Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities

Martin W. Traunmueller[a,*], Nicholas Johnson[b,*], Awais Malik[c,*], Constantine E. Kontokosta[c,*]

[a] *New York University Center for Urban Science & Progress, United States*
[b] *University of Warwick & New York University Center for Urban Science & Progress, United States*
[c] *Dept. of Civil & Urban Engineering, Center for Urban Science & Progress, United States*

ARTICLE INFO

Keywords:
Modeling urban mobility
WiFi probe data
Spatial network analysis
Big data

ABSTRACT

City governments all over the world face challenges understanding mobility patterns within dense urban environments at high spatial and temporal resolution. While such measures are important to provide insights into the functional patterns of a city, novel quantitative methods, derived from ubiquitous mobile connectivity, are needed to provide policy-makers with better insights to improve urban management and planning decisions. In this paper, we develop a model that uses large-scale WiFi probe request data to model urban mobility trajectories in dense urban environments. We collect probe request data from a public Wifi network with 54 access points in the Lower Manhattan section of New York City over one week, accounting for more than 30 million observations and over 800,000 unique devices. First, we aggregate unique entries per access point and per hour, demonstrating the potential to use WiFi data to approximate local population counts by type of user. We then use a spatial network analysis to identify edge frequencies and directions of journeys between the network nodes, and apply the results to the road and pedestrian sidewalk network to identify usage intensity levels and trajectories for individual street segments. We demonstrate the significant potential in the use of WiFi probe request data for understanding mobility patterns in cities, while highlighting non-trivial issues in data privacy raised by the growing availability of public WiFi networks.

1. Introduction

With an annual growth of 60 million new city dwellers every year (U. WHO, 2010), the world is experiencing a rapid population shift of people moving from rural areas into urban environments over the last several decades. Driven by technological innovations and increasing economic opportunities (Dargay, Gately, & Sommer, 2007), this situation has led to a steady increase in motorized and pedestrian mobility activity in cities all over the world (Millard-Ball & Schipper, 2010). For city governments, this increased demand has lead to challenges in managing city services and infrastructure, and in maintaining quality–of–life standards for its population, as congestion and overcrowding of areas can negatively affect the city's economy (Sweet, 2014), sustainability (Zhao, 2014) and its population's health (Hansson, Mattisson, Bjoerk, Oestergren, & Jakobsson, 2011).

To address these challenges, city managers need to understand patterns of urban mobility to enable targeted and "smart" interventions to limit overcrowding, improve service delivery, and ensure effective emergency response. In many cases, methods to measure mobility dynamics focus on reporting traffic counts at specific points in the city at discrete times, typically using rather simple technologies (Slack, 2017) that are limited in terms of scalability and real–time feedback, and that can be cost–intensive when applied to large areas. With the rise of remote and in–situ sensing technologies, the analysis of closed–circuit–television (CCTV) footage using computer vision machine learning techniques offers a new, and increasingly popular, approach for computer scientists and urbanists to count not only motor, but also pedestrian traffic on a large scale (Slack, 2017).

However, these "counting–gate" methods are limited to traffic counts at specific locations for a specific time period, and thus they do not offer data about trajectories of pedestrians between them. Current work in data mining aims to fill this gap by using mobile phone data to model urban mobility (Calabrese, Diao, Di Lorenzo, Ferreira Jr., & Ratti, 2013; Jiang et al., 2016), but shows limitations in terms of population representation by capturing only mobile users of a specific network provider, and typically with low spatial granularity. In

* Corresponding authors at: New York University, Center for Urban Science & Progress, 370 Jay St, 12th Floor, Brooklyn, NY 11201, United States.
E-mail addresses: martin.traunmueller@nyu.edu (M.W. Traunmueller), nicholas.johnson@nyu.edu (N. Johnson), awais.malik@nyu.edu (A. Malik), ckontokosta@nyu.edu (C.E. Kontokosta).

https://doi.org/10.1016/j.compenvurbsys.2018.07.006
Received 12 December 2017; Received in revised form 21 July 2018; Accepted 21 July 2018

addition, computer vision techniques create significant concerns around confidentiality and privacy, as facial recognition methods become more widely applied.

Data that are independent of specific network providers are able to capture a larger sample of the population at any given place and time. One example of this is smart device probe requests to WiFi access points (APs) in public urban space. With an increasing number of public WiFi APs and networks in cities, these networks can provide dense coverage across the cityscape, particularly at the neighborhood or district scale. Each AP continuously “senses” its surroundings in terms of potential users equipped with WiFi–enabled mobile devices, which send probe requests to available networks and proximate APs at regular frequencies. With the increasing market penetration of WiFi connectible mobile devices, such as tablets or smartphones (64% of all U.S. citizens and approximately 80% in New York City owned a smartphone in 2016 (Smartphone Users, n.d.)), computer and urban scientists have begun to use WiFi probe data with the aim to understand human behavior and mobility (Kontokosta & Johnson, 2017). However, while many of the large–scale studies focus on indoor activities (Abedi, Bhaskar, & Chung, 2014; Meneses & Moreira, 2012), less has been done discussing every–day movement patterns in open public spaces at the neighborhood scale using Wifi probe data. This has largely been the result of the lack of data available at necessary spatial and temporal granularity, and the computational challenges in processing these data. Mobility data, however, must be handled with appropriate data management and access protocols, as concerns about data privacy become paramount. In addition, using such data can also raise equity issues, as sampling bias caused by differential access and technology adoption rates can exclude certain demographic groups (Kontokosta, Hong, & Korsberg, 2017).

In this work, we hypothesize that WiFi probe data can be used to analyze outdoor mobility and human trajectories in a large and densely populated urban area at high spatial resolution and temporal frequency. We use a dataset of WiFi probe requests collected by 54 public-access WiFi APs over the duration of one week in Lower Manhattan in New York City, NY, collected through the” Quantified Community” urban test–bed (Kontokosta, 2016). First, we show how WiFi probe data can be used to report hyperlocal, real-time counts at each AP, similar to “counting gates” methods described above, and be used to understand localized population segmentation. Second, we conduct network analysis to describe a spatial network that can be applied to street and sidewalk segments. We demonstrate how these data can be used to analyze common paths of travel and trajectories, indicating the intensity of street activity over time. We begin by presenting recent literature on measuring urban mobility, and then present our data and data processing steps. We introduce our methodology and describe our results for pedestrian counts and trajectories. We conclude with an in–depth discussion of the findings, including limitations, privacy concerns, and applications to city management and planning.

2. Literature review

2.1. Capturing urban mobility

The most commonly used method to capture urban mobility by city agencies is the installment of “counting–gates” at pre–defined locations, such as intersections or heavily–used main roads. While technology has improved over the years, the method has remained relatively the same by using, for instance, pneumatic road tubes, Piezo–electric sensors or infrared sensors (Slack, 2017) to count primarily motor traffic. While these methods offer an easy way to quantify traffic aggregations on a street for a specified time period, such as per hour or day, they are limited in terms of temporal and geographical scalability and rather expensive to run due to installation and service charges, compared to the output they provide.

More current work uses advances in computer vision to analyze closed–circuit–television (CCTV) feeds to count motorized and pedestrian traffic at lower costs. In doing so, researchers and city governments are now able to count traffic at places with CCTV–coverage, like high–volume intersections, by applying computer vision algorithms, such as blob–detection (Trafficvision, n.d.). Focusing primarily on motor traffic, this approach has been extended over the years to also count pedestrians (Placemeter, n.d.). The analysis of CCTV footage offers effective ways to aggregate traffic quantitatively and is only limited by the number of CCTV–camera locations (with appropriate resolutions and fields of view) in a city. As the usage of CCTV cameras in the urban environment is growing due to congestion and security concerns, the method becomes increasingly applicable to count traffic on a large scale. However, in focusing on traffic counts, it does not offer any insight into the routes people take between their locations and provides little ancillary information about activity patterns, and hence do not generate critical information for city managers.

The increasing availability of open data has offered researchers novel opportunities to study traffic routes, in particular for public transport, on a large scale using a data mining approach. In doing so, metro journeys (Tfl Study, n.d.), the use of public bike sharing schemes (Woodcock, Tainio, Cheshire, & Goodman, 2014), or GPS traces of taxis (Ferreira, Poco, Vo, Freire, & Silva, 2013), for instance, have been visualized and the time–dependent frequencies of routes through cities detected.

While the results of such studies can contribute to the efficiency of public transport systems, these open data sources do not include information about the population who do not use public transport. As many people in U.S. cities travel by car or increasingly walk (Milne, 2014), using these data sources excludes a large portion of the urban population and are therefore not fully representative. The focus on individual transport modalities also limits valuable information about human behavior and activity at the micro- and meso-scales in various urban environments.

A data source that includes these populations are call detail records (CDR). With the increased use of mobile phones over the last decade, CDR data from mobile phone providers have become a popular source for urban mobility research. For instance, (Yuan & Raubal, 2012) extracted dynamic mobility patterns in urban areas using a ‘Dynamic Time Wrapping’ algorithm, and were able to classify areas according to the observed patterns. (Calabrese et al., 2013) combined mobile phone traces and odometer readings from annual vehicle safety inspections to map mobility as averaged individual total trip lengths for the case of Boston. In doing so, researchers found, for instance, that the two most important factors for regional variations in mobility are accessibility to work and non–work destinations, while population density and mix of land–use showed less significance. Other work uses CDR data to model urban flows. (Gonzalez, Hidalgo, & Albert-Laszlo, 2008), for example, studied 100.000 mobile phone user trajectories over six months and found that human trajectories show a high degree of temporal and spatial regularity. Furthermore, findings suggest that humans follow rather simple, reproducible mobility patterns.

These studies demonstrate the opportunities for using CDR data to study human mobility at the urban scale. However, at the same time, telecommunication data can be sensitive and often difficult to access for researchers. One possible way to gain access to such data is to take part in a data mining challenges (Competition Example, n.d.) where providers make parts of their data publicly available. However, as available data are pre–processed, their accuracy often suffer due to unknown data processing steps (Traunmueller, Quattrone, & Capra, 2014). Furthermore, when data are provided by individual mobile phone providers, they offer limited representativeness of the urban population by excluding various groups of people that use other providers, or people that do not have a cell phone contract, such as the elderly or lower–income populations, or those that use Pay–As–You–Go options.

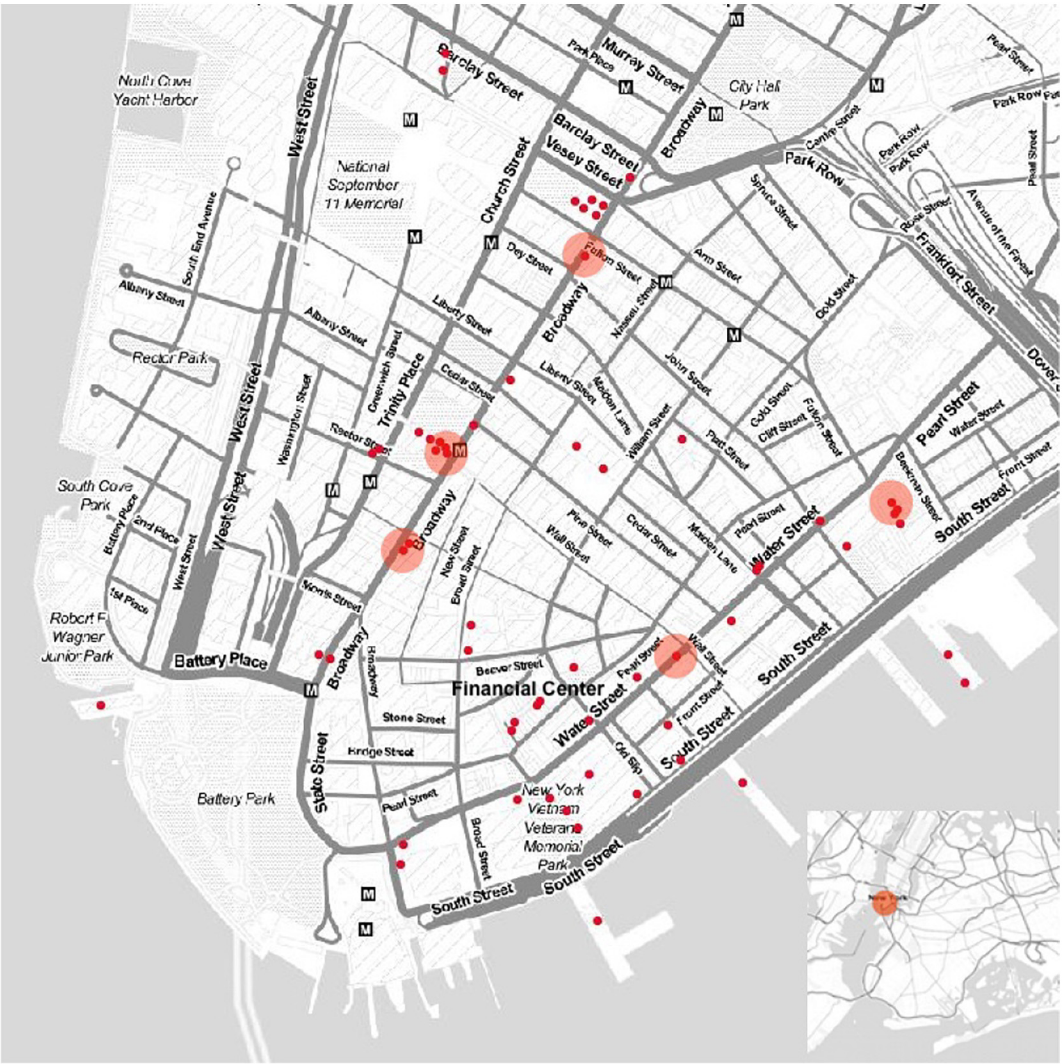


Fig. 1. WiFi AP locations in Lower Manhattan, New York City. Highlighted APs are compared to official counts.

2.2. WiFi data and urban mobility

A dataset that provides a more complete representation of the urban population is WiFi probe data, which is provider independent. This has been demonstrated by (Kontokosta & Johnson, 2017), who developed a dynamic alternative to census data by estimating real–time population by different categories (e.g., workers, residents, visitors) for a similar testbed in New York City as used in this paper.

By default, a mobile device is configured to steadily scan its environment for available WiFi APs and continues to transmit data, including time, geolocation and the device MAC address. Such data have been used by researchers to model people's behaviour, such as queuing activity and movement in indoor environments (Wang et al., 2013). An early study (Sevtsuk, Huang, Calabrese, & Ratti, 2006) used WiFi data of logged–in users, collected via 3000 AP within MIT university buildings, to visualize and detect activity patterns and hence, draw relationships to mobility within the campus. Also using the university environment as a testbed, (Meneses & Moreira, 2012) were able to describe movement of people in various academic buildings on two campuses, using data from more than 550 APs over a duration of several months and highlighting the relationship between spatial properties found in the environment and resulting human motion. (Abedi et al., 2014) use a similar approach to discuss complex spatial–temporal dynamics of movement in indoor shared zones, such as lounges and office areas, at another university and were able to extract different staff usage patterns of space (such as, frequency, duration, and utilization peaks). Another study (Prentow, Ruiz-Ruiz, Blunck, Stisen, & Kjrgaard, 2015) uses WiFi data to derive information to improve facility management and planning for large building complexes, such as a hospital, by quantifying area densities and flows.

Other work extends the usage of WiFi data to capture mobility for outdoor spaces. Following a probabilistic approach that uses visited destinations according to WiFi traces, (Danalet, Bierlaire, & Farooq, 2012) show the potential of WiFi data to predict user destinations and routes. More recent work (Sapiezynski, Stopczynski, Gatej, & Lehmann, 2015) uses WiFi data collected through a large scale study to capture human mobility for the case of Copenhagen. Based on a previous scientific study, the data contain a high level of information about the users and uses of data for logged–in users. In reality, WiFi probe data that have been collected 'in the wild' are less detailed and the relative number of people logging–in to public WiFi networks is limited. These circumstances lead to questions about the suitability of an approach relying on authentication of the network device for quantitative mobility modeling. (Chilipirea, Petre, Dobre, & van Steen, 2016) show in their work the potential of WiFi data that has been collected throughout a 3–day festival. However, as such festivals attract uncommon masses of people for its duration, the study does not represent a normal day in a city and, hence, it is unclear how the approach applies to typical situations.

A recent study (Tfl Study, n.d.) shows the usability of WiFi probe data to detect different routes and trains people take to reach their destination on the London Underground network. We show in this paper how similar data, collected by public Wifi APs without the need for network log-in, can be used to model urban mobility on a large scale and in a densely populated, unconstrained public area for a” normal” week in New York City.

3. Methodology & results

3.1. Dataset description

The method we propose requires access to WiFi probe request data that contains information about detected client devices and their proximity to surrounding WiFi APs at a specific time. For this study, we use a dataset provided by the ‘Alliance for Downtown New York’ (Downtown Alliance, n.d.), the non–profit management entity for Downtown NY–Lower Manhattan's Business Improvement District. Our dataset includes observations from 54 WiFi AP locations throughout the study area covering the whole of Lower Manhattan south of Barclay and Franklin Streets, and includes APs on two of the East River piers, as shown in Fig. 1. With its mix of high density residential, retail, and commercial buildings (Land–use, n.d.), Lower Manhattan represents a very dynamic and diverse environment, allowing us to observe and model mobility behavior of different groups of people as they activate the area at different times throughout the day and the week.

The majority of APs are located on building rooftops (approximately 2–3 stories high) and capture WiFi enabled devices, regardless of whether or not the device is associated with the network, within a distance of 100 ft. or more, depending on the environment (Johnson, Bonczak, & Kontokosta, 2018).

Spatially, APs are densely distributed along Broadway (especially at Vesey and Wall Streets) and in the Financial District towards the East River, focusing on more active corridors, such as Wall Street and John Street. Density decreases towards the primarily residential area at the West-side of Lower Manhattan. Locations for each AP were stored in a separate Shapefile by the provider and were visualized on the NYC street grid using mapPLUTO (mappluto, n.d.). Data were collected from Friday, 2017/04/14, 0.00 AM to Friday, 2017/04/21, 12.00 PM, containing a total of 30,862,317 data points collected by 1,175,039 unique users. Each data point includes the MAC address of the observing AP, the client device's MAC address, the device's received signal strength indicator as seen by the AP (rssi), and the observation time.

3.2. Data pre–processing

Mobility data such as our WiFi probe dataset are highly sensitive given the provision of MAC addresses of clients. Combined with logged WiFi AP locations, such data can lead to privacy issues as it opens opportunities for tracking individuals. To ensure anonymity, we first de–identified our dataset by replacing MAC addresses of clients with an anonymous identifier, consisting of a unique incremental integer starting from 1. Following the addition of the unique identifier, the original MAC address is deleted, and no linkage keys are provided to re–merge the unique identifier and the MAC address. In doing so, the data could not be traced back to the individual and hence, would eliminate many underlying privacy concerns. (The study received NYU IRB approval with IRB No.: #IRB-FY2017–526.)

Using the de–identified dataset, we conduct a descriptive analysis to identify possible missing inputs or other data errors to ensure the completeness of the data for our analysis. We identify 628 data points that were counted multiple times, and 3019 data points that were missing geographical or temporal information. We remove these data points from the dataset. Furthermore, we remove close to 1 million entries from our dataset that were captured by multiple WiFi APs simultaneously, as this would confound our model and bias results. To minimize spatial inaccuracy for these simultaneous entries, we retain the observation with the highest rssi–value, indicating the closest WiFi AP location to the device, and exclude the remaining entries.

Next, we detect and remove all devices captured in the dataset that exhibit connection patterns indicative of stationary or fixed devices, as they would skew the mobility model by increasing an AP's use frequency. We examine the connection patterns for each client and observe that while a majority of clients connected to various APs over the study period – indicating movement – some clients were only captured by a single AP during the entire period. These data are removed from the dataset. We then identify and remove data points that showed no movement for longer than 20 min, the average walking time for the maximum distance between all APs within the testbed (1.3 miles), as suggested by (Barton, Grant, & Guise, 2003). Such data are considered to represent non-active users and are removed from the dataset as they would skew the mobility model by increasing an AP's use frequency. Overall, we exclude a total of 15,987,026 (51.80%) observations from our raw data resulting in 14,875,289 observations from 835,797 unique clients. We then joined the location and observation datasets by assigning AP addresses to each observation, according to the individual AP ID number.

The most common way for city governments to model urban mobility is the installment of “counting–gates”, providing traffic aggregations at specific locations for a defined temporal unit. By scanning and detecting WiFi–enabled devices in hyper–local environments, we show how WiFi APs can be used in a similar way, and enable deeper analyses of trajectories and activity patterns. First, to detect variances depending on time of day and differences between weekday and weekend, we aggregate the total number of captured clients per hour of the day across all APs. Fig. 2 presents these distributions and differences in client counts between weekend (2017/04/15 and 2017/04/16) and weekday are clearly visible. From a cursory examination of the client activity patterns, we find that, as expected, counts are higher during the work week than on weekends. Weekday distributions show three peaks of increased client activity, one in the morning (6.00 AM–9.00 AM), one during lunch time (12.00 PM–1.00 PM) and one in the early evening (4.00 PM–6.00 PM). On the weekend, the distribution was found to be smoother, increasing from the later morning (9.00 AM) and decreasing in the early afternoon (5.00 PM). We interpret these observations as reflections of the predominant commercial land use of Lower Manhattan, which is characterized by office workers commuting to and from the area in the morning and evening and going to restaurants near–by during the lunch hour. On the weekends, the area attracts many visitors and tourists, due to its landmarks (such as the New York Stock Exchange or One World Trade Center) and retail establishments that populate the area throughout the day. At the network level, these results reinforce the findings of (Kontokosta & Johnson, 2017).

3.3. Aggregation

Next, we aggregate these client observations per AP and per day to show the usefulness of WiFi probe data for population counts at specific

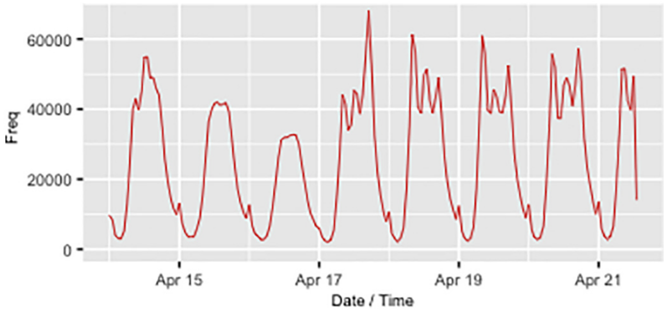


Fig. 2. Weekly distribution of WiFi requests over all APs for the cleaned dataset.

Table 1
Comparison between official manual daily pedestrian counts and WiFi probe requests.

Location	Manual counts	WiFi requests	Offset %
Broadway at Pine Street	20,431	18,183	−11.37
Broadway south of Fulton Street	37,158	34,845	−6.33
Broadway and Exchange Place	17,842	19,978	10.36
Water and Wall Street	7447	6885	−9.21
Water and Fulton Street	22,751	20,157	−11.1

locations. We validate our aggregations with manually-collected daily pedestrian counts at various locations as defined by the Downtown Alliance (Downtown Alliance, n.d.). As these locations are not necessarily co–located with our WiFi AP locations, we select five locations that are spatially proximate to specific AP locations. Fig. 1 highlights these locations with a larger red dot, indicating three contiguous APs on Broadway (at Fulton, Wall Street and Exchange Place) and two along Water Street (at Fulton and Wall Street).

Table 1 shows the result of the comparison between the manual counts per 24 h and our average counts per 24 h of unique clients per AP. We find that our numbers vary between 6 and 11% from the manual counts and tend to under–count individuals. This difference can be explained by several factors. First, it is unclear of the exact methodology used for the manual counts, and manually counting is, in general, an unreliable measure given the potential for human error. Therefore, it is difficult to ascertain whether the observed variance, especially between over– and under–counting, is caused by errors in the WiFi method or in the manually counting process. Second, the under–counting found in the WiFi observations could be a function of the prevalence of WiFi–enabled devices. As was mentioned, approximately 80% of the NYC population has a smart device, and it is possible that the under-counting reflects this penetration rate (although we would assume, given the demographic and socioeconomic characteristics of the population in the study area, that WiFi–enabled device use would be higher than the city-wide average). This effect may also be minimized because of individuals with more than one smart device, who may be counted more than once. However, while these issues lead to some uncertainty in interpretation, this result shows the potential for using WiFi probe request data to estimate the number of people in specific areas, similar to "counting–gate" approaches. In addition, this method has numerous advantages over other counting methods, including being able to identify unique visitors to a particular location and, as we will demonstrate, paths of travel.

To detect spatial variance, we next map the WiFi–derived count aggregations based on individual AP locations per total day, from Friday, 2017/04/14 to Thursday, 2017/04/20. We observe in Fig. 3 geographic differences in usage intensities by AP location depending on the day of the week. The increased activity on weekdays compared to the weekend is clearly visible and highlighted by the proportional dot plot for each AP location. Analysis of the spatial distributions reveal three "hot–spots" of increased activity located along Broadway, in the Financial Center south of Wall Street, and at the intersection of Water Street and Maiden Lane on the East Side. This pattern again reflects the land–use of the immediate area: in addition to commercial buildings driving weekday population, the area attracts a large number of tourists and visitors to destinations such as Wall Street, Trinity Church and historic Front Street at the East River. Broadway is one of NYC's main north–south traffic routes in Manhattan, and attracts both locals and visitors due to high retail density and substantial public transit infrastructure.

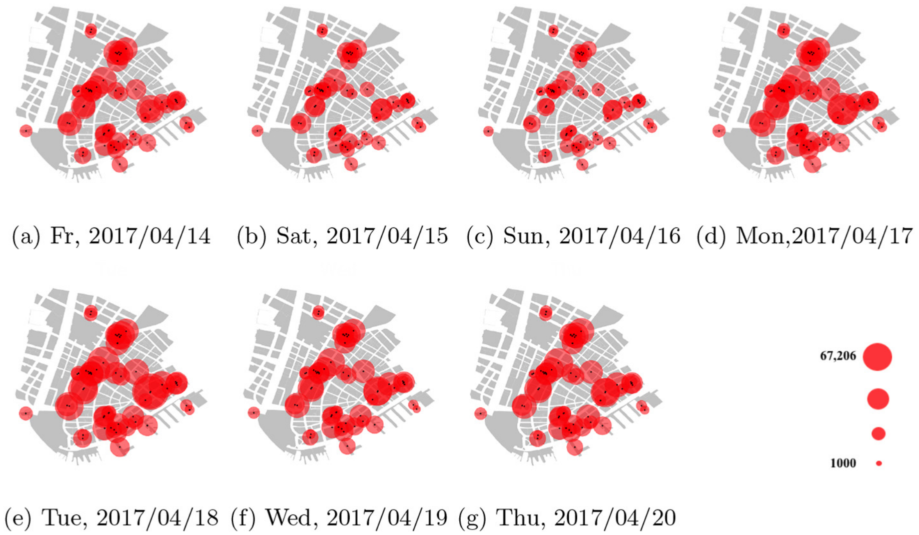
These results show how WiFi APs can be used to measure population counts at given locations, offering a more efficient and comprehensive alternative to common approaches as described previously. However, by measuring client aggregations at WiFi locations, one can only assume mobility patterns (as for instance, if the movement is linear or clustered). Therefore the output offers only a limited picture as it does not include any information about *where* or *how* people move between these locations. Looking at differences in the land–use of the urban environment and calculated WiFi counts, we can identify two types of urban mobility patterns in the area. On Broadway, we would expect a predominant linear pattern along the street from north to south (or vice versa). In the historic Financial District, with its web of small streets and large office buildings, we expect a more diffuse network pattern of movement, intersecting more street segments and offering more options for route–finding based on individual preference. We explore these spatial trajectories in detail in the next section. We demonstrate how the same WiFi probe data can be used to not only to generate aggregations at specific points, but also to evaluate actual street network usage by capturing movements between APs, allowing us to model urban mobility at the individual street segment level.

3.4. Trajectories between WiFi AP locations

Applying methods from network science, we turn our set of client activities into a graph, examining which nodes (or vertices) are connected to each other via edges. By considering all journeys, we create a network where the nodes are the WiFi AP locations, and the edges are the client flows of consecutive AP–entries in each direction between them, as defined by our 835,797 clients, and weighted by the number of journeys carried out on each edge. Results show a highly connected network of 54 WiFi APs on most of the 2916 edges with an average degree of 38. A majority (78%) of all journeys were direct between departure and destination point, without passing intervening nodes.

As indicated above, we would expect a high variance in activity between *Weekend* and *Weekdays*, as well as during different times of the day, such as *Early Morning* (0.00 AM–6.00 AM), *Morning* (6.00 AM-12.00 PM), *Afternoon* (12.00 PM–6.00 PM) and *Night* (6.00 PM-12.00 AM). Fig. 4 presents the network graphs for each of these time bins with edge frequencies for each journey between the WiFi AP node. Frequencies show a long–tail distribution: while the most–frequented journey of the week has 781,073 trips, the least–frequented journeys are single trips between two WiFi APs and the mean frequency (J_m) among observed journeys overall is 102,452 ($Q_1 = 58,677$, $Q_3 = 164,988$). In the graph, we indicate the frequency of journeys between the nodes by the width of the edges – the wider an edge, the more journeys detected, ranging from 1 to 21,387 journeys per day. As indicated from our aggregations, we observe a high variance in mobility patterns between *Weekend* and *Weekdays*: more journey activity is detected throughout the week ($J_m = 161,191$) compared to the weekend ($J_m = 73,087$), when, in addition to visitors, a high number of office workers populate the area. Furthermore, we observe that weekday journeys start earlier and end sooner when compared to weekends. While on weekdays many journeys are detected in the *Morning* (46%, as people commute to work) and less at *Night* (24%, as people may go to sleep earlier), we find the opposite pattern for weekends due to reduced activity in the *Morning* (21%, as people might sleep later) and more activity at *Night* (37%, as people might partake in social activities).

Looking at the type of movement found in the area, the network graph supports our prior assumptions of linear (Broadway) and clustered movements (Financial Center), showing a similar pattern for both weekends and weekdays. While results indicate the highest journey frequencies ($J_m = 324,096$) along Broadway (as nodes on Broadway/Fulton Street in the north are mostly connected to nodes at Broadway/Wall Street in the south), we observe journeys with lower, but in all directions evenly distributed, frequencies ($J_m = 54,096$) for nodes in the Financial District (for instance, Water Street/Maiden Lane is accessed from all sides). As described above, we believe that this outcome might result from differences in the scale of the built environment, such as the smaller block sizes in this area, allowing more options for route–finding. We also can see that WiFi APs located at the piers of the East River, where water taxis stop, show stronger connections to APs in the



(a) Fr, 2017/04/14 (b) Sat, 2017/04/15 (c) Sun, 2017/04/16 (d) Mon,2017/04/17

(e) Tue, 2017/04/18 (f) Wed, 2017/04/19 (g) Thu, 2017/04/20

Fig. 3. Daily client frequencies per WiFi AP, as indicated by circle size. (a) Fr, 2017/04/14 (b)Sat, 2017/04/15 (c) Sun, 2017/04/16 (d) Mon,2017/04/17 (e) Tue, 2017/04/18 (f) Wed, 2017/04/19 (g) Thu, 2017/04/20.

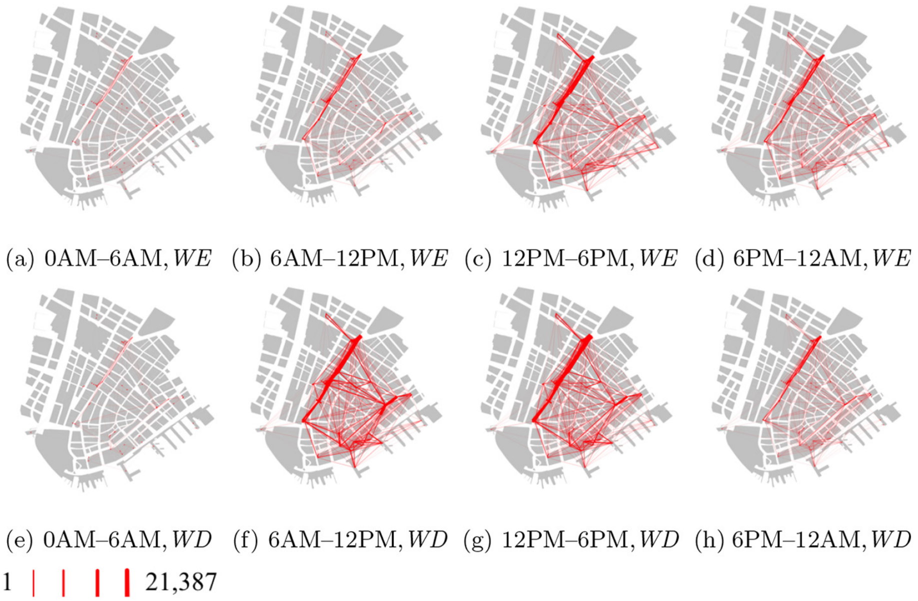
Financial District, as for instance on weekday mornings, when 62% of all journeys from these APs lead into the Financial District. We interpret this result as indication that people coming to the area by ferry are most likely commuters going to and from work.

Different patterns of movement in the area become more clear when plotting the directions of these journeys, such as where people come from and where they go, to detect attractors in the area.

In Fig. 5, we show for the busiest time bin (Weekday Mornings, 6 AM–12 PM) the direction of each journey, as indicated by saturation of blue arrows, and their frequencies, as indicated by edge width in red: resulting blue patches indicate nodes of destinations within the network. The output shows the dominance of linear journeys along Broadway in both directions, north to south and south to north, indicated by the larger red edge width and more directed blue arrow heads between WiFi APs along Broadway. The two locations at the corner of Broadway/Wall Street (in–degree $D_i = 33$, out–degree $D_o = 31$) and Fulton Street ($D_i = 31$, $D_o = 29$) show strong connections mainly to other locations along the street segment, which might be the result of two proximate subway stations. In contrast, we see how the Financial Center (for instance, Water Street/Old Slip; $D_i = 42$, $D_o = 34$ as shown in Fig. 1) attracts people from multiple origins, as indicated by thinner edges with a lower saturation of arrows, but coming from multiple angles. This outcome might result from office and business locations and their employees coming from public transport hubs, such as the Water Taxi landings on the East River piers ($D_i = 7$, $D_o = 25$) and subway stations.

To see how these movements affect the usage of street segments, we next applied our graph to the street network. We used street network data that are publicly available for New York City (New york city streets, n.d.) and overlay it on our AP locations. As most of the locations were offside road vectors, we first assigned them to the nearest road segments, using the k–nearest neighbor classification algorithm (Cover & Hart, 1967). Having assigned APs to road segments allowed us to then generate the shortest path journeys between each WiFi AP location as described by the network graph on the street segment level, using the breadth–first search algorithm (Lee, 1961).

In Fig. 6, we show for each of our defined time bins the outcome of our model. The opacity of red indicates the usage intensity for each segment, normalized by day and week: the more opaque, the greater number of journeys on that road segment. Compared to Fig. 4, we now



(a) 0AM–6AM, WE (b) 6AM–12PM, WE (c) 12PM–6PM, WE (d) 6PM–12AM, WE

(e) 0AM–6AM, WD (f) 6AM–12PM, WD (g) 12PM–6PM, WD (h) 6PM–12AM, WD

Fig. 4. Network graphs of client activities between WiFi APs, for Weekends *WE* (a-d) and Weekdays *WD* (e-g). Edges widths represents frequencies between the nodes. (a) 0 AM–6 AM, *WE* (b) 6 AM–12 PM, *WE* (c) 12 PM–6 PM, *WE* (d) 6 PM–12 AM, *WE* (e) 0 AM–6 AM, *WD* (f) 6 AM–12 PM, *WD* (g) 12 PM–6 PM, *WD* (h) 6 PM–12 AM, *WD.*

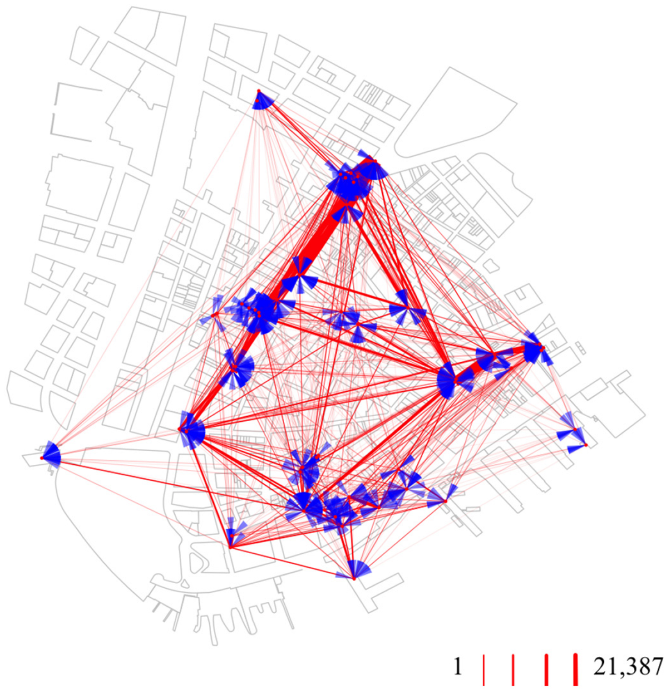


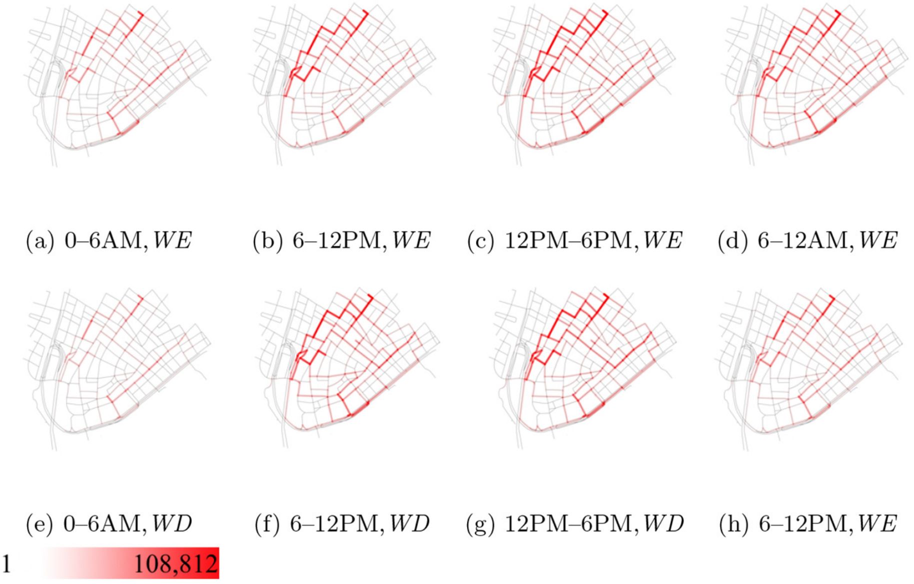
Fig. 5. Network graph, showing directions of movements between nodes for Weekday Mornings, 6 AM–12 PM, indicated by saturation of blue arrows, and their frequencies, indicated by edge width in red: Blue patches indicate nodes of destinations within the network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

see how people use roads for their journeys between AP locations, based on the shortest path algorithm. We observe that there are no major changes in routing between weekend and weekdays or different times of day, reflecting the result from our network graph. In addition, we see that there is a strong relationship between the number of detected routes and the time of day, as we observe an increase throughout the day compared to nighttime hours. Overall, the results confirm the importance of Broadway and near–by parallel roads (such as Trinity Place and Church Street) as pedestrian routes. We see that the increased activity throughout the Financial District is concentrated along Water and Pearl Streets, connecting most of the AP locations along the East River. Wall and Broad Streets also show a high frequency of journeys, which we interpret as the preferred routes between the ferry landing points at the East River piers and the office buildings in the Financial District. As a preferred east–west connector between Broadway and Financial Center, we can identify Pine and Fulton Streets, showing the highest frequencies compared to other parallel streets.

These results of this street intensity analysis reflect official statistics (Downtown Alliance, n.d.) on the most-used streets in the area, such as Broadway and Fulton Streets, and support our hypothesis that WiFi probe data can be used to estimate high spatiotemporal resolution counts of pedestrians, but also to model street usage intensity and paths of travel in a time–efficient and cost–effective way. Next, we will discuss the limits of our approach, potential applications, and future research.

4. Discussion and implications

In this paper, we have presented a novel method to model urban mobility in public space on a large scale using a data mining and network analysis approach. The method requires access to WiFi probe data, including information about client activities and location information of APs, which are becoming increasingly available as the number of public and municipal WiFi networks grows. From the dataset, we extracted aggregations about client frequencies per WiFi AP location, showing the method's potential to capture similar data as commonly-used “counting–gates”, while employing a more efficient process that also harnesses data to model pedestrian flows. We then apply these WiFi probe data to a network graph and overlay open source street boundary geodata to evaluate journeys and their frequencies at the street segment level based on the shortest routes. Typically, network analyses assume a lower level of connectivity requiring intermediate nodes (in subway networks, for instance) and therefore include betweenness centrality measures of nodes and edges located on the shortest paths between nodes. Our network of 54 WiFi APs is highly connected on most of its 2916 edges (mean node degree of 38). Unlike subway networks, our public WiFi AP network does not require clients to pass intervening nodes due to its representation in the physical environment (in our case 78% of all journeys were direct between departure and destination point, without passing intervening nodes), so this measure was not considered. As publicly-available and municipal WiFi networks in cities are continuously expanding, the use of WiFi probe data to model urban mobility is becoming more precise. This development suggests that the proposed methodology will become increasingly applicable in the coming years.


(a) 0–6AM, *WE* (b) 6–12PM, *WE* (c) 12PM–6PM, *WE* (d) 6–12AM, *WE*

(e) 0–6AM, *WD* (f) 6–12PM, *WD* (g) 12PM–6PM, *WD* (h) 6–12PM, *WE*

Fig. 6. Street network showing road usage per segment for Weekends *WE* (a-d) and Weekdays *WD* (e-g). The opacity of red indicates the frequency usage for each street segment: the more opaque, the more journeys happen on the road segment. (a) 0–6 AM,*WE* (b) 6–12 PM,*WE* (c) 12 PM–6 PM,*WE* (d) 6–12 AM,*WE* (e) 0–6 AM,*WD* (f) 6–12 PM,*WD* (g) 12 PM–6 PM,*WD* (h) 6–12 PM,*WE.* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.1. Limitations

We acknowledge that this work is constrained by a number of limitations. First, our method is highly dependent on the overall number and density of AP locations within the network. The higher the density of APs, the more spatially accurate the model's performance can become. In our case, we have shown the potential of the approach for cities with a relatively high penetration of WiFi APs, such as New York City. However, even within this testbed, we observe differences in AP distribution (as for instance a higher AP density was found along Broadway, compared to secondary streets) that might affect the outcome, such as linearity vs. clustering. We do not know how the same approach would perform in other cities with a lower density of WiFi APs, as for instance, in less developed countries. As cities in developing countries show the highest rise in urban population (D. o. E. United Nations, 2014), our method could contribute to urban planning in such environments, which typically lack more formal infrastructure to collect similar data, such as public transit RFID cards.

In addition to AP network density, the method we propose is highly dependent on the accuracy of collected data. Accuracy could suffer, for instance, due to shielding problems (e.g. signals impeded by buildings) especially in dense urban areas. Therefore, researchers using this method need to be aware of WiFi APs effective capture radii, and how to account for variations in rssi. In our case, we also had to exclude a rather large part – approximately half – of our data points that were identified as stationary devices (such as a desktop computer in an office building) to ensure the model's accuracy. Such devices are constantly detected by the same surrounding WiFi APs, and hence captured in the dataset throughout the day. Accurately detecting stationary devices is thus critical to the accuracy of the mobility model.

Representativeness is also a concern when using WiFi probe data. As described in the literature, we can expect a high capture rate in the case of New York City due to a high penetration of WiFi-enabled devices (Smartphone Users, n.d.). Although we validate our results against manual," ground-truth" pedestrian counts, which indicate our model's relative performance, we are unable to quantify with certainty the exact number of unaccounted for individuals. This limitation can lead to bias in the model outputs, particularly as it pertains to understanding the mobility of children, the elderly, and other population sub-groups less likely to possess a WiFi–enabled device.

In this work, we do not differentiate between pedestrian or motorist. As some roads are for motor traffic only, this might lead to inaccuracies in route generation. Furthermore, here we apply the shortest–route algorithm to generate our result. As people do not always follow the shortest route on the street network or take the subway, we need to consider the application of other routing algorithms to be more suitable and detect users of the public transport.

Finally, we use data from a relatively short time period in the Spring 2017. While we believe this analyzed dataset is sufficient to develop and test our methodology, it is not possible to generalize observed patterns in counts or trajectories. Therefore, our interpretation of the results is presented as a descriptive analysis of mobility behavior to present the potential application of our approach.

4.2. Implications

We consider the potential of our approach to benefit a number of user groups either researching urban mobility patterns or operating and planning urban infrastructure and transport systems. These include:

Researchers in urban and computational studies can use the method to model and analyze people flows in public space and detect to what degree urban mobility patterns are impacted by the design and condition of the urban environment. In doing so, the output can be used to describe possible relationships between human activity and urban phenomena, such as crime, congestion, or property values, by integrating other data sources, such as localized crime data, road construction locations, or house prices.

Other researchers can use the method to further understand the relationship between urban mobility and land–use, and discuss opportunities to identify and describe communities based on their movement patterns and behavior. Our method, and the type of data used here, can enable new investigations into the dynamics of socio–economic differentiation in a city. Results of such work could further reinforce exploratory studies (Kontokosta & Johnson, 2017) suggesting WiFi data as dynamic alternative to census data.

For city governments and urban policy-makers, tools can be built using our method to inform urban planning processes and real–time operational decisions. Such tools can help detect mobility patterns in a city over time, and how these patterns might change in case of emergencies or other anomalous events. Other tools based on this method could inform planning decisions by modeling urban mobility on different road network configurations to see how mobility flow would be impacted, such as when converting a street segment to a pedestrian-only zone. Design and policy alternatives can be defined, modeled, and tested to forecast the impacts of various land use or development scenarios on mobility patterns and city service demand.

The method can also support local retailers by understanding hyperlocal patterns of pedestrian activity. The model could be used, for instance, to suggest locations that have high rates of passers–by, and thus large numbers of potential customers. Furthermore, as the method is able to detect different temporal patterns of activity related to an area's land–use composition, it can further inform locational decisions and be used to evaluate the impacts of space programming and complementary (or competing) uses.

For residents and visitors, mobile applications can be developed to improve pedestrian navigation and wayfinding. Such applications can suggest routes based on preferences to avoid or seek crowds, depending on the user's mood in real–time. In offering user–defined routes, such applications have the potential to support urban walkability, impacting urban quality-of-life and individual well-being.

5. Conclusion and future work

We have presented a novel methodology to efficiently model urban mobility in public spaces at scale using WiFi network probe request data, and to couple detected movement patterns to the local street network to measure usage intensities. Our analysis demonstrates the potential to use WiFi probe request data as an alternative to other traditional technologies, one that allows for a hyperlocal understanding of mobility patterns and human behavior in heterogeneous urban environments. With the increasing ubiquity of WiFi-enabled mobile devices and municipal WiFi networks, our approach can be scaled to cover large urban areas without the need for new or dedicated pedestrian counting technologies.

To improve the accuracy of the model, we will continue to differentiate between pedestrians and motor–traffic by using the temporal information provided in the data as proxy for individual's movement speed. By combining output with additional information about road types (if pedestrian–friendly or not), we will be able to inform routing procedures of the model accordingly and will provide a more accurate picture of urban mobility. To minimize flaws due to route generation, we will review and test routing algorithms other than the shortest–route. Finally, to more completely explore mobility patterns and the potential causal links between activity and urban environmental conditions, we will next include data capturing a longer time frame and a larger geographic area. As cities all over the world face challenges of congestion and livability exacerbated by growing populations, the ability to measure, model, and predict urban mobility at high resolution is a critical tool for city managers and planners responsible for improving quality-of-life for city dwellers.

It must be acknowledged and emphasized that these data can be highly sensitive, and the potential for technology providers to track individuals creates a significant concern that should be at the forefront of public and regulatory discussions of data privacy and transparency. Our research is intended to show how the responsible use of high resolution spatiotemporal data can inform non-trivial urban operational and planning decisions that could have meaningful impacts on existing and future cities. By highlighting privacy concerns, we also hope that this work can help to raise awareness about what data cities and corporations are collecting, and how these data can be used.

Acknowledgements

This research was funded, in part, by the John D. and Catherine T. MacArthur Foundation (Grant no. 16-0875) and by the U.S. Department of Transportation, C2SMART research center. Data were provided by the Alliance for Downtown New York, and we would like to thank Jeremy Schneider for his support of our research partnership. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the view of any supporting institution. The authors are responsible for any errors.

References

Abedi, N., Bhaskar, A., & Chung, E. (2014). Tracking spatio-temporal movement of human in terms of space utilization using media-access-control address data. *Applied Geography, 51*, 72–81.

Barton, H., Grant, M., & Guise, R. (2003). *Shaping neighbourhoods: A guide for health, sustainability and vitality.* New York: Spon Press.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Jr., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C, 26*, 301–313.

Chilipirea, C., Petre, A., Dobre, C., & van Steen, M. (2016). Presumably simple: monitoring crowds using wifi. *17th IEEE international conference on mobile data management* (pp. 220–225). .

Competition Example http://www.d4d.orange.com/en/Accueil, Accessed date: 10 May 2017.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.

D. o. E. United Nations (2014). *P. D. Social Affairs, World urbanization prospects: The 2014 revision*.

Danalet, A., Bierlaire, M., & Farooq, B. (2012). Estimating pedestrian destinations using traces from wifi infrastructures. *Pedestrian and Evacuation Dynamics,* 1341–1352.

Dargay, J., Gately, D., & Sommer, M. (2007). Vehicle ownership and income growth, worldwide: 1960-2030. *Energy Journal, 28*(4), 143–170.

Downtown Alliance https://www.downtownny.com, Accessed date: 20 February 2017.

Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of New York city taxi trips. *Transactions on Visualization and Computer Graphics, 19*(12), 2149–2158.

Gonzalez, M. C., Hidalgo, C. A., & Albert-Laszlo, B. (2008). Understanding individual human mobility patterns. *Nature, 453*, 779–782.

Hansson, E., Mattisson, K., Bjoerk, J., Oestergren, P. O., & Jakobsson, K. (2011). Relationship between commuting and health outcomes in a cross-sectional population survey in southern sweden. *Public Health, 11*(834).

Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., & Gonzales, M. C. (2016). The timegeo modeling framework for urban mobility without travel surveys. *PNAS, 113*(45), E5370–E5378.

Johnson, N. E., Bonczak, B., & Kontokosta, C. E. (2018). Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. *Atmospheric environment*.

Kontokosta, C. (2016). The quantified community and neighborhood labs: A framework for computational urban science and civic technology innovation. *Urban Technology, 23*, 67–84.

Kontokosta, C., Hong, B., & Korsberg, K. (2017). *Equity in 311 reporting: Understanding socio-spatial differentials in the propensity to complain.* (arXiv preprint arXiv:1710.02452).

Kontokosta, C., & 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.

Land–use http://maps.nyc.gov/doitt/nycitymap/, Accessed date: 10 March 2017.

Lee, C. Y. (1961). An algorithm for path connections and its applications. *IRE Transactions on Electronic Computers, 3*, 346–365.

mappluto http://www1.nyc.gov, Accessed date: 20 February 2017.

Meneses, F., & Moreira, A. (2012). Large scale movement analysis from wifi based location data. *2012 International conference on indoor positioning and indoor navigation (IPIN)* (pp. 1–9). .

Millard-Ball, A., & Schipper, L. (2010). Are we reaching peak travel? trends in passenger transport in eight industrialized countries. *Transport Reviews, 31*(3).

Milne, A. (2014). Melin, Bicycling and walking in the united states 2014 – benchmarking report. *Tech. rep*. Washington, DC: Alliance for Biking and Walking.

New york city streets http://www1.nyc.gov/site/planning/data-maps/open-data.page, Accessed date: 10 February 2017.

Placemeter http://www.placemeter.com, Accessed date: 5 July 2017.

Prentow, T., Ruiz-Ruiz, A., Blunck, H., Stisen, A., & Kjrgaard, M. (2015). Spatio–temporal facility utilization analysis from exhaustive wifi monitoring. *Pervasive and Mobile Computing, 16*, 305–316.

Sapiezynski, P., Stopczynski, A., Gatej, R., & Lehmann, S. (2015). Tracking human mobility using wifi signals. *PLoS ONE, 10*(7).

Sevtsuk, A., Huang, S., Calabrese, F., & Ratti, C. (2006). *Mapping the mit campus in real time using wifi*.

Slack, B. (2017). *Traffic counts and traffic surveys.* The Geography of Transport Systems.

Smartphone Users https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/, Accessed date: 30 April 2017.

Sweet, M. (2014). Traffic congestion's economic impacts: Evidence from us metropolitan regions. *Urban Studies, 51*(10).

Tfl Study, http://wgallia.com/#!underground, visited:2017/01, accessed: 2017-04-20.

Trafficvision http://www.trafficvision.com, Accessed date: 15 May 2017.

Traunmueller, M., Quattrone, G., & Capra, L. (2014). Mining mobile phone data to investigate urban crime theories at scale. In: *Proc., SocInfo 2014: Social informatics* (pp. 396–411). Springer.

U. WHO (2010). Hidden cities: unmasking and overcoming health inequities in urban settings. *Tech. rep*World Health Organization WHO library Cataloguing–in–Publication Data.

Wang, Y., Yang, J., Liu, H., Chen, Y., Gruteser, M., & Martin, R. (2013). Measuring human queues using wifi signals. *19th International conference on mobile computing and networking* (pp. 235–238). .

Woodcock, J., Tainio, M., Cheshire, J., & Goodman, A. (2014). Health effects of the london bicycle sharing system: health impact modelling study. *BMJ, 348*.

Yuan, Y., & Raubal, M. (2012). Extracting dynamic urban mobility patterns from mobile phone data. *Geographic Information Science. 7478*, 354–367.

Zhao, P. (2014). Private motorised urban mobility in china's large cities: The social causes of change and an agenda for future research. *Journal of Transport Geography, 40*.