Computers, Environment and Urban Systems 64 (2017) 144–153

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Urban phenology: Toward a real-time census of the city using Wi-Fi data

Constantine E. Kontokosta [a,*], Nicholas Johnson [b]

[a] Center for Urban Science and Progress & Tandon School of Engineering, New York University, 1 Metrotech Center, 19th Floor, Brooklyn, NY 11201, United States
[b] Center for Urban Science and Progress, New York University & University of Warwick, 1 Metrotech Center, 19th Floor, Brooklyn, NY 11201, United States

ARTICLE INFO

Article history:
Received 22 June 2016
Received in revised form 18 January 2017
Accepted 21 January 2017
Available online xxxx

Keywords:
Urban studies
Human dynamics
Population
Community
Wi-Fi
Census

ABSTRACT

New streams of data are being generated by a range of in-situ instrumentation, mobile sensing, and social media that can be integrated and analyzed to better understand urban activity and mobility patterns. While several studies have focused on understanding flows of people throughout a city, these data can also be used to create a more spatially and temporally granular picture of local population, and to forecast localized population given some exogenous environmental or physical conditions. Effectively modeling population dynamics at high spatial and temporal resolutions would have significant implications for city operations and policy, strategic long-term planning processes, emergency response and management, and public health.

This paper develops a real-time census of the city using Wi-Fi data to explore urban phenology as a function of localized population dynamics. Using Wi-Fi probe and connection data accounting for more than 20,000,000 data points for the year 2015 from New York City's Lower Manhattan neighborhood – combined with correlative data from the U.S. Census America Community Survey, the Longitudinal Employer-Household Dynamics survey, and New York City administrative records – we present a model to create real-time population estimates classified by residents, workers, and visitors/tourists in a given neighborhood and localized to a block or geolocation proximate to a Wi-Fi access point. The results indicate that the approach has merit: we estimate intra-day, hourly worker and resident population counts within 5% of survey validation data. Our building-level test case demonstrates similar accuracy, estimating worker population to within 1% of the reported building occupancy.

1. Introduction

Our collective understanding of how people interact with cities and the built environment has historically been limited to surveys with low spatial and temporal resolution. The definition and estimation of urban population and population density at various scales has been a persistent problem in the planning and demography literature, and in the study of city dynamics more broadly (Foley, 1954; Schmitt, 1956; McDonald, 1989; Sutton, Roberts, Elvidge, & Baugh, 2001). Population estimates are used to forecast future demand for critical infrastructure and resources (such as energy, food and water), to measure the need for social services, and to effectively plan for future land use changes and transportation needs. In many cases, population estimates drive policy decisions and the prioritization of public investment across regions, cities, and neighborhoods. Census data, both the decennial census and the American Community Survey, currently provide our most robust glimpse of spatial population at a single point in time. However, Census data are limited not only by temporal and spatial constraints, but also by the lag time to release and gain access to these data for analysis purposes. Relying on annual surveys for planning and policy design decreases opportunities for urban innovation by extending the time needed to implement and evaluate planning decisions.

Measuring urban population at high spatial granularity and temporal frequency is critical to the understanding of socio-economic dynamics, public health, and emergency planning and response, among other factors (Dobson, Bright, Coleman, Durfee, & Worley, 2000; Hay, Noor, Nelson, & Tatem, 2005; Smith, Martin, & Cockings, 2014). However, current methods to collect population data constrain the ability of policymakers, researchers, and local stakeholders to know how many people are present at a specific time of day at a specific location, and how population counts and human activity levels are affected by diurnal rhythms and perturbations to "normal" routines (Candia et al., 2008; Kontokosta, 2016). These variations may depend on range of factors in the urban landscape, such as proximity to cultural attractions, large office buildings, and popular restaurants or entertainment spots. We would also expect population density levels to vary due to abnormal or atypical events, such as significant weather conditions (e.g. snow or heavy rainfall), disasters and other emergencies and localized community events such as music festivals, organized public events, and other public gatherings.

The increasing ubiquity of Wi-Fi enabled mobile devices, and the proliferation of public wireless networks in urban environments, creates new opportunities to understand population dynamics across

* Corresponding author.
E-mail address: ckontokosta@nyu.edu (C.E. Kontokosta).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.01.011

space and time (Ratti, Frenchman, Pulselli, & Williams, 2006). In the U.S., the market penetration of Wi-Fi-enabled smartphones has more than doubled between 2011 and 2015 (Anderson, 2015). In New York City, and other dense cities, nearly 80% of residents own such a device (New York City Department of Consumer Affairs, 2015). While the representativeness of these data must be considered, and the potential biases in analyses derived from these data need to be acknowledged, it can be expected that over time all or nearly all individuals will own or have access to a Wi-Fi-enabled mobile device. The expansion of public networks and free, high-speed Wi-Fi access in cities such as New York, Paris, and Tel Aviv increases the potential to use connection data to better understand population fluctuations in time and space, while directly accounting for privacy concerns and protecting confidentiality of users (Lane, Stodden, Bender, & Nissenbaum, 2014).

Our research seeks to develop a real-time census of the city within the context of describing an urban phenology, defined here as the cyclic and seasonal variations of human behavior in urban environments. Using unique access to Wi-Fi probe and connection data accounting for more than 20,000,000 data points for the year 2015 from the Lower Manhattan neighborhood of New York City – combined with correlative data from the American Community Survey, the Longitudinal Employer-Household Dynamics survey, and New York City administrative records – we present a model to create real-time population estimates of residents, workers, and visitors/tourists in a given neighborhood and localized to a specific block or geolocation. The goal of this work is to develop a real-time population estimate at the street scale that will serve as a baseline for understanding neighborhood activity levels and provide critical insight into population density dynamics to inform city operations, policy, and planning. The paper follows with a review of relevant literature, a discussion of the data and methods, our approach to model validation, and the results of the model for the Water Street sub-network in Lower Manhattan. We conclude with a discussion of the findings and applications to urban planning and city management.

2. Background

The Census Bureau's Population Estimates Program (PEP) produces estimates of the population for the United States, its states, counties, cities, and towns, as well as for the Commonwealth of Puerto Rico and its municipios (Yowell & Devine, 2014). Demographic components of population change (births, deaths, and migration) are produced at the national, state, county, and sub-county levels of geography. Additionally, housing unit estimates are produced for the nation, states, and counties. These estimates are used in federal funding allocations, as survey controls, as denominators for vital rates and per capita time series, and as indicators of demographic change. With each new release of annual estimates, the entire time series of estimates is revised for all years back to the last census. All previously published estimates are superseded and archived. This method of producing population estimates has both a time-lag and limited temporal and spatial resolution, which can introduce significant underestimates for dense and rapidly developing urban areas like New York City (Mulder, 2002).

In addition, census survey data only capture resident population based on place of primary residence. This obscures the true population density in a neighborhood at any given time, as it ignores temporal fluctuations in worker and visitor populations and only accounts for a survey respondent's stated primary or "usual" residence, which may not be the actual place of residence for a household (U.S. Census Bureau, Population Division). Attempts at improving population estimates have relied on the integration of multiple survey data sources, which can significantly reduce the reliability of the estimates by compounding underlying margins of error within and across individual surveys (Moss & Qing, 2012).

The U.S. Census does provide estimates of daytime population, using two methods that account for resident population and worker populations that do not live in the area. The lowest level of geography for these estimates is the Census Place (McKenzie, Koerber, Fields, Benetsky, & Rapino, 2010). Other census data products can be used to infer daytime population through worker estimates, including the Longitudinal Employer Household Dynamics (LEHD) survey and tabulations of the American Community Survey data through Census Transportation Planning Products (CTPP) (Abowd, Haltiwanger, & Lane, 2004; Kobayashi, Medina, & Cova, 2011). However, these approaches are limited as they do not capture intraday or weekend/weekday variations, they exclude visitors/tourists and others not characterized as workers or residents, and they suffer from typical challenges of small area estimation from survey data, resulting in large margins-of-error for small geographies (Rao & Molina, 2015).

New streams of data are being generated by a range of in-situ instrumentation, mobile sensing, smart card data, and social media that can be integrated and analyzed to better understand urban activity and mobility patterns (Bettencourt, 2014; Bhaduri, Bright, Coleman, & Urban, 2007; Kontokosta, 2016; Lansley & Longley, 2016; Long & Thill, 2015; Reades, Calabrese, Sevtsuk, & Ratti, 2007). These data can also be used to create a more spatially and temporally granular picture of local population at a specific point in time, and to forecast localized population given some exogenous environmental or physical conditions. For example, Bhaduri et al. (2007) use aerial imagery to estimate population dynamics through dasymetric modeling. Dasymetric modeling integrates population data, typically from census surveys, with ancillary data sources, such as land cover, road density, or parcel data, to develop finer resolution population estimates (Alahmadi, Atkinson, & Martin, 2013; Mennis, 2003; Reibel & Bufalino, 2005; Tapp, 2010; Nagle, Buttenfield, Leyk, & Spielman, 2014). These approaches are limited by the uncertainties underlying the original sources of data and data collection methods, as well as those inherent to the model created to link the various datasets. More common approaches, such as areal interpolation, suffer from the assumption of linear distributions of population across space. While the method proposed by Bhaduri et al. (2007) holds promise, it only describes average working day population distributions. Other methods are also constrained by focusing on achieving estimates of finer spatial resolution, rather than temporal periodicity. Even when the time dimension is accounted for, resolutions at real-time or near-real-time are not possible without ancillary data collected at similar frequencies.

Catalyzed by the availability of big data, the modeling of human dynamics in urban environments is an increasingly important, though nascent, component of urban science research (Becker et al., 2011; Jiang, Ferreira, & González, 2012; Song, Qu, Blumm, & Barabási, 2010; Gonzalez, Hidalgo, & Barabasi, 2008a, 2008b). In the biological sciences, the study of phenology – the seasonal and cyclic patterns in plants and animals and the timing of natural events – is a means to understand the impact of endogenous and exogenous changes to local ecology and habitats (Edwards & Richardson, 2004; Menzel, 2002; Walther et al., 2002). A similar approach can be applied in the urban context, as longitudinal observations of real-time population dynamics and mobility patterns can be used to understand changes in the physical, natural, and social aspects of the built environment and their effects on individual and collective human behavior. We introduce the concept of urban phenology here in relation to the observational and data-driven understanding of patterns of human activity at the micro- (building, street) and meso- (neighborhood, district) scale in cities.

3. Data and methodology

3.1. Data sources and description

The primary data used in this study are supplied by the Alliance for Downtown New York (ADNY) who own and maintain a publicly-accessible wireless network consisting of 53 access points (APs) distributed throughout lower Manhattan in New York City (see Fig. 1). The APs