Application of Social Media Data to High Resolution Mapping of a Special Event Population

K. M. Sims, E. M. Weber, B. L. Bhaduri, G. Thakur and D. R. Resseguie

Oak Ridge National Laboratory, PO Box 2008 MS6017, Oak Ridge, TN 37831
Telephone: (+1) 865-576-9366
Email: {simskm, weberem, bhaduribl, thakurg, resseguiedr}@ornl.gov

Abstract

Society’s increasing participation in social media provides access to new sources of near real-time data that reflect our activities in space and in time. The ability for users to capture and express their geolocation through their phones’ global positioning system (GPS) or through a particular location’s hashtag or Facebook Page provides an opportunity for modeling spatiotemporal population dynamics. One illustrative application is the modeling of dynamic populations associated with special events such as sporting events. To demonstrate, Twitter posts and Facebook check-ins were collected across a 24-hour period for several football game days at the University of Tennessee, Knoxville, during the 2013 season. Population distributions for game-hours and non-game-hours of a typical game day were modeled at a high spatial resolution using the spatiotemporal distributions of the social media data.

Keywords: social media, population, population modeling.

1. Introduction

Modeling population distributions at high spatial and temporal resolution requires accounting for the dynamic nature of human populations. Models and representations of population that rely on census counts necessarily miss these dynamics. However, there have been recent efforts to incorporate daytime or diurnal distributions (Kobayashi et al. 2011; Bhaduri et al. 2007), and episodic or tourist populations (Jochem et al. 2012) in order to better capture population dynamics. To continue these efforts, publicly available data feeds from social media offer an additional opportunity to improve population models (Bukhari et al. 2012).

The world’s two leading social media platforms are Twitter, with 230 million monthly active users, and Facebook, with 874 million monthly active users. Recognizing the value of social media activity to geographic analysis, Goodchild (2007) has referred to those posting on these platforms as “Citizen Sensors.” Only a portion of social media data has associated location information, however. For example, although an estimated 500 million tweets were sent per day as of October 2013 (Twitter, Inc.), estimates of the portion of tweets that are geo-located have ranged from 0.47% (Cheng 2010) to 3.17% (Morstatter et al. 2013).

A natural application of the geo-located subset of social media data is the modeling of episodic populations associated with special events with high attendance and a significant presence on social media; in particular, this study focuses on game day college football fans at The University of Tennessee (UT), Knoxville. Using tweets (from Twitter) and check-ins (from Facebook), this research integrates this new form of data in a high-resolution dasymetric population distribution model.

2. Methods & Results

The area within a 1.5 mile radius around The University of Tennessee football stadium was chosen for this study, and the population associated with football game days was modeled. Geo-located tweets and check-ins were collected for the 24-hour period surrounding the scheduled kickoff for each home game in 2013. Seven terms associated with the university were used to filter tweets from Twitter's streaming API (table 1). A cumulative count of Facebook check-ins was captured every 30 minutes for 95 establishments associated with game day activities (e.g., restaurants and tailgating locations).

UT Term/Phrase	Number of Geocoded Tweets
"Tennessee"	11,582
"Vols"	7,837
"GBO"	2,495
"VFL"	1,612
"Neyland"	1,144
"Football Time In Tennessee"	1,135
"Big Orange"	418

Table 1. The total seasonal count of game-day geocoded tweets associated with The University of Tennessee, 2013.

Two scenarios were modeled: 1) a "non-game-hours" scenario and 2) a "game-hours" scenario. Each model outputs a population estimate for each cell in a raster grid with 3 arc-second resolution (~90 m). The 2012 version of the LandScan USA (Bhaduri et al. 2007) gridded nighttime dataset was used as a baseline population distribution to which the new modeled distributions could be added to create the final output grids. The LandScan USA nighttime distribution better represents the study area on a Saturday (all game-days were Saturdays) than the daytime dataset, since the daytime dataset assumes a weekday distribution of workers and students very different from what would be expected in the area on a weekend.

A consistent measure of social media activity was required for each raster cell. First, the tweets and check-ins (collectively referred to as "posts" from here on) were divided into two sets based on their timestamps. Posts were considered for the "game-hours" scenario if they occurred less than two hours prior to kickoff or less than 3 hours after kickoff. All posts outside of those hours were considered for the "non-game-hours" scenario. A count of tweets and a count of check-ins were computed for each raster cell for each of the two scenarios, resulting in four raw count rasters.

Because of the limited amount of geo-located posts and the spatial errors in the associated location information, it is possible that some locations that attract event populations might have no representation in the social media data. To overcome this limitation, kernel density estimation with a radius of two grid cells was performed on each of the four raw count rasters to estimate tweet densities and check-in densities across the grid. The densities were then scaled so that each value represented posts per cell and could be interpreted as interpolated counts.

For each scenario, a linear relationship between social media activity and event population is assumed. For each raster cell i, the special event population (y) is modeled as:

$$y_i = \beta_T w_i \tag{1}$$

where β_T is a linear coefficient specific to the scenario T and w_i is the number of posts at cell i. The β coefficient is the number of fans represented by each geo-located post. If available, an observed value or estimate of the total population in the study area associated with a scenario can be used to estimate β_T:

$$\beta_T = E_T / \sum_{i=1}^{n} w_i \tag{2}$$

where E_T is the estimated total special event population for scenario T. The study area total for the game-hours and non-game-hours scenarios can be estimated by summing two separate components of the special event population: A, the ticketed fans, which is estimated by averaging the recorded attendance for each game, and α, all the other (non-ticketed) people in the area specifically for the event:

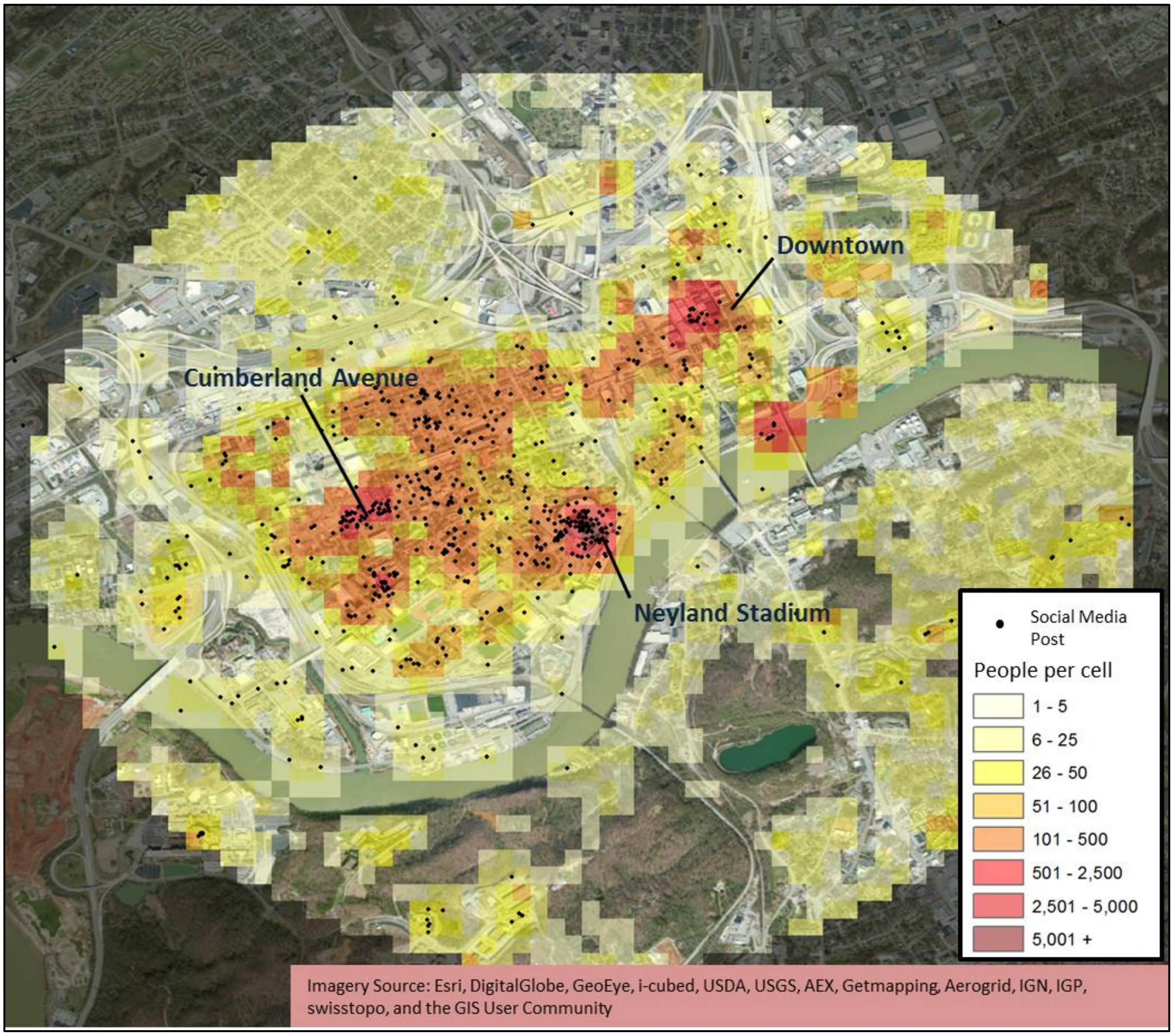
$$E_T = \lambda_T (A + \alpha) \tag{3}$$

where λ is a parameter representing the estimated portion of the peak game day population. Ultimately, the final population estimate for each cell i is the sum of the baseline (LandScan USA nighttime) population (L_i) and the special event population (y_i):

$$P_i = L_i + y_i \tag{4}$$

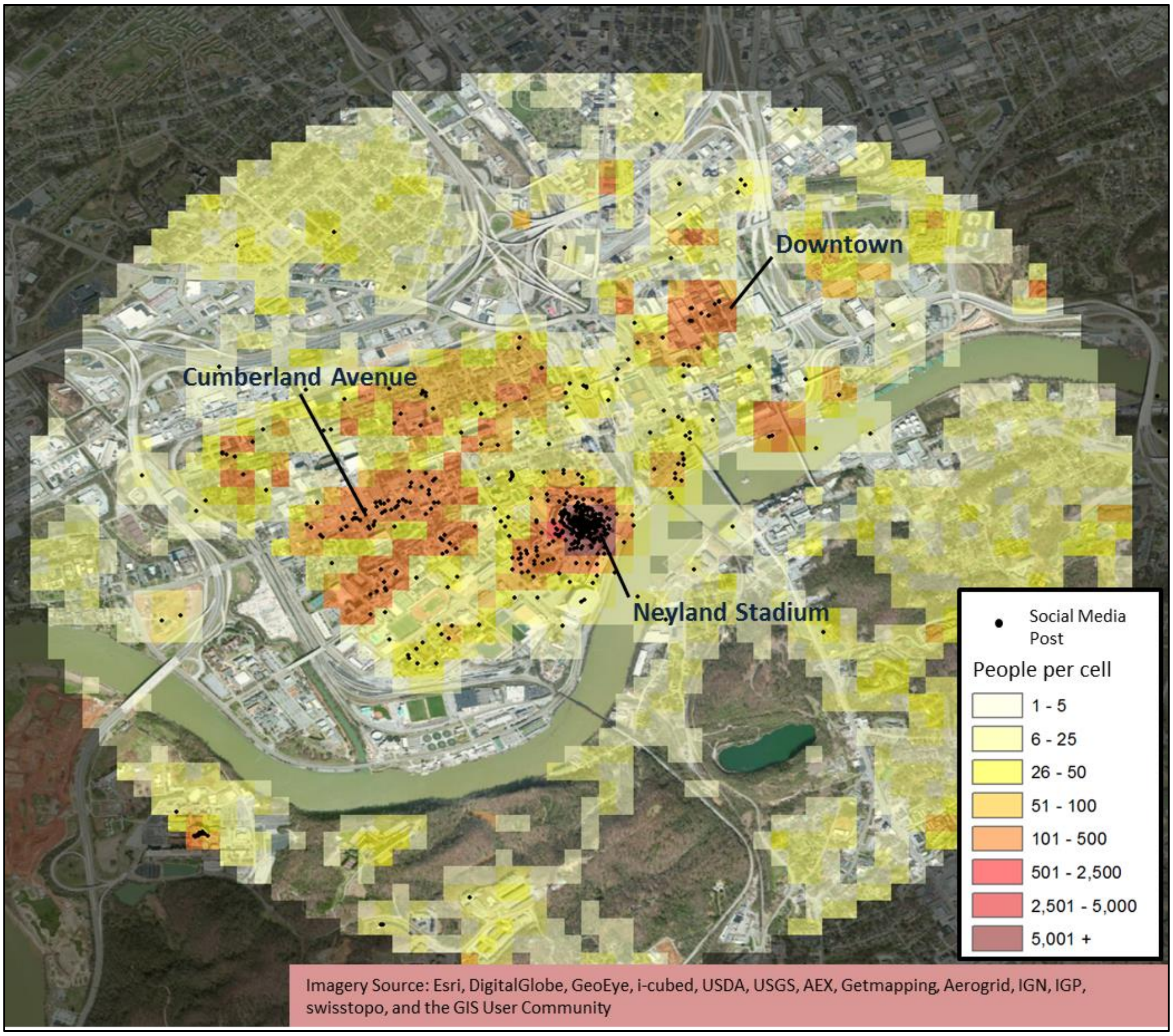
In most special event situations, there is unlikely to be data to support precise estimates of the parameters, α and λ, from equation 3. But event officials will often have expert knowledge and be privy to information that allows reasonable estimates of these parameters. Ultimately, a software solution aimed at event officials would allow such knowledge to be incorporated in the parameterization. Fig. 1 and fig. 2 show example representations of the non-game-hours and game-hours scenarios using rough estimates of the parameters. The estimate for α was 30,000. For the game-hours scenario, λ was set at 1, assuming the population peaks during the game. The non-game-hours scenario is meant to represent a moment approximately 3 hours before kickoff, for which λ was set at two-thirds.

High populations can be seen in and near the stadium in both fig. 1 and fig. 2, but with much greater concentration in fig. 2. Fig. 1 shows greater concentrations in areas on and near campus that are popular for tailgaters as well as along Cumberland Avenue and in the downtown area, where restaurants, bars, and shops are concentrated. Different parameterizations (for α and λ) would lead to different absolute population values, but the overall pattern and the ratios among the values would remain the same (since the spatial distribution is based only on the social media data).



Imagery Source: Esri, DigitalGlobe, GeoEye, i-cubed, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 1. Example game day population distribution around the University of Tennessee, Knoxville, during non-game-hours.



Imagery Source: Esri, DigitalGlobe, GeoEye, i-cubed, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 2. Example game day population distribution around the University of Tennessee, Knoxville, during game-hours.

3. Acknowledgments

This manuscript has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

4. References

Bhaduri, B.L., Bright, E., Coleman, P. and Urban, M., 2007. LandScan USA: A High-Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics. *GeoJournal*, 69(1), pp.103-17.

Bukhari, I., Wojtalewicz, C., Vorvoreanu, M. and Dietz, J.E., 2012. Social Media Use for Large Event Management: The Application of Social Media Analytic Tools for the Super Bowl XLVI. *Homeland Security (HST), 2012 IEEE Conference on Technologies for Homeland Security*.

Cheng, Z., Caverless, J. and Lee, K., 2010. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 759-68.

Goodchild, M., 2007. Citizens as sensors: The world of volunteered geography. *GeoJournal,* 69(4), pp.211-21

Jochem, W., Sims, K., Bright, E., Urban, M., Rose, A., Coleman, P. and Bhaduri, B., 2013. Estimating Traveler Populations at Airport and Cruise Terminals for Population Distribution and Dynamics. *Natural Hazards*, (68), pp.1325-42.

Kobayashi, T., Medina, R. and Cova, T., 2011. Visualizing Diurnal Population Change in Urban Areas for Emergency Management. *The Professional Geographer*, 63(1), pp.113-30.

Morstatter, F., Pfeffer, J., Liu, H. and Carley, K.M., 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming Api with Twitter's Firehose. *In Proceedings of ICWSM*. Cambridge, MA: AAAI Press.

Twitter, Inc., 2013. S-1 Paperwork: US Security and Exchange Commission (SEC), [online] Available at: <http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm> [Accessed 10 December 2013].