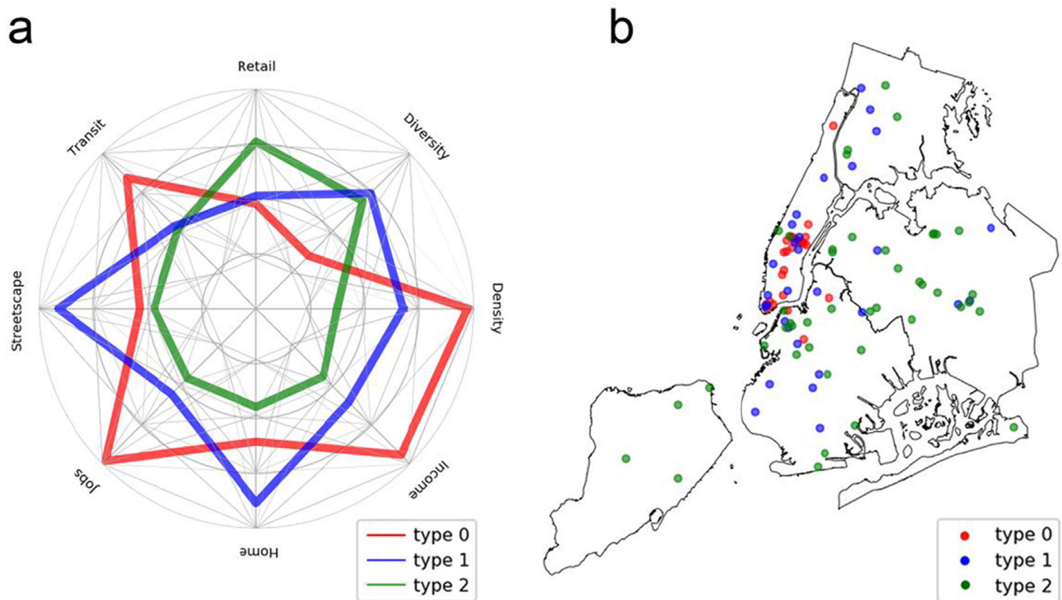

Fig. 3. Radar chart visualizing classified urban typology (3a) and classified places in NYC (3b).

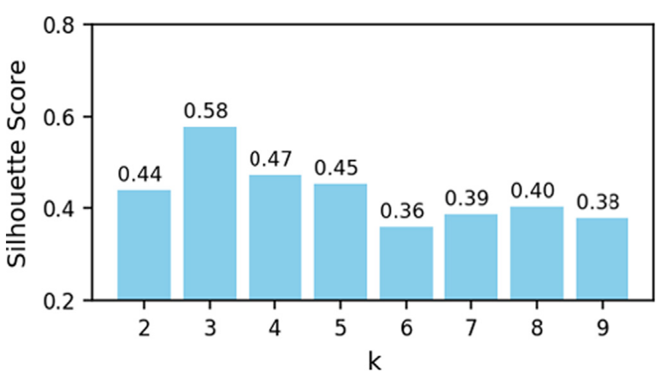

Fig. 4. Silhouette scores evaluating the optimal number of clusters ($K = 3$).

Table 4
Situational variable, parameter and quantification.

Situational Variable	Parameter
Crime	*Violent crime per capita* $= \frac{\sum(assault, rape, murder)}{\sum residents + \sum employees}$
Noise	*Noise Complaints per Capita* $= \frac{\sum(NoiseComplaints)}{\sum residents + \sum employees}$
Environmental Health	*Health-Related Complaints per Capita* $= \frac{\sum(ComplaintsAirQualityAndSanitation)}{\sum residents + \sum employees}$
Temperature	*Citywide Hourly Average Temperature (F)*
Humid (1/0)	*A binary variable, Humid = 1 if citywide Dew Point (F) > 60, else 0.*
Rainy (1/0)	*A binary variable, Rainy = 1 if precipitation > 0, else 0.*

volume (Miranda-Moreno, 2011; Schneider et al., 2009; Schneider, 2013). A log-linear model is appropriate for this analysis as it provides a reasonable approximation of the relationships in the data and allows for the interpretation of individual factors and their coefficients.

5. Findings

5.1. Variance in pedestrian volume

A descriptive analysis presents the differences in pedestrian activity across location, year, day of week, and time of day, and the panel analysis reveals spatial–temporal variances of pedestrian count data from 2009 to the present (Fig. 5). The results indicate an overall pedestrian count increase over the last seven years. We note that weekday afternoon counts exhibit an unexpected dip in 2012. We suspect this was due to an unclassified event at or near the time of the count observations, or a systematic error in the manual counting procedure during that period. Overall, when comparing pedestrian counts for specific locations and times between spring and fall, we find relative stable local pedestrian volumes regardless of season (Fig. 6a).

Looking at intra-week temporal variations, weekend pedestrian volumes are generally lower than weekday counts, and do not reveal a consistent pattern across locations (Fig. 6b). We would expect different contextual and situational factors to drive pedestrian volumes on weekdays and weekends, particularly given previous mobility research using cell-phone data (Hvel, Simini, Song, & Barabsi, 2014). A comparison of weekday morning and afternoon counts reveals a linear relationship between morning and afternoon pedestrian volumes on the same day (Fig. 6c).

When visualizing this relationship based on our three urban contextual typologies, we observe places of type 0 (in red) have a weaker correlation between AM and PM pedestrian counts. This indicates that places with high building density, large local populations, and well-connected transit have large intra-day fluctuations in pedestrians (Fig. 6d). This variance reflects a 'home-work-activity' dynamic by commuting workers, where the presence of certain amenities and services (shopping, dining, commercial services) result in a notable increase in pedestrian volumes during after-work hours. These descriptive findings reinforce the need to analyze pedestrian activity independently by time of day.

5.2. Drivers of pedestrian activity

The regression model results demonstrate fundamental differences between measuring walkability and predicting pedestrian volume. Walkability measures quantify a location's pedestrian attractiveness, which relates to environmental quality and travel preference. However, when estimating pedestrian volume at a specific place and time, walkability is just one of multiple factors involved, and destinations become more significant. Results from our multivariate regression model reveal a complex interaction between pedestrian activity, urban

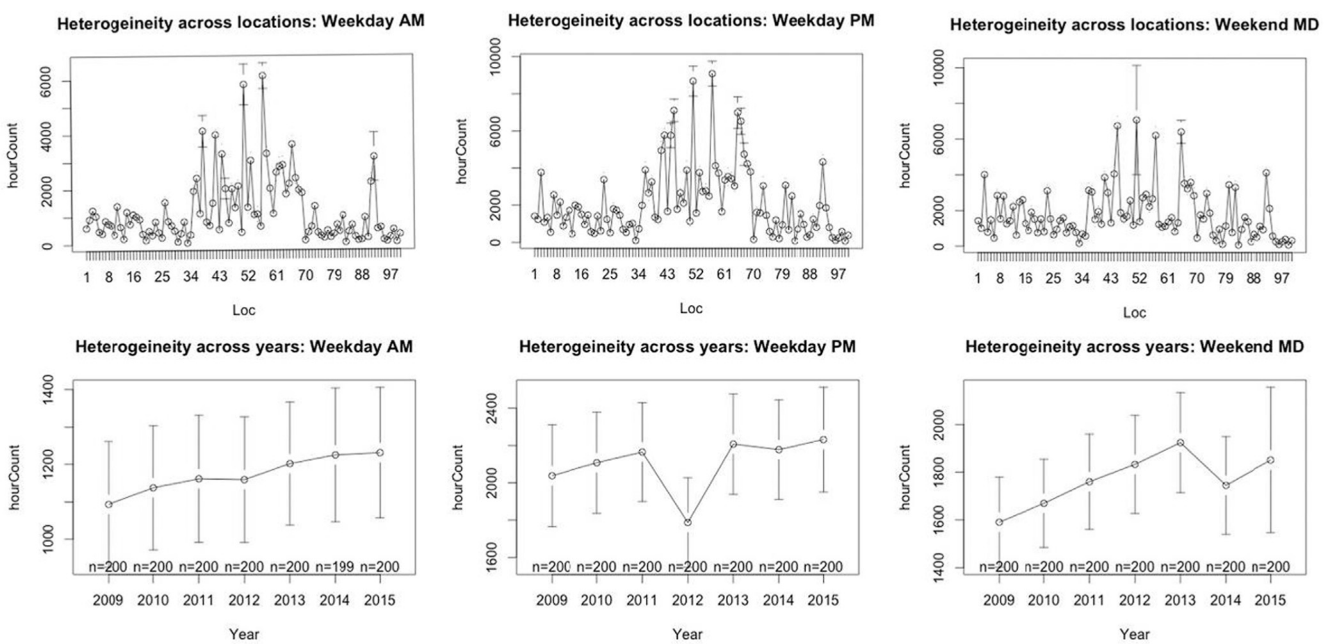

Fig. 5. Panel analysis of pedestrian observations.

context, and situational factors (Table 5). Weekday pedestrian activity is more predictable using urban contextual factors than weekend activity, based on R-squared values. Both *p* values and correlation coefficients indicate statistically significant relationships between pedestrian volumes and urban context during the three time periods.

As expected, building density, transit access, and the proportion of local residents are found to drive pedestrian activity regardless of the specific day or time of day. In contrast, walkability measures are only relevant during certain times of day. Looking at 'land use diversity', for example, the results show a weaker influence of land use mix on weekday mornings and weekends, indicating weekday afternoon pedestrian volumes are more sensitive to mixed-uses and driven by working, living, and recreation/leisure activities. The different correlation coefficients for a given variable across the three models reveal how that factor affects pedestrian volumes at different times of day. For instance, streetscape quality has a stronger impact on pedestrian activity during weekends than weekdays. The ratio of retail establishments in a given area, as another illustration, has a weak correlation with morning pedestrian activity, which demonstrates both the influence of typical business operating hours and more limited retail activity during workday mornings. Local crime events have a constant negative impact on pedestrian activity, potentially reflecting pedestrians' perception on neighborhood safety.

The results of our pedestrian volume predictive model indicate nontrivial variations in prediction errors across the identified urban typologies (Fig. 7). The distributions of percentage errors indicate pedestrian activity in Type 0 (red, high density and transit oriented) is quite predictable during weekdays, but not on weekends, while Type 2 (green, low density) activity is more difficult to estimate regardless of time period. We measure partial dependence to investigate how features interact during different times. Two-way partial dependence between local residents and workers reveals how they drive local pedestrian volumes collectively (Fig. 8). The result indicates the weekday morning populations are primarily driven by local workers, especially when the worker population reaches 5000 within a 1/8-mile buffer. There is a complex interaction between local workers and residents during weekday afternoons, as one would expect more interactions between the two given commuting patterns and after-work leisure activities. In contrast, local working population has no impact on weekends, as shown in the right-hand frame of Fig. 8. Such findings reinforce the regression model results and the relative prediction errors across our defined place typologies.

6. Discussion

In addition to a methodology for objectively quantifying places based on a range of attributes, this study explores new approaches to analyze multiple drivers of pedestrian activity varying by location and time in the urban environment. This location-based knowledge supports a deeper understanding of hyperlocal urban dynamics and enhances decision-making and activity prediction based on the measured urban context (Pantsar-Syva¨niemi et al., 2010). Our typology classification enables us to compare places with both absolute (location-specific conditions) and relative measures (comparative conditions). Our quantification method reveals the relative difference in multiple dimensions of place – including built density, transit, and local population composition – which explain, in part, varying pedestrian activity profiles and real-time local population flux at high-spatial resolution.

To illustrate the potential application of our model in planning and design, we select one location per typology classification to investigate how local pedestrian volumes are impacted by a specific change in urban land use and population (Fig. 9). Fulton Street-Broadway (type 0) is located in Lower Manhattan and serves as a transit hub connecting the World Trade Center and Financial District to the rest of the city. Chambers Street-Hudson Street (type 1), located in Tribeca, is an upscale neighborhood in Lower Manhattan with high real estate prices, cobblestone streets, and celebrity residents. Red Hook (type 2) is located on a peninsula at the southern edge of Carroll Gardens and Downtown Brooklyn, which is known for sparse public transit, limited development, and a high proportion of residents living in poverty. These three places illustrate the diversity in our typology classifications. Local workers drive pedestrian activity on weekday afternoons at Fulton Street-Broadway and Chambers Street-Hudson Street, where commercial space dominates trip destinations. In contrast, there is no dominating driver of pedestrian volumes in Red Hook.

Our regression model provides a quantitative tool to estimate the impact of future development, or changes to the design or programming of the urban environment, on local pedestrian activity (Table 6). For example, if we assume a building re-use project converts office space to retail, while holding other factors constant, we can measure its impact

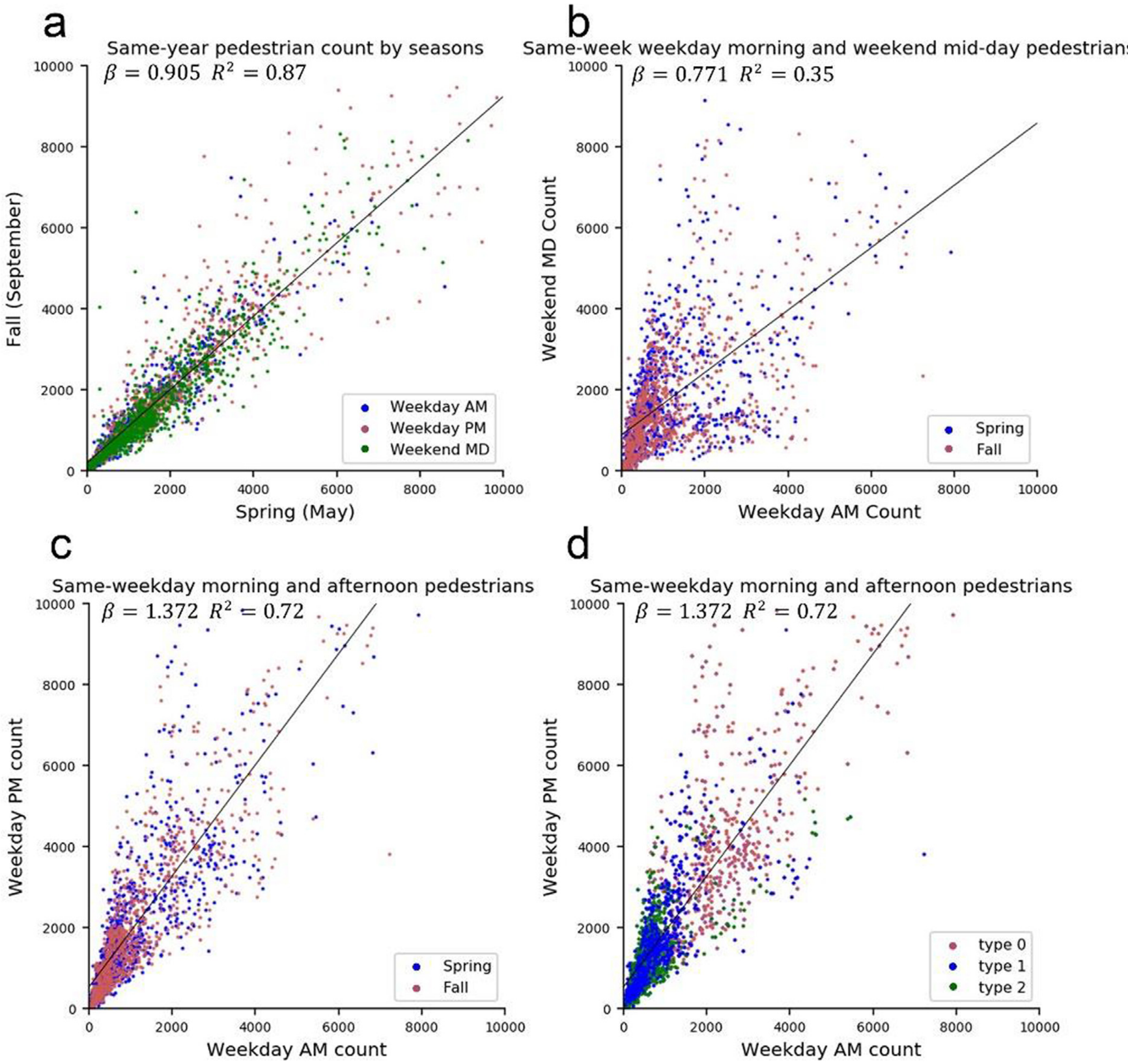

Fig. 6. Comparison of pedestrian counts at the same location and different times.

on local pedestrian activity as:

$$\ln\left(\frac{y'}{y}\right) = \ln(y') - \ln(y) = \beta_1(\text{Retail}' - \text{Retail}) + \beta_2(\text{Diversity}' - \text{Diversity}) \tag{6}$$

$$HourlyPedestrianImpact\,(\%) = \left(e^{\ln\left(\frac{y'}{y}\right)} - 1\right) \times 100 \tag{7}$$

A 20,000 square foot office space to retail conversion, for instance, would generate a greater relative impact on pedestrian activity in Red Hook (+13 per hour) than a similar change at Chambers Street (+5 per hour) on weekends. Although typical pedestrian volumes are higher around Chambers Street than in Red Hook, our methodology reveals how similar changes in urban development result in different magnitudes of impact on the local community based on its current development density and population composition. While a modest redevelopment scenario, this illustrates the importance of accounting for differences in contextual factors and underlying place typology when evaluating the impact of future design and planning decisions.

One limitation of this analysis is the low temporal resolution of the pedestrian count data, since DOT only conducts counting six times per year. This results in challenges to identifying variances across different time units, including time of the day, day of the week, and season. To ensure consistent longitudinal observations, DOT conducts its pedestrian counts during a typical day with normal weather condition and no special events. This data collection method constrains our capability to fully analyze the impacts of dynamic or situational factors, due to the limited variation in weather conditions and the absence of anomalous events. In addition, count data only represent volumes, which limits us to inferring the reason for the observed pedestrian behavior. For instance, previous studies categorized pedestrian activities into three types – necessary activities, optional activities, and social activities – and where these types tend to occur (Gehl, 1987). In order to analyze such dynamics, we need more detailed data to capture length of stay or frequent visitors beyond simple measures of volume. However, despite the limitations discussed above, we selected our count data to ensure quality, reliability, and consistency of our dependent variable. Other data sources, such as Wi-Fi probe request data, wearable accelerometer data, and geotagged social media data hold promise, but errors in linking these methods to actual counts of people can be significant (Böhm, 2016; Jiang et al., 2015; Kontokosta & Johnson, 2017; Kostakos, Juntunen, Goncalves, Hosio, & Ojala, 2013; Traunmueller et al., 2018).

Table 5
Multivariate OLS linear panel regression for data from 2010 to 2015.

Model Variable		Weekday AM		Weekday PM		Weekend MD	
		Coeff.	(PCSE)	Coeff.	(PCSE)	Coeff.	(PCSE)
Dependent Variable = ln (Hourly Pedestrian Count)							
(Intercept)		6.05	(0.047)***	7.23	(0.447)***	6.49	(0.559)***
Contextual Variable	Density	0.41	(0.065)***	0.64	(0.068)***	0.49	(0.037)***
	Diversity	0.11	(0.032)***	0.21	(0.034)***	0.13	(0.011)***
	Retail	−0.02	(0.034)	0.24	(0.036)***	0.29	(0.012)***
	Transit	0.25	(0.035)***	0.25	(0.036)***	0.24	(0.010)***
	Streetscape	0.06	(0.033)*	0.17	(0.035)***	0.20	(0.020)***
	Jobs	0.18	(0.056)***	0.13	(0.058)**	−0.02	(0.030)
	Home	0.11	(0.037)***	0.16	(0.038)***	0.16	(0.009)***
	Income	−0.07	(0.044)	−0.10	(0.047)**	−0.13	(0.019)***
Situational Variable	Crime	−0.36	(0.062)***	−0.01	(0.030)	−0.05	(0.020)**
	Noise	−0.18	(0.113)	−0.04	(0.031)	−0.02	(0.013)
	Env. Health	−0.04	(0.047)	0.05	(0.040)	−0.03	(0.032)
	Temperature	0.01	(0.005)	0.00	(0.006)	0.01	(0.007)
	Humid (1/0)	−0.14	(0.110)	−0.19	(0.088)**	−0.17	(0.198)
	Rainy (1/0)	−0.01	(0.107)	0.08	(0.143)	NA	NA
Overall Model							
Sample Size (N)		1200		1100		1200	
Adjusted R^2		0.59		0.58		0.42	
F- test		46.74***		43.51***		25.01***	

NOTES: Coeff. = coefficient and PCSE = Panel Corrected Standard Errors. *** = significant at 99% (p ≤ 0.01); ** = significant at 95% (p ≤ 0.05); * = significant at 90% (p ≤ 0.10).

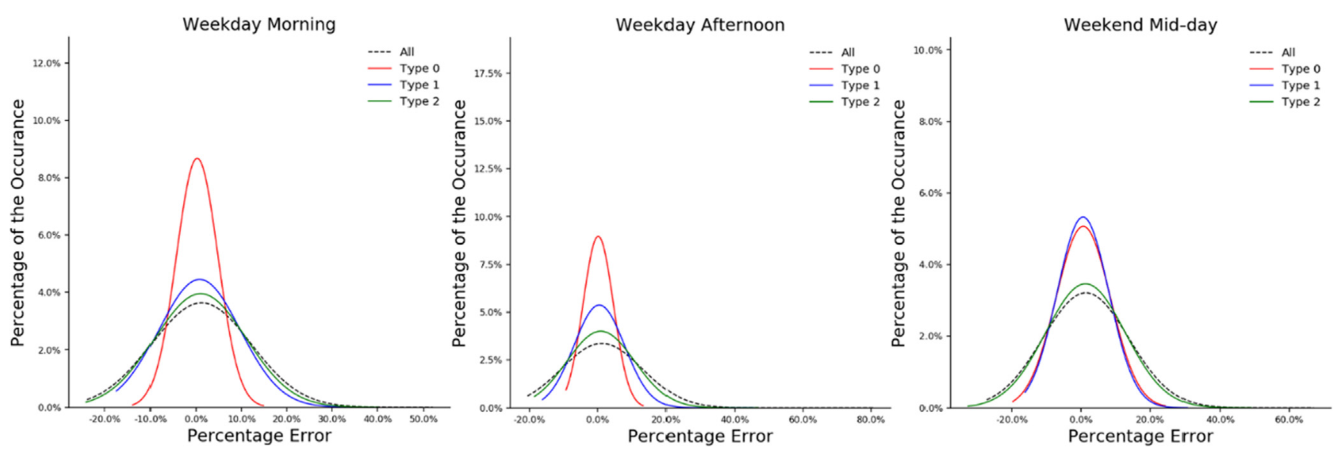

Fig. 7. Histogram of prediction percentage errors (%).

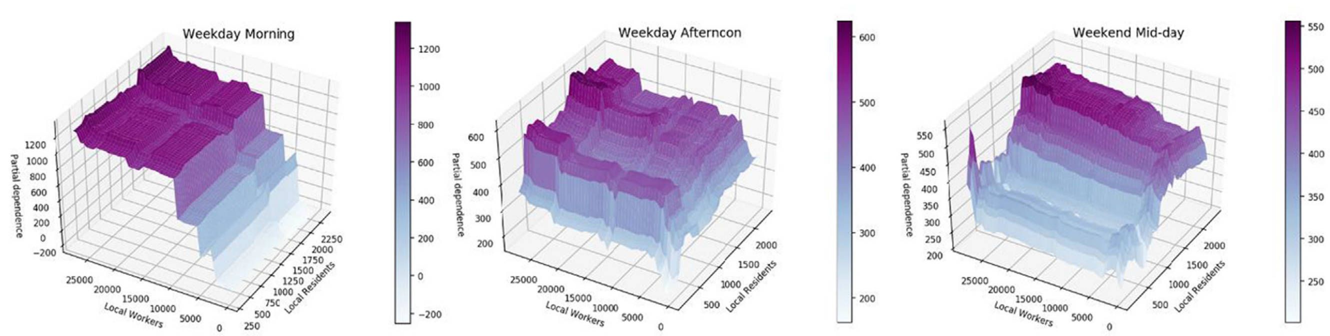

Fig. 8. The two-way partial dependence plot shows the dependence of local pedestrian volume on joint effects of local total residents and workers during different time periods.

7. Conclusion

Our exploratory analysis attempts to better understand the determinants of local pedestrian activity in urban environments. We develop a scalable and reproducible approach to quantify the physical, environmental, and socio-economic characteristics of a place for a given geo-location through heterogeneous data mining and classification, using an automated spatial query and unsupervised clustering algorithms. We are able to integrate disparate data sources, both in terms of attributes and spatial resolution, for a location to create an objective,

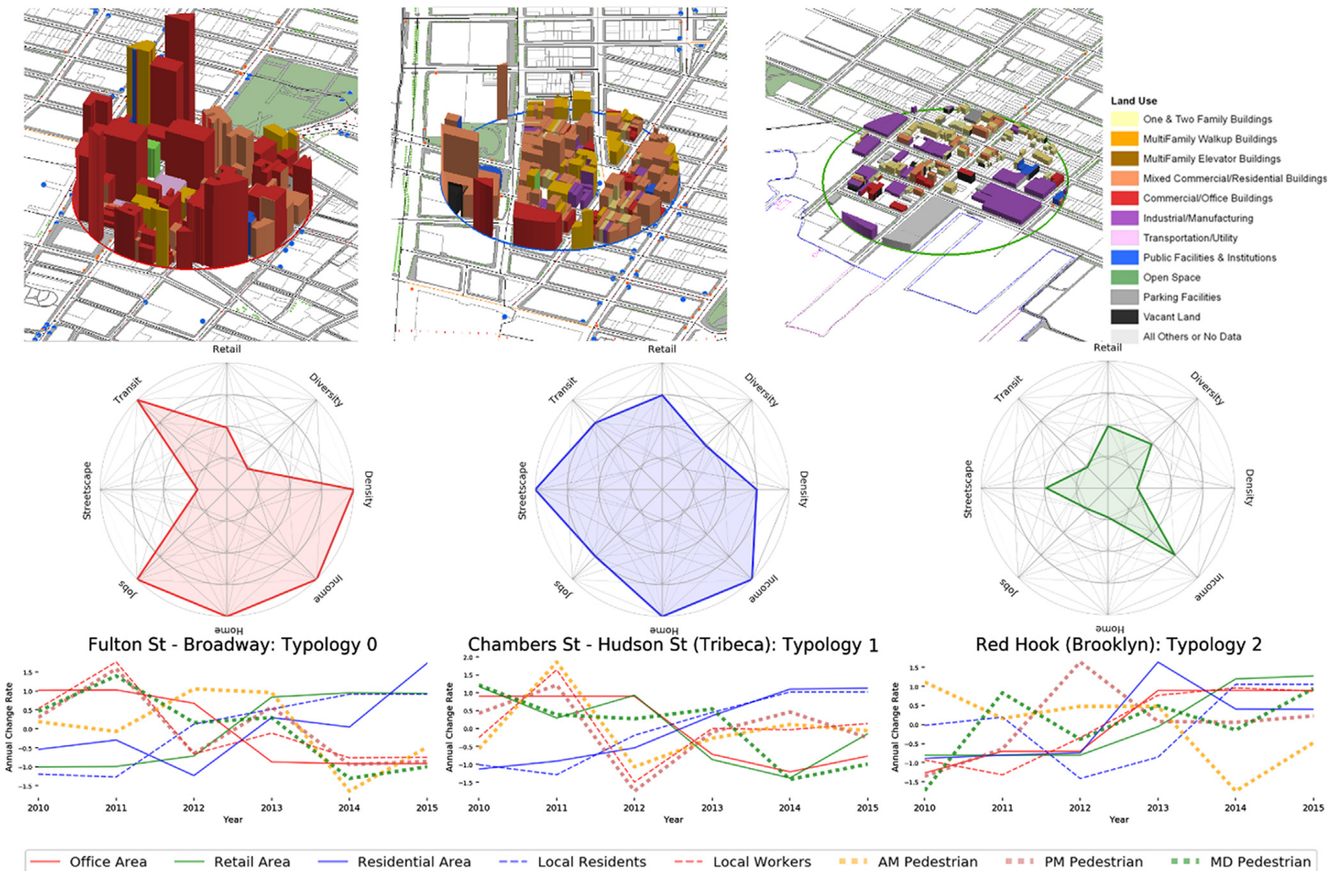

Fig. 9. Quantified urban context; examples for each place type.

Table 6
Location comparison based on data from Fall 2015.

	Fulton St.	Chambers St.	Red Hook
Hourly Pedestrian (AM)	2064	4092	42
Hourly Pedestrian (PM)	2346	3086	109
Hourly Pedestrian (MD)	1354	849	175
Total Built Area (Sq. Ft.)	9,447,820	4,411,112	824,620
Total Local Workers	23,217	4235	585
Total Local Residents	1675	1734	223
Office Percentage	63%	15%	5%
Retail Percentage	9%	13%	7%
Development Scenario: Convert 20,000 sq.ft. of office space to retail space			
AM Pedestrian Impact	−4 (−0.20%)	−4 (−0.10%)	0 (−0.40%)
PM Pedestrian Impact (%)	+3 (0.14%)	+28 (0.92%)	+8 (6.96%)
MD Pedestrian Impact (%)	+5 (0.35%)	+10 (1.13%)	+13 (7.40%)

transparent method for quantifying urban places. We combine contextual factors that represent hyperlocal physical infrastructure and land use characteristics to define the urban form, and link these to dynamic aspects of the urban environment – situational factors - that describe social, behavioral, and environmental conditions that vary with time and space.

This study contributes to both the objective quantification of a 'place' using large-scale urban data at high spatial resolution and to the understanding of the links between urban landscape and pedestrian volumes. Our findings provide empirical evidence for the urban environmental attributes that influence walkability. Building density, local population, and public transit are found to be consistent drivers of pedestrian activity in New York City. Other characteristics of the urban landscape impact pedestrian activity differently based on time of day, day of the week, and surrounding context. Our analysis not only provides insight into the drivers of pedestrian activity in cities, but also enables policy-makers and designers with the ability to measure the impact of changes in the built environment. In future work, we will explore non-traditional datasets to potentially achieve better predictive performance by training models with more dynamic pedestrian data at high spatial–temporal resolution. Additional opportunities exist to enhance our typology classification by quantifying more places across cities, and inspect how other activities, such as economic transactions, vary by urban typology.

References

Alexander, C. (1977). *A pattern language: Towns, buildings, construction.* Oxford University Press.

Aultman-Hall, L., Lane, D., & Lambert, R. (2009). Assessing impact of weather and season on pedestrian traffic volumes. *Transportation Research Record: Journal of the Transportation Research Board, 35*–43.

Batty, M., Jiang, B., & Thurstain-Goodwin, M. (1998). Local movement: agent-based models of pedestrian flows.

Beck, N., & Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review, 89*(3), 634–647.

Boehm, R. G., & Petersen, J. F. (1994). An elaboration of the fundamental themes in geography. *Social Education, 58*, 211.

Böhm, M. F. (2016). *Evaluating the usage of Wi-Fi and bluetooth based sensors for pedestrian counting in urban areas. European Transport Conference*. Barcelona: Association for European Transport (AET).

Brown, B. B., Yamada, I., Smith, K. R., Zick, C. D., Kowaleski-Jones, L., & Fan, J. X. (2009). Mixed land use and walkability: Variations in land use measures and relationships with BMI, overweight, and obesity. *Health & Place, 15*, 1130–1141.

Brownson, R. C., Hoehner, C. M., Day, K., Forsyth, A., & Sallis, J. F. (2009). Measuring the built environment for physical activity: State of the science. *American Journal of Preventive Medicine, 36*, 99–123.

Calabrese, F., Pereira, F. C., Di Lorenzo, G., Liu, L., & Ratti, C. (2010). The geography of taste: analyzing cell-phone mobility and social events. *International Conference on Pervasive Computing* (pp. 22–37). Springer.

Cervero, R., & Duncan, M. (2003). Walking, bicycling, and urban landscapes: Evidence from the San Francisco Bay area. *American Journal of Public Health, 93*, 1478–1483.

Clifton, K. J., & Kreamer-Fults, K. (2007). An examination of the environmental attributes associated with pedestrian–vehicular crashes near public schools. *Accident Analysis &*

Prevention, 39, 708–715.
Croner, C. M., Sperling, J., & Broome, F. R. (1996). Geographic information systems (GIS): New perspectives in understanding human health and environmental relationships. *Statistics in Medicine, 15*, 1961–1977.
Davis, S. E., King, L. E., & Robertson, H. D. (1988). Predicting pedestrian crosswalk volumes. *Transportation Research Report, 1168*, 25–30.
Day, K. (2016). Built environmental correlates of physical activity in China: A review. *Preventive Medicine Reports, 3*, 303–316.
Day, K., Boarnet, M., Alfonzo, M., & Forsyth, A. (2006). The Irvine-Minnesota inventory to measure built environments: Development. *American Journal of Preventive Medicine, 30*(2), 144–152.
Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*, 743–761.
DOT (2016). *Bi-annual pedestrian count data feed.* New York City, NY: Department of Transportation.
Ewing, R., & Clemente, O. (2013). *Measuring urban design: Metrics for livable places.* Island Press.
Ewing, R., Hajrasouliha, A., Neckerman, K. M., Purciel-Hill, M., & Greene, W. (2016). Streetscape features related to pedestrian activity. *Journal of Planning Education and Research, 36*, 5–15.
Frank, L. D., Sallis, J. F., Conway, T. L., Chapman, J. E., Saelens, B. E., & Bachman, W. (2006). Many pathways from land use to health: Associations between neighborhood walkability and active transportation, body mass index, and air quality. *Journal of the American Planning Association, 72*, 75–87.
Frank, L. D., Sallis, J. F., Saelens, B. E., Leary, L., Cain, K., Conway, T. L., et al. (2010). The development of a walkability index: Application to the neighborhood quality of life study. *British Journal of Sports Medicine, 44*, 924–933.
Frank, L. D., Schmid, T. L., Sallis, J. F., Chapman, J., & Saelens, B. E. (2005). Linking objectively measured physical activity with objectively measured urban form: Findings from SMARTRAQ. *American Journal of Preventive Medicine, 28*, 117–125.
Freeman, L., Neckerman, K., Schwartz-Soicher, O., Quinn, J., Richards, C., Bader, M. D., et al. (2013). Neighborhood walkability and active travel (walking and cycling) in New York City. *Journal of Urban Health, 90*, 575–585.
Garment District (2016). Pedestrian counts report, the Garment District NYC. Available from: http://garmentdistrict.nyc/real-estate/pedestrian-counts/.
Gehl, J. (1987). *Life between buildings: Using public space.* Island Press.
Gehl, J. (2008). World class streets: Remaking New York City's public realm." New York City Department of Transportation.
Golledge, R. G. (1992). Place recognition and wayfinding: Making sense of space. *Geoforum, 23*(2), 199–214.
Greenwald, M., & Boarnet, M. (2001). Built environment as determinant of walking behavior: Analyzing non-work pedestrian travel in Portland, Oregon. *Transportation Research Record: Journal of the Transportation Research Board,* 33–41.
Griffin, M., Nordstrom, B. W., Scholes, J., Joncas, K., Gordon, P., Krivenko, E., et al. (2016). A case study: Analyzing city vitality with four pillars of activity: Live, work, shop, and play. *Big Data, 4*, 60–66.
Hajrasouliha, A., & Yin, L. (2014). The impact of street network connectivity on pedestrian volume. *Urban Studies, 52*(13), 2483–2497.
Hediyeh, H., Sayed, T., Zaki, M. H., & Mori, G. (2014). Pedestrian gait analysis using automated computer vision techniques. *Transportmetrica A: Transport Science, 10*, 214–232.
Hillier, B., & Hanson, J. (1989). *The social logic of space.* Cambridge University Press.
Hvel, P., Simini, F., Song, C., & Barabsi, A. L. (2014). Computational models of mobility: A perspective from mobile phone data. In D. Offenhuber, & C. Ratti (Eds.). *Decoding the city: Urbanism in the age of big data* (pp. 110–124). Birkha¨user.
Jacobs, J. (1961). *The death and life of great American cities.* Vintage.
Jiang, S., Ferreira Jr, J., & González, M. C. (2015). Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. In Int. Workshop on Urban Computing.
Klein, R. W., Koeser, A. K., Hauer, R. J., Hansen, G., & Escobedo, F. J. (2016). Relationship between perceived and actual occupancy rates in urban settings. *Urban Forestry and Urban Greening, 19*, 194–201.
Kockelman, K. (1997). Travel behavior as function of accessibility, land use mixing, and land use balance: Evidence from San Francisco Bay area. *Transportation Research Record: Journal of the Transportation Research Board, 1607*, 116–125.
Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal, 79*(1), 1–14.
Kontokosta, C. E. (2016). The quantified community and neighborhood labs: A framework for computational urban science and civic technology innovation. *Journal of Urban Technology, 23*(4), 67–84.
Kontokosta, C. E. (2014). Mixed-income housing and neighborhood integration: Evidence from inclusionary zoning programs. *Journal of Urban Affairs, 36*, 716–741.
Kontokosta, C. E., & Johnson, N. (2017). Urban phenology: Toward a real-time census of the city using Wi-Fi data. *Computers, Environment and Urban Systems, 64*, 144–153.
Kostakos, V., Juntunen, T., Goncalves, J., Hosio, S., & Ojala, T. (2013). Where am I? Location archetype keyword extraction from urban mobility patterns. *PLoS One, 8*, e63980.
Leslie, E., Coffee, N., Frank, L., Owen, N., Bauman, A., & Hugo, G. (2007). Walkability of local communities: Using geographic information systems to objectively assess relevant environmental attributes. *Health & Place, 13*, 111–122.
Lukermann, F. (1964). Geography as a formal intellectual discipline and the way in which it contributes to human knowledge. *The Canadian Geographer/Le Géographe Canadien, 8*, 167–172.
Lynch, K. (1960). *The image of the city.* MIT press.
Ma, Z., & Chan, A. B. (2013). Crossing the line: Crowd counting by integer programming with local features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2539–2546.
MapPLUTO (2016). PLUTO and MapPLUTO data. New York City Department of City Planning. Available from: https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page.
Miranda-Moreno, L. F. (2011). *Pedestrian activity modelling at signalized intersections: land use, urban form, weather and spatio-temporal patterns. 90th Annual Meeting Transportation Research Board (Vol. 1)*. Washington D.C.: Transportation Research Board (TRB).
MPO (2016). Bicycle-pedestrian count program. Boston Region Metropolitan Planning Organization. Available from: http://www.ctps.org/livability.
Neckerman, K. M., Purciel-Hill, M., Quinn, J. W., & Rundle, A. (2013). Urban design qualities for New York City. *Measuring urban design* (pp. 63–82). Springer.
Owen, N., Cerin, E., Leslie, E., Coffee, N., Frank, L. D., Bauman, A. E., et al. (2007). Neighborhood walkability and the walking behavior of Australian adults. *American Journal of Preventive Medicine, 33*, 387–395.
Pantsar-Syväniemi, S., Simula, K., & Ovaska, E. (2010). Context-awareness in smart spaces. Computers and Communications (ISCC), 2010 IEEE Symposium on, IEEE, 1023–1028.
Petersen, J. F. (1994). The guidelines for geographic education: A ten-year retrospective. *Social Education, 58*, 206.
Pikora, T. J., Bull, F. C., Jamrozik, K., Knuiman, M., Giles-Corti, B., & Donovan, R. J. (2002). Developing a reliable audit instrument to measure the physical environment for physical activity. *American Journal of Preventive Medicine, 23*, 187–194.
Placemeter (2017). Placemeter: Quantify the world. Available from: http://www.placemeter.com.
Qian, C., Kats, P., Malinchik, S., Hoffman, M., Kettler, B., Kontokosta, C., & Sobolevsky, S. (2017). Geo-Tagged Social Media Data as a Proxy for Urban Mobility. *International Conference on Applied Human Factors and Ergonomics* (pp. 29–40). Cham: Springer.
Sallis, J. F. (2009). Measuring physical activity environments: A brief history. *American Journal of Preventive Medicine, 36*, 86–92.
Schneider, R., Arnold, L., & Ragland, D. (2009). Pilot model for estimating pedestrian intersection crossing volumes. *Transportation Research Record: Journal of the Transportation Research Board,* 13–26.
Schneider, R. J. (2013). *Development and application of the San Francisco pedestrian intersection volume model.* Safe Transportation Research and Education Center, UC Berkeley. University of California.
Southworth, M. (2005). Designing the walkable city. *Journal of Urban Planning and Development, 131*, 246–257.
Traunmueller, M. W., Johnson, N., Malik, A., & Kontokosta, C. E. (2018). Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities. *Computers, Environment and Urban Systems*.
Times Square (2016). Times square alliance market research data. Times Square Alliance. Available from: http://www.timessquarenyc.org.
Tresidder, M. (2005). *Using GIS to measure connectivity: An exploration of issues.* Portland, OR, USA: School of Urban Studies and Planning, Portland State University.
Tuan, Y. F. (1975). Place: An experiential perspective. *Geographical Review,* 151–165.
Villanueva, K. K.-C. (2014). The impact of neighborhood walkability on walking: Does it differ across adult life stage and does neighborhood buffer size matter? *Health & place, 25*, 43–46.
Whyte, W. H. (1980). *The social life of small urban spaces.* Conservation Foundation.
Yin, L., Raja, S., Li, X., Lai, Y., Epstein, L., & Roemmich, J. (2013). Neighborhood for playing: Using GPS, GIS and accelerometry to delineate areas within which youth are physically active. *Urban Studies, 50*, 2922–2939.
Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD International conference on knowledge discovery and data mining, ACM, 186–194.
Zheng, Y. (2015). Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data, 1*, 16–34.

Data Sources

NYC DOT (2016). Bi-annual pedestrian count data. New York City Department of Transportation. Available from: http://www.nyc.gov/html/dot/html/about/datafeeds.shtml.
PLUTO (2016). New York City Department of City Planning. Available from: https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page.
LEHD (2016). Longitudinal employer-household dynamics origin-destination employment statistics (LODES). U.S. Census Bureau. Available from: https://lehd.ces.census.gov/data/.
MTA Subway (2016). Subway stations and ridership data by Metropolitan Transportation Authority. Available from: https://data.cityofnewyork.us/Transportation/Subway-Stations/arq37z49/data.
MTA Bus (2016). Bus stops location data by Metropolitan Transportation Authority. Available from: http://web.mta.info/developers/developer-data-terms.html.
DOT Bench (2016). City bench location data by New York City Department of Transportation. Available from: http://www.nyc.gov/html/dot/html/about/datafeeds.shtml.
DOT Bicycle (2016). Bicycle facility data by New York City Department of Transportation. Available from: http://www.nyc.gov/html/dot/html/about/datafeeds.shtml.
Single line street map (LION) (2015). New York City Department of City Planning. Available from: https://www1.nyc.gov/site/planning/data-maps/open-data/dwnlion.page.
DOITT (2015). Sidewalk area map data by New York City Department of Information Technology and Telecommunications. Available from: https://data.cityofnewyork.

us/City-Government/Sidewalk/vfx9-tbb6/data.
NYC Park (2015). Park properties Shapefile data by New York City Department of Parks and Recreation. Available from: https://data.cityofnewyork.us/City-Government/Parks-Properties/rjajzgq7/data.
NYC Tree (2015). Street tree census by New York City Department of Parks and Recreation. Available from: https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/pi5s-9p35/data.
NYPD (2015). 7 major felony incidents. New York City Police Department. Available from: https://data.cityofnewyork.us/Social-Services/nypd-7-major-felony-incidents/k384-xu3q/data.
NYC 311 (2015). 311 service requests. New York City 311. Available from: https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9.
Weather Underground (2016). Historic hourly weather data. Data queried from - https://www.wunderground.com.