Estimating Hourly Population Distribution Patterns at High Spatiotemporal Resolution in Urban Areas Using Geo-Tagged Tweets and Dasymetric Mapping

Jaehee Park
Department of Geography, San Diego State University, CA, USA
jpark1200@sdsu.edu

Hao Zhang
HDMA center, San Diego State University, CA, USA
zhanghaoshogo@gmail.com

Su Yeon Han
Department of Geography, San Diego State University, CA, USA
shunny1004@gmail.com

Atsushi Nara
Department of Geography, San Diego State University, CA, USA
anara@sdsu.edu

Ming-Hsiang Tsou[1]
Department of Geography, San Diego State University, CA, USA
mtsou@sdsu.edu

Abstract

This paper introduces a spatiotemporal analysis framework for estimating hourly changing population distribution patterns in urban areas using geo-tagged tweets (the messages containing users' geospatial locations), land use data, and dasymetric maps. We collected geo-tagged social media (tweets) within the County of San Diego during one year (2015) by using Twitter's Streaming Application Programming Interfaces (APIs). A semi-manual Twitter content verification procedure for data cleaning was applied first to separate tweets created by humans from non-human users (bots). The next step was to calculate the number of unique Twitter users every hour within census blocks. The final step was to estimate the actual population by transforming the numbers of unique Twitter users in each census block into estimated population densities with spatial and temporal factors using dasymetric maps. The temporal factor was estimated based on hourly changes of Twitter messages within San Diego County, CA. The spatial factor was estimated by using the dasymetric method with land use maps and 2010 census data. Comparing to census data, our methods can provide better estimated population in airports, shopping malls, sports stadiums, zoo and parks, and business areas during the day time.

2012 ACM Subject Classification Human-centered computing → Social media

Keywords and phrases Population Estimation, Twitter, Social Media, Dasymetric Map, Spatiotemporal

Digital Object Identifier 10.4230/LIPIcs.GIScience.2021.I.10

Funding This material is partially based upon work supported by the National Science Foundation under Grant No. 1416509 and Grant No. 1634641.Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

[1] Corresponding author

1 Introduction

The widespread use of social media and mobile phone data provides a great research opportunity for researchers to map and analyze dynamic human behaviors, communications, and movements [27, 8, 24, 25]. People use smartphones, mobile devices, and personal computers, leaving their digital footprints on the Internet. These human-made digital records provide a foundation for human dynamics research. Human dynamics is a new transdisciplinary research field attracting scientists and researchers from different domains, including complex systems [3], video analysis [6, 28], spatial diffusion of events [18], human mobility and network [14, 15], public health [22] and geography [13, 26]. One key research topic of human dynamics is to estimate the dynamic change of population distribution in urban areas. Although the census provides the detailed population statistics covering age, sex, and race, it does not reflect the dynamic change of population since census population is based on the location of residence. Therefore, estimating the dynamic change of population is crucial for evacuation planning, disaster management, epidemic management, event planning, and urban planning. For example, dynamic population estimation at finer scales can be useful for a stage-based evacuation planning during emergency situation[23]. Conventionally, the change of population distribution is estimated from the census survey by using data sampling and forecasting techniques. Recently, scientists have started using satellite images [5], mobile phone data [4, 8], or vehicle probe data [16] to estimate the dynamic change of population distribution at small area level. One example is to use mobile phone-based call detail records (CDR) to detect spatial and temporal differences in everyday activities among multiple cities [1]. Another example is to estimate seasonal, weekly, and daily changes in population distribution over multiple timescales with aggregated and anonymized mobile phone data [8].

In Geographic Information Systems (GIS) and cartographic research, dasymetric mapping methods have been applied to estimate population density using census data and ancillary data sources [29, 12, 17]. In the previous studies, the authors have identified that it is a challenging problem to integrate vector-based census tracks and raster-based land cover data and satellite images for dasymetric mapping. To improve the traditional problems of binary value in categorical data and areal weighting, [21] introduced an intelligent dasymetric mapping technique (IDM) with a data-driven methodology to calculate the ratio of class densities. Similar to the IDM method, this study utilizes social media data (geo-tagged data), other GIS data sources (land use and census data), and dasymetric mapping techniques to estimate the hourly change of population distribution. There are several advantages of using social media for population estimation[19]. The real-time updates of social media messages can better reflect dynamic changes of population than remote sensing imageries, which are often more expensive in cost and time to collect and process data [9]. Alternatively, mobile phone data, such as CDR, are also very expensive and inaccessible. Another drawback of CDR is that it is not possible to identify the content of communications in each phone call. In contrast, social media data are easy-to-collect, free (using public access methods), content-rich, and updated in real-time [25, 18].

In this study, we estimate hourly population distribution patterns at a high spatiotemporal resolution in urban areas using geo-tagged tweets and dasymetric mapping. The remainder of this paper follows the process as shown in Figure 1.

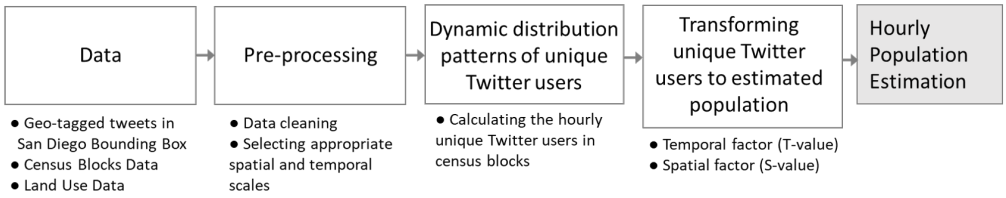


Figure 1 Overview of the process.

2 Data and pre-processing

2.1 Data collection

This study utilized public Twitter Application Programming Interfaces (APIs) to collect geo-tagged Twitter messages (tweets) through customized Python programs. The geo-tagged tweets were downloaded via the Twitter Streaming APIs and stored in a NoSQL database (MongoDB). We collected geo-tagged tweets within the bounding box of San Diego County for one year (from 2015/1/1 to 2015/12/31). There are total 7,884,806 geotagged tweets. Among the collected data, 2,601,560 (33.2%) tweets do not contain the exact coordinates and 2,355,945 (30.1%) were created outside the San Diego County. This study only utilized the remaining 2,927,301 (37.7%) geo-tagged tweets within San Diego County for population estimation. We noticed that the number of monthly geo-tagged tweets in San Diego County in 2015 fluctuated. The months of March and April 2015 have the biggest number of geo-tagged tweets. A similar trend reported by other researchers, such as Business Insider [11] suspecting that the causes might be due to Twitter's systematic updates. Figure 2 illustrates the spatial distribution of geotagged tweets from 12am to 1am in downtown, San Diego during weekdays in July 2015 (over one month).

To apply dasymetric mapping based on different types of land use, the 2017 parcel land use data was downloaded from the San Diego Association of Governments (SANDAG) website (`http://www.sandag.org`). The census blocks and their population estimates in San Diego County were obtained from the 2010 Decennial Census data.

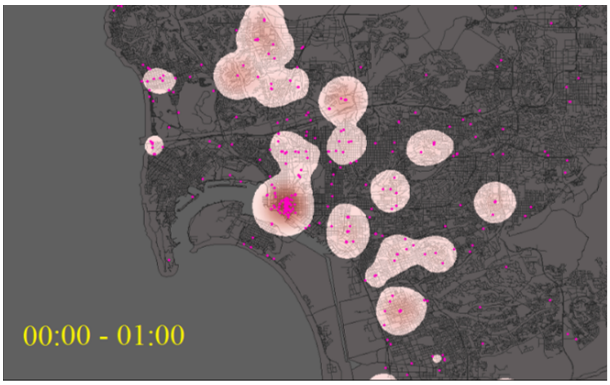


Figure 2 The distribution of geo-tagged Twitter messages (tweets as red dots) in San Diego downtown from 12am to 1am during weekdays in July of 2015 (26 days combined).

2.2 Data cleaning

Previous research has identified some major types of data noises in Twitter data, including spams, bots, and cyborgs [30, 7]. Spams and bot messages are created for reaching more users and increasing the financial gain for spammers. Since spam and bots messages can not represent the actual locations of human beings, we removed all the identifiable spams and bots based on the source field in Twitter metadata and some general bot detection rules (for example, removing tweets from TweetMyJOBS and others based on a black list of the source field). The major portion of the noise (spams and bots) in San Diego dataset includes job posting (9.07% of the total geo-tagged tweets, such as TweetMyJOBS), advertisements (1.60%, such as dlvr.it), and earthquake (1.06%) in San Diego County. The earthquake event-related tweets are geo-tagged in the localities of the earthquakes. In this study, 13.01% of geo-tagged tweets were identified as noises and removed. After removing these spams and bot posts, 2,546,385 tweets were used for calculating the unique Twitter users in each census block within one hour by filtering multiple messages posted by a single user for weekdays and weekends.

2.3 Selecting appropriate spatial and temporal scales for population estimation

For spatial units, the U.S. Census block was selected to estimate the distribution of the population. A census block is the smallest geographic unit defined by the U.S. Census Bureau for demographic analysis and therefore, it can be aggregated to census tract or other spatial units for the purpose of analysis. For example, census blocks can be aggregated to traffic analysis zones(TAZ), which is a special area formalized by local transportation officials for analyzing traffic-related data and evacuation planning. Researchers can utilize TAZ to create disaster evacuation plans and emergency response procedures. We selected one hour as our temporal resolution for estimating population density in San Diego County to meet the need for evacuation planning. In Figure 3, during weekdays, the unique Twitter user activities of posting Twitter messages decrease from midnight to 4 am. From 4 am to noon, the user activity starts to climb up. We assume that relatively a large number of Twitter activities around noon are due to tweets related to lunch time activities posted by residents and visitors. The peak of the tweeting activities comes at around 8 pm when people are getting dinner or enjoying leisure time with friends or family members. We also noticed that tweeting activities show different patterns between weekdays (Monday to Friday) and weekends (Saturday and

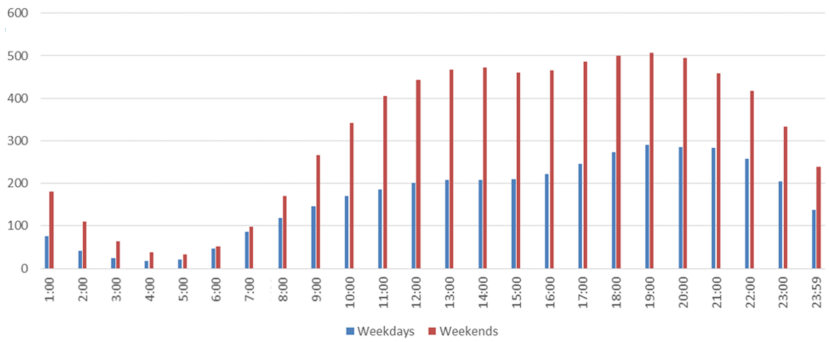


Figure 3 Comparison of hourly average numbers of unique Twitter users in San Diego County on weekdays (Monday to Friday) and weekends (Saturday to Sunday) in 2015.

Sunday). In general, the tweeting activities are more active during the weekends comparing to weekdays. Despite the similar pattern found on the weekdays where people tweeted most around 8 pm, the tweeting rate is high at around 2 pm during weekends. Therefore, we distinguish weekdays from weekends for the hourly population density estimation.

3 Methodology

3.1 Dynamic distribution patterns of unique Twitter users

3.1.1 Calculating the hourly unique Twitter users in census blocks

Within each geographical unit of census blocks, we estimate the population during a specific hourly time slot by calculating the frequency of the unique user IDs. Since one Twitter user can post several tweets within an hour from the same region (a census block), we counted one unique user ID once within an area for one hour rather than the total number of tweets. Figure 4(a) and (b) represent the distribution of unique Twitter users from 6 am – 6:59 am (a) and from 8 pm – 8:59 pm (b) respectively during weekdays in 2015 in San Diego County. The unique Twitter user density was calculated by using the total unique Twitter users within one census block during the specific hour, divided by the area of the census block. Figure 4(c) displays the 2010 population census data to visually compare its geographical distribution to that of unique Twitter users. In these maps, we selected the quantile classification method at

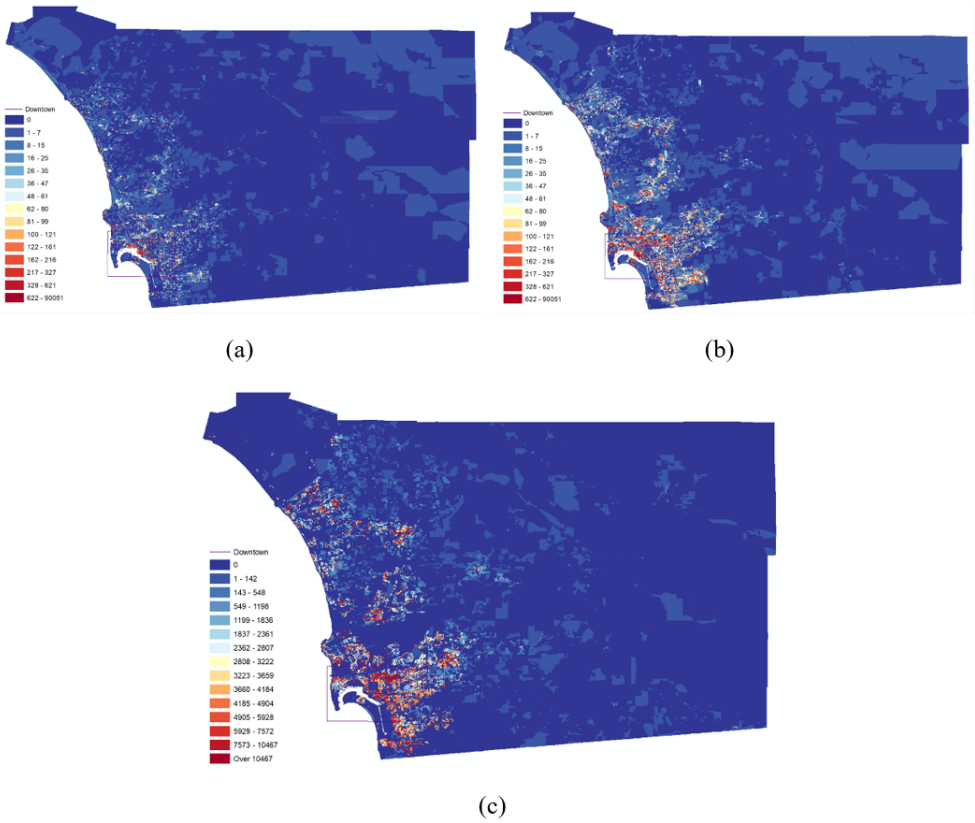


Figure 4 Spatial distribution patterns of unique Twitter users using census blocks in San Diego County from 6am – 6:59am (a) and from 8pm – 8:59pm (b) with 2015 geo-tagged tweets for weekdays. The (c) map displays the population density using 2010 census data.

8pm as the classification framework (applied to other time slots) in order to compare their spatial patterns. Figure 4(a) and (b) show an increase in unique Twitter users from 6 am to 8 pm in Western urbanized areas. The geographical distribution of unique Twitter users from 8 pm – 8:59 pm (Figure 4(b)), when has the highest average number of unique Twitter users in 2015 in San Diego County, is similar to that based on the 2010 census data (Figure 4(c)).

Maps in Figure 5 are enlarged views of Figure 4 exhibiting San Diego City downtown areas. Figure 5(a) and (b) highlight the increase of the number of unique Twitter users in areas shopping malls in Fashion Valley and Mission Valley, Balboa Park and San Diego Zoo, and the downtown Gaslamp area. The dynamic changes in these areas are reflecting the real world activities in San Diego downtown area. By comparing the 8 pm map (b) with the 2010 census block population map (c), we found that the large number of unique Twitter users in areas where there is no population in the census data. These areas are governmental and commercial lands including the (A) San Diego international airport, (B) the downtown Gaslamp quarter area, (C) Balboa Park and San Diego Zoo, (D) shopping malls in Fashion Valley and Mission Valley, and (E) Qualcomm stadium. Since the census population is considered as nighttime population estimated from residential addresses, this example shows the capability of utilizing social media data to estimate daytime population distribution at a finer spatio-temporal scale.

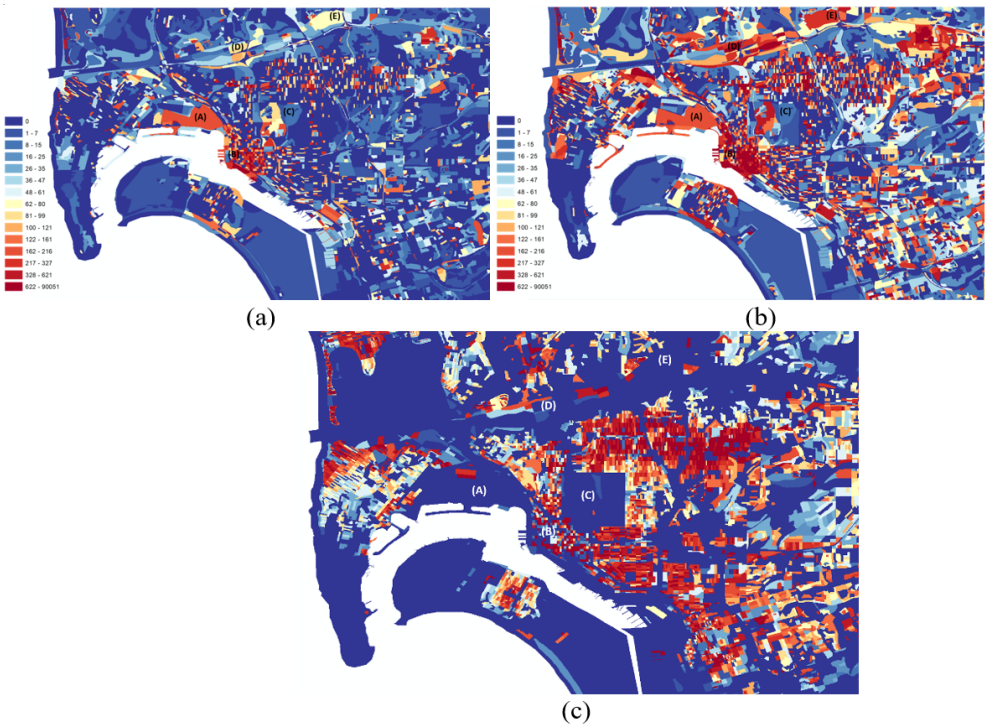


Figure 5 The spatial distribution of unique Twitter users in census blocks of San Diego downtown areas from 6 am to 6:59 am (a) and from 8 pm to 8:59 pm (b) with 2015 geo-tagged tweets for weekdays. The (c) map displays the 2010 census data in San Diego downtown areas.

3.1.2 Comparing the population change patterns of unique Twitter users between weekdays and weekends

With the hourly unique Twitter users density maps being produced (Figure 4 and Figure 5) based on weekdays and weekends, some human movement patterns can be detected and further analyzed. One of the advantages of visualizing dynamic Twitter user population patterns is that their dynamic changes can reflect the real-world situation with a high spatial resolution (census blocks) and a high temporal resolution (hourly). The following example introduces a case study in the Qualcomm Stadium with a comparison between weekdays and weekends (Figure 6). The Qualcomm Stadium is a multi-purpose stadium located in San Diego City, CA. The Qualcomm Stadium events data is archived through their official website in the events calendar. During the weekdays, the stadium usually hosts one to three events per day from 15:00 to 20:30. The events held on weekends usually started from 10:30 and ended at 17:30. The population density of unique Twitter users in Qualcomm Stadium during the weekdays shows the highest peak of Twitter user activities at 6pm. The high peaks of weekend's activities are from 1pm to 5pm. These patterns match the real-world situations since most football game events are happening between 1pm to 5pm on weekends. Figure 7 illustrates the comparison of the unique Twitter user density patterns in the Qualcomm's census blocks between weekday (a) and weekends (b) from 12pm to 12:59pm with its surrounding area. Qualcomm Stadium has a higher density of population at 12pm during weekends (comparing to weekdays).

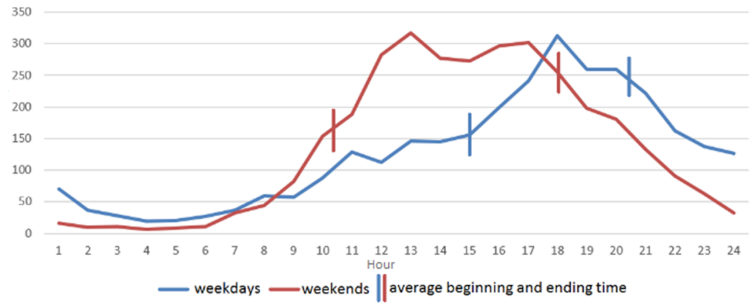
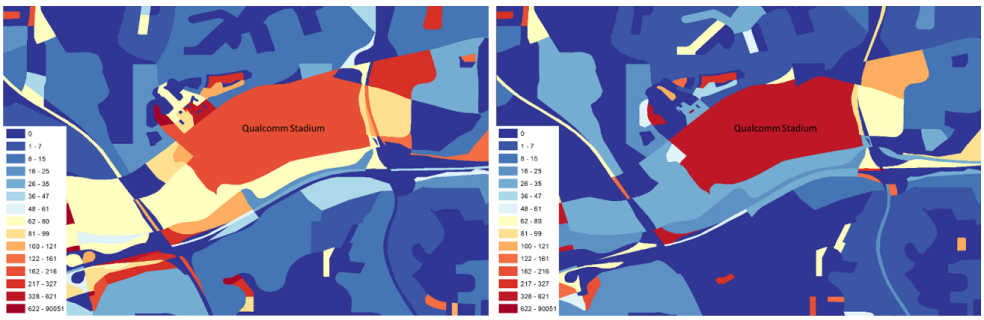


Figure 6 Comparing weekdays (blue) and weekends (red) hourly unique Twitter user density in the Qualcomm Stadium census block using 2015 geo-tagged tweets.



(a) (b)

Figure 7 Hourly Unique Twitter User Density from 12pm to 12:59 pm at the Qualcomm Stadium census block for Weekdays (a) and Weekends (b) in 2015.

3.1.3 Comparing unique Twitter population with census data

Comparing the weekdays and weekends unique Twitter user density map in Census Block polygon with census population can reveal the fact whether Twitter population can be used to represent the human mobility and real human population during different period of time in a day. The Census population represents the population distribution during the nighttime since it collects the number of people living in their household.

The Table 1(a) presents the $Z_{hx \cap pop}$ values in San Diego County area which compares the similarity of census block with unique Twitter user in different time slot from H1 to H24. Each Z value represents for the sum of absolute difference (SAD value) of two sets of data within range 0 to 1 based on formula 1.

$$Z_{hx \cap pop} = \sum \left| \frac{P_{A \cap hx}}{P_{hx_{\max}}} - \frac{P_{A \cap pop}}{P_{pop_{\max}}} \right| \tag{1}$$

Where:
$Z_{hx \cap pop}$ = the sum of the absolute difference of number of population between time slot hx and census population pop;
$P_{A \cap hx}$ = the value of unique Twitter population in time slot hx in Polygon P_A;
$P_{hx_{\max}}$ = the maximum value of unique Twitter population in time slot hx.

Note that *sd* refers to San Diego, *cb* refers to census block polygon, *wd* refers weekdays, and *we* refers to weekends. Thus, the intersection between H1 (0:00 to 0:59) and $Z_{sd\_cb\_wd}$ stands for the SAD Value of comparing the unique Twitter user density map with census block population density in the scale of San Diego County during weekdays. Based on the results showed in the table for census block polygon, the H5 (4:00 to 4:49) in weekdays and H6 (5:00 to 5:59) in weekends are the two time slot where the unique Twitter user is the closest to the census block population. The census block population records the number of human population in the residential area in detail. Meanwhile, 4:00 to 5:59 is usually the time when people get up during the morning time. Thus, it is possible to reflect the human residential area by using Twitter data.

Table 1(b) presents the $Z_{hx \cap pop}$ values in San Diego downtown area by comparing the census block population with unique Twitter user in downtown area, San Diego. Note that *dt* refers to downtown area of San Diego, H5 (4:00 to 4:49) for both weekdays and weekends is the time slot where the unique Twitter user is the closest to the census block population. On the other side, from the perspective of dissimilarity, H24 (23:00 to 23:49) and H1 (0:00 to 0:59) have the most dissimilar unique Twitter user distribution comparing to the census block population.

3.2 Transforming unique Twitter users to estimated population with spatial and temporal variation factors

The previous sections illustrate how to calculate the dynamic changes of unique Twitter users in high spatial and temporal resolution units. The next step is to create a dynamic population model to transform the numbers of unique Twitter users into estimated population. We proposed a simplified population estimation model using census blocks, land use data, and dasymetric mapping methods like the following:

$$\widehat{D}_{hx \cap a} = UserNumber_{hx \cap A} * (T_{hx}) * (S_{hx \cap A}) \tag{2}$$

Table 1 The sum of absolute difference between the number of hourly unique twitter data(from 0:00 to 23:59) with census block population during weekdays and weekends in (a) San Diego County and (b) San Diego Downtown.

		(a) San Diego County		(b) San Diego Downtown	
Time Slot	Description	Weekdays $Z_{sd\_cb\_wd}$	Weekends $Z_{sd\_cb\_we}$	Weekdays $Z_{dt\_cb\_wd}$	Weekends $Z_{dt\_cb\_we}$
H1	00:00 to 00:59	402.1	412.3	131.3	120.0
H2	01:00 to 01:59	399.5	403.9	126.4	116.0
H3	02:00 to 02:59	430.7	408.9)	121.3	1116.1
H4	03:00 to 03:59	377.6	402.6	109.0	113.1
H5	04:00 to 04:59	366.7	367.9	97.5	98.3
H6	05:00 to 05:59	367.9	367.0	97.8	98.6
H7	06:00 to 06:59	381.1	377.4	101.9	102.4
H8	07:00 to 07:59	387.4	386.5	104.4	106.4
H9	08:00 to 08:59	391.7	381.8	106.5	105.3
H10	09:00 to 09:59	390.8	388.8	106.5	108.0
H11	10:00 to 10:59	391.6	388.9	107.2	108.3
H12	11:00 to 11:59	391.7	397.5	107.3	111.6
H13	12:00 to 12:59	393.1	396.1	108.2	111.0
H14	13:00 to 13:59	394.2	399.9	108.5	112.9
H15	14:00 to 14:59	392.1	398.4	108.0	112.1
H16	15:00 to 15:59	392.1	396.3	108.0	111.3
H17	16:00 to 16:59	398.4	398.1	110.9	112.4
H18	17:00 to 17:59	411.2	392.7	116.7	110.3
H19	18:00 to 18:59	387.9	396.4	107.2	111.6
H20	19:00 to 19:59	390.7	392.3	108.0	110.1
H21	20:00 to 20:59	405.4	394.6	114.1	110.6
H22	21:00 to 21:59	428.3	397.5	123.4	111.8
H23	22:00 to 22:59	441.2	392.1	129.2	110.0
H24	23:00 to 23:59	428.0	409.3	132.8	118.6

3.2.1 Temporal variation factor (t-value)

The temporal variation factor (T-value) is defined as a value of factor multiples with the frequency number of hourly average Twitter user in each census block or land use polygon. A temporal factor was based on hourly frequency changes of unique Twitter users within the County of San Diego. Figure 8 illustrated the creation of temporal variation factor (T-value). First of all, we calculate the total number of unique Twitter users in the whole San Diego County at each hour (from 0am, 1am, 2am ...). Then we select the highest number (at 18:00-18:59 or H19, 75690) as the base number (T-value = 1). Each T-value is calculated using the base number (75690) divided by the total unique Twitter user numbers in each time slot. For example, the T-value at 4am will be 75690 / 5481 = 13.81.

Figure 9 shows the original unique Twitter user density map (a) and the estimated population density map (b) with temporal variation factor (T-value = 3.82) from 0:00 to 0:59 in San Diego downtown for Weekdays in 2015. As Figure 9(b) shows, estimated population with temporal variation factor at H1(0:00-0:59) is the result of the population in every census block increased by T-value times. Given that people less likely tweet during nighttime, temporal variation factor tends to be exaggerated during those hours.

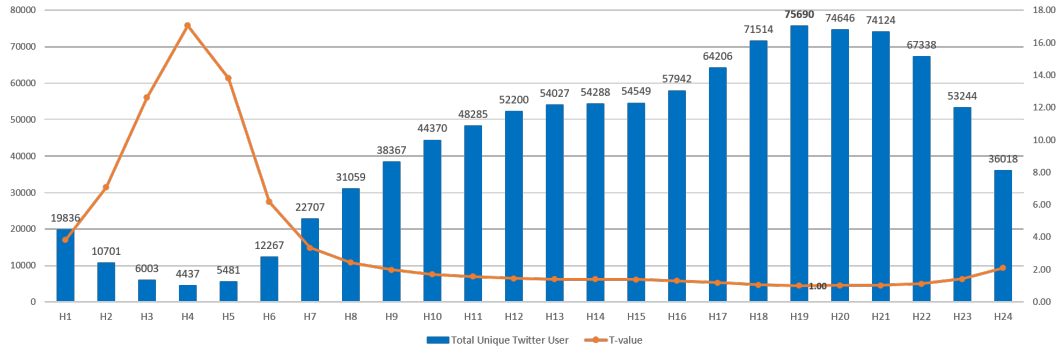


Figure 8 The total unique Twitter user numbers in each time slot and their T-values.

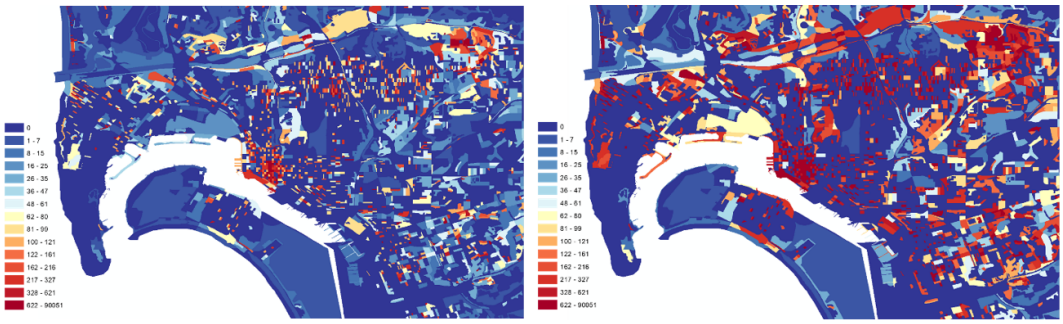


Figure 9 The original unique Twitter user density map (a) and the population density estimation (b) with temporal variation factor (T-value = 3.82) from 0:00 to 0:59 in San Diego County during weekdays in 2015.

3.2.2 Spatial change factor using dasymetric mapping method (s-value)

We utilized dasymetric mapping technique to redistribute the unique Twitter user population based on the ratio of average census population and the average hourly unique Twitter user population in each type of land use categories. Various human activities happen at a certain time in a certain land use type. For example, people would shop at shopping malls during its open hours, meaning the population in commercial land use type during daytime. Therefore, the goal is to refine the population density maps by taking different types of land use data (residential areas, commercial areas, etc.) and census data into consideration.

The census block boundaries (43,326 polygons in San Diego County) were overlaid with the 2016 parcel land use data (189,635 polygons) which created a union map with 740,843 polygons. The parcel land use data contains 10 types of land use which include unzoned, single- family, minor multiple, restricted multiple, multiple residential, restricted commercial, commercial, industrial, agricultural, and special. We downgraded the 10 types of land cover into 6 categories which are unzoned, residential, commercial, industrial, agricultural, and special. The road section were added into the parcel shapefile by extracting the road polygons from SANDAG's land use shapefile which shares the same dimension with parcel data. The new land use map ended up with 7 types of land use in total (see Table 2). Both census population and unique Twitter user population are re-distributed from the larger census

block polygon to the finer polygons (subareas) in the overlaid map. The following formula (3) were applied to calculate the number of census population with certain land use type (a) as:

$$\widehat{SCP}_a = CP_{\mathrm{A}} \left(\frac{SA_{\mathrm{A}(a)}}{A_{\mathrm{A}}} \right) \tag{3}$$

Where:
$\widehat{SCP}_a$= the estimated count of census population in subarea of land use a;
CP_{A} = the count of census population in census block A;
$SA_{\mathrm{A}(a)}$ = the area of subarea a under census block A;
A_{A} = the area of census block A;
a = the land use type;
A = census block ID.

The method of calculating unique Twitter population (formula 3) is similar to the way of re-distributing census population, while adding the temporal variation variable (T-value) into consideration. The count of unique Twitter population in census block A during time slot hx, $TP_{hx \cap A}$ is acquired by multiplying average unique Twitter user with T-Value as:

$$TP_{hx \cap A} = tp_{hx \cap A} \left(T_{hx} \right) \tag{4}$$

Where:
$tp_{hx \cap A}$ = the count of original Twitter population in census block A during time slot hx;
T_{hx} = T-Value for certain time slot hx.

The estimated count of unique Twitter population in each subarea is then calculated based on the ratio of the size of subarea and area of census block A.

$$\widehat{STP}_{hx \cap a} = TP_{hx \cap A} \left(\frac{SA_{\mathrm{A}(a)}}{A_{\mathrm{A}}} \right) \tag{5}$$

Where:
$\widehat{STP}_{hx \cap a}$ = the estimated count of unique Twitter population during time slot hx in subarea of land use A;
$TP_{hx \cap A}$ = the count of unique Twitter population in census block A during time slot hx.

The estimated population density $\widehat{D}_{hx \cap a}$ aims to estimate the hourly human population based on the ratio of the sum of census population in land use Type a and the sum of hourly unique Twitter user population in land use Type a. The ratio (R_{A}) is defined as:

$$R_{\mathrm{A}} = \frac{\sum \widehat{SCP}_a}{\sum \widehat{STP}_{hx \cap a}} \tag{6}$$

$$\widehat{D}_{hx \cap a} = R_{\mathrm{A}} \left(\frac{\widehat{STP}_{hx \cap a}}{SA_{\mathrm{A}(a)}} \right) \tag{7}$$

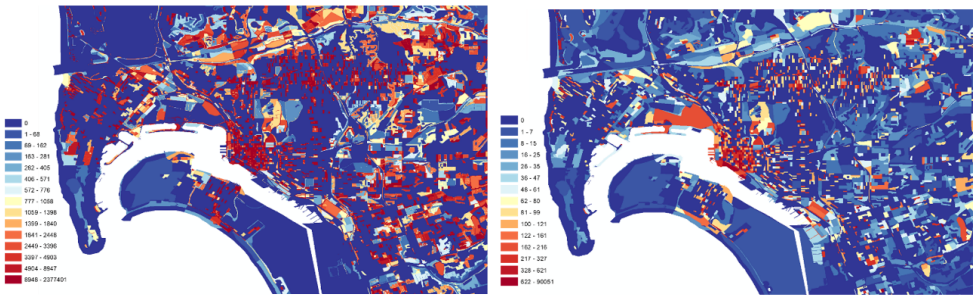
While the estimated population density $\widehat{D}_{hx \cap a}$ for certain land use type is the estimated count of unique Twitter population with R_A and divided by the size of the corresponding subarea as formula (7). Table 2(a) shows the area of 7 land use types in square kilometer, the total number of estimated unique Twitter population during H7 (6:00 to 6:59) and H21 (20:00 to 20:59) after applying T-value, and the estimated census population based on different types of land use. Table 2(b) shows the ratio (R_{A}) which was calculated based on the division of cenpop ($\sum \widehat{SCP}_a$) with twepop_h7 ($\sum \widehat{STP}_{h7 \cap a}$)) and twepop_h21 ($\sum \widehat{STP}_{h21 \cap a}$).

Table 2 (a) The area of seven types of land use, the total number of estimated unique Twitter user population during 6:00 to 6:59 (twepop_h7) and 20:00 to 20:59 (twepop_h21), and the total number of estimated census population (cenpop) based on land use. (b) The Ratio for estimating the h7 (6:00 to 6:59) and h21 (20:00 to 20:59) real population and its corresponding land use type.

		(a)				(b)	
LC	Land Use	Area(km^2)	twepop_h7	twepop_h21	cenpop	ratio_h7	ratio_h21
0	Unzoned	6437.72	65.61	46.37	99553.74	1517.25	2147.05
1	Residential	1626.62	96.56	109.85	1128499.76	11686.73	10272.84
2	Commercial	394.29	42.07	49.32	112381.49	2671.32	2278.48
3	Industrial	322.65	21.50	18.15	24372.34	1133.52	1342.97
4	Agricultural	1704.09	2.97	2.73	15602.81	5252.44	5708.23
5	Special	291.88	8.43	7.13	24342.04	2889.01	3413.78
6	Road	285.68	52.55	55.82	392448.08	7467.56	7030.45

4 Results

Figure 10 and Figure 11 show the preliminary result of applying dasymetric mapping equations (4) and (5) to adjust and re-distribute hourly unique Twitter user population into estimated population density. The purpose of the comparison between maps is not to examine the difference in numbers in each census. Instead, the focus is to visually compare the relative distribution of areas with high and low frequency between the two maps.



(a) (b)

Figure 10 (a) Population density estimation with spatial variation factor and the dasymetric mapping method from 6:00 to 6:59 in San Diego downtown areas during Weekdays in 2015; (b) the original hourly unique Twitter user density from 6:00 to 6:59 in San Diego downtown areas during Weekdays in 2015.

Table 2(b) shows that the value of residential area is higher than the values of the rest 6 types of land use types due to the influence brought by census block data. Therefore, when the estimated population is calculated by reflecting temporal variation factors and spatial change factor, more population can be redistributed to the residential area. Based on the side by side comparison of estimated population density and the original unique Twitter user population, the estimated population is transformed by the landuse types and more population is redistributed on the residential area than the rest 6 types of land use due to the influence brought by overlaying landuse with census data. Since census data represents the count of population at home, the dasymetric mapping methods could improve the estimations using Twitter density maps and to adjust the shortage of the people who may not tweet much when they are sleeping or at home. Figure 10(a) shows more population in residential area instead of the original situation where downtown areas have higher density population.

In Figure 11, the maps show the comparison of the estimated population density map (a) and the original unique Twitter user population density map (b) from 20:00 to 20:59 during weekdays in San Diego downtown areas, with the 2010 population density based on 2010 census data (c). The result shows that dasymetric mapping technique could provide a balanced population estimation comparing to the hourly unique Twitter user density and the census (night-time only) population. Comparing to the Twitter density map in the same time slot (from 20:00 to 20:59), high population density areas, such as Balboa Park and San Diego Zoo, shopping malls, and San Diego International Airport, are better estimated by using dasymetric maps with Twitter user population density data and landuse data.

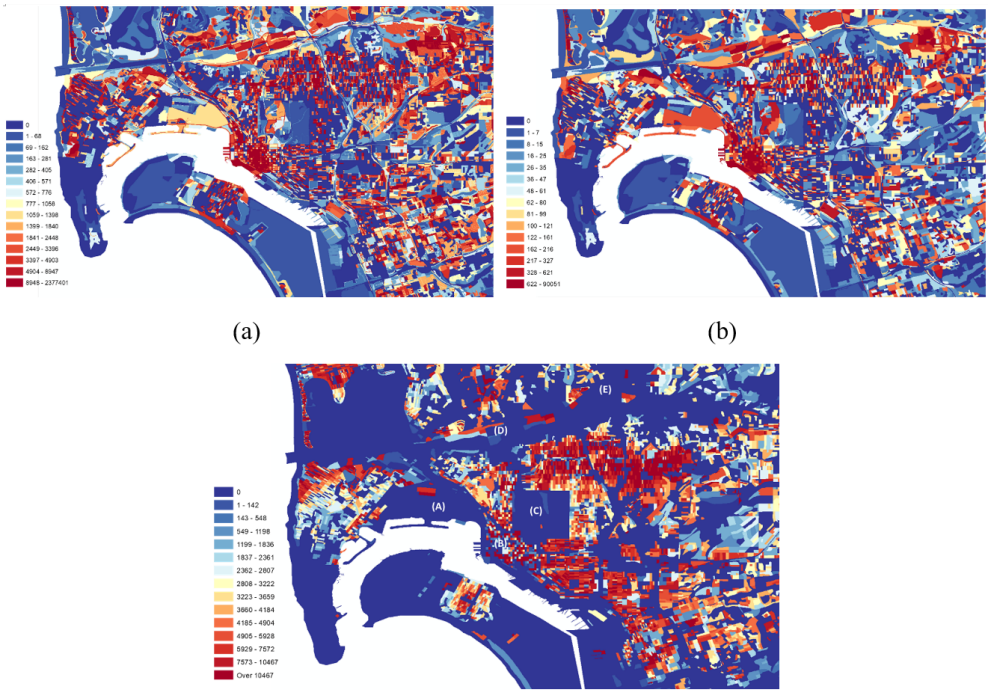


Figure 11 (a) Population density estimation map with the spatial variation factor and dasymetric mapping method from 20:00 to 20:59 in San Diego downtown areas during Weekdays in 2015.; (b) the original hourly unique Twitter user density map (middle) from 20:00 to 20:59 in San Diego downtown areas during Weekdays in 2015; (c) the 2010 census block population density map using census data.

5 Limitations and future study

There are several research limitations in our study as the following:

(a) Geo-tagged Twitter users can not represent the total population. In general, social media users are younger comparing to the general population, and more users live in urban areas than rural areas [10].

(b) It is very difficult to validate our dynamic population model because there is no similar data existed in San Diego County. We can only estimate the night time population to compare to the actual 2010 census data. However, these data are not created originally for displaying the dynamic hourly population density and may not be suitable for the validation purpose.

(c) Spatial and temporal factors in population estimation are usually correlated and should be considered together [2]. Our simplified model does not consider the autocorrelation between the spatial and temporal factors.

(d) This study only utilizes one single social media data (Twitter) among many kinds of them. For sustainability, we should consider combining other social media, such as Instagram, Facebook check-in, Foursquare, and other possible digital footprints to enhance our population model. However, different types of social media platforms and digital footprints may have different types of spatiotemporal patterns, which will be another challenge research question.

(e) The public Streaming APIs provided by Twitter is not very stable. We found that unequal number of tweets collect in different months and days, which may create some biases in our estimation of population density. For example, the Twitter use activities during March and April may more influence to the final population estimation result.

To improve and refine our future study of population density models, we are planning to use more complicated dasymetric mapping methods similar to intelligent dasymetric mapping technique (IDM)[21] to calculate the probability of population distribution in a more detailed land use category and census blocks using other spatial statistic methods, such as Weighted Linear Combination (WLC). We recognized that validation is a key challenge to evaluate our dynamic population estimation model. While collecting dynamic population from real world in a large area is extremely difficult, it might be possible to partially compare the estimate during a certain temporal duration with existing data. For example, Census American Community Survey (ACS) provides a daytime population estimate [20]. Therefore, we can measure the goodness of fit between the estimates from the model and ACS during daytime (e.g., 9 am to 3 pm, a core work hour). However, it is necessary to carefully consider the validation process since social media data are drawn from potentially biased population and the data may include not only local residents but also visitors whereas ACS data account for residents and workers. Taking visitors in San Diego into consideration can be helpful for revealing the real pattern of human dynamic. Therefore, further social media data filtering procedures should be applied to identify local residents for validation. While data at finer spatial and temporal scales can provide better understanding of human movement, it can raise privacy concerns. Population estimation needs to find balance between privacy and accuracy. Within the context of social media studies, fine scale results are not the most appropriate because they can reveal users' location. Results should be aggregated to the point at which they show significance without jeopardizing users' privacy. Therefore, researchers should ask how fine does the data need to be to protect users' privacy while also providing meaningful results. Methods such as data anonymization and using aggregated data to mitigate privacy risks can be considered.

The finalized framework, with frontend web design and backend database, can be applied with real-time data as well in the future by upgrading the current 1 hour temporal resolution to 10 minutes or even higher scale. To summarize, although the Twitter data cannot perfectly represent the entire population, this study has revealed the potential research framework using social media data and dasymetric maps to calculate the dynamic change of population distribution patterns. Our proposed methods can provide a better estimation of hourly population patterns in airports, sports stadiums, shopping malls, downtown areas, parks and other tourist locations comparing to traditional census data or ACS data.

The combination of multiple social media data, mobile phone records, and other digital footprints created by human beings will be a great source to study human dynamics and help us to understand different types of human behaviors, movements, and activities in high spatial

and temporal resolution. This integration of utilizing multiple sources of information would be able to increase the demographic comprehensiveness of this research. This information can facilitate the improvement of our transportation systems, emergency evacuation procedures, and urban planning in the future.

References

1 Rein Ahas, Anto Aasa, Y Yuan, Martin Raubal, Zbigniew Smoreda, Yu Liu, Cezary Ziemlicki, Margus Tiru, and Matthew Zook. Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in harbin, paris, and tallinn. *International Journal of Geographical Information Science*, 29(11):2017–2039, 2015.

2 Li An, Ming-Hsiang Tsou, Stephen ES Crook, Yongwan Chun, Brian Spitzberg, J Mark Gawron, and Dipak K Gupta. Space–time analysis: Concepts, quantitative methods, and future directions. *Annals of the Association of American Geographers*, 105(5):891–914, 2015.

3 Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

4 Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan Von Schreeb. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti. *PLoS medicine*, 8(8), 2011.

5 Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Marie L Urban. Landscan usa: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69(1-2):103–117, 2007.

6 Christoph Bregler. Learning and recognizing human dynamics in video sequences. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 568–574. IEEE, 1997.

7 Zi Chu, Indra Widjaja, and Haining Wang. Detecting social spam campaigns on twitter. In *International Conference on Applied Cryptography and Network Security*, pages 455–472. Springer, 2012.

8 Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

9 Pinliang Dong, Sathya Ramesh, and Anjeev Nepali. Evaluation of small-area population estimation using lidar, landsat tm and parcel data. *International Journal of Remote Sensing*, 31(21):5571–5586, 2010.

10 Maeve Duggan, Nicole B Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. Social media update 2014. *Pew research center*, 19, 2015.

11 Jim Edwards. Leaked twitter api data shows the number of tweets is in serious decline. *Business Insider*, February 2016. URL: `http://www.businessinsider.com/tweets-on-twitter-is-in-serious-decline-2016-2`.

12 Cory L Eicher and Cynthia A Brewer. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, 28(2):125–138, 2001.

13 Su Yeon Han, Ming-Hsiang Tsou, and Keith C Clarke. Do global cities enable global views? using twitter to quantify the level of geographical awareness of us cities. *PloS one*, 10(7), 2015.

14 Su Yeon Han, Ming-Hsiang Tsou, and Keith C Clarke. Revisiting the death of geography in the era of big data: the friction of distance in cyberspace and real space. *International Journal of Digital Earth*, 11(5):451–469, 2018.

15 Su Yeon Han, Ming-Hsiang Tsou, Elijah Knaap, Sergio Rey, and Guofeng Cao. How do cities flow in an emergency? tracing human mobility patterns during a natural disaster with big data and geospatial data science. *Urban Science*, 3(2):51, 2019.

16 Yusuke Hara and Masao Kuwahara. Traffic monitoring immediately after a major natural disaster as revealed by probe data–a case in ishinomaki after the great east japan earthquake. *Transportation research part A: policy and practice*, 75:1–15, 2015.

17 James B Holt, CP Lo, and Thomas W Hodler. Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science*, 31(2):103–121, 2004.

18 Elias Issa, Ming-Hsiang Tsou, Atsushi Nara, and Brian Spitzberg. Understanding the spatio-temporal characteristics of twitter data with geotagged and non-geotagged content: two case studies with the topic of flu and ted (movie). *Annals of GIS*, 23(3):219–235, 2017.

19 Bin Jiang, Ding Ma, Junjun Yin, and Mats Sandberg. Spatial distribution of city tweets and their densities. *Geographical Analysis*, 48(3):337–351, 2016.

20 Brian McKenzie, William Koerber, Alison Fields, Megan Benetsky, and Melanie Rapino. Commuter-adjusted population estimates: Acs 2006-10. *Washington, DC: Journey to Work and Migration Statistics Branch, US Census Bureau*, 2010.

21 Jeremy Mennis and Torrin Hultgren. Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, 33(3):179–194, 2006.

22 Anna C Nagel, Ming-Hsiang Tsou, Brian H Spitzberg, Li An, J Mark Gawron, Dipak K Gupta, Jiue-An Yang, Su Han, K Michael Peddecord, Suzanne Lindsay, et al. The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. *Journal of medical Internet research*, 15(10):e237, 2013.

23 Atsushi Nara, Xianfeng Yang, Sahar Ghanipoor Machiani, and Ming-Hsiang Tsou. An integrated evacuation decision support system framework with social perception analysis and dynamic population estimation. *International journal of disaster risk reduction*, 25:190–201, 2017.

24 Tao Pei, Stanislav Sobolevsky, Carlo Ratti, Shih-Lung Shaw, Ting Li, and Chenghu Zhou. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28(9):1988–2007, 2014.

25 Ming-Hsiang Tsou. Research challenges and opportunities in mapping social media and big data. *Cartography and Geographic Information Science*, 42(sup1):70–74, 2015.

26 Ming-Hsiang Tsou, Ick-Hoi Kim, Sarah Wandersee, Daniel Lusher, Li An, Brian Spitzberg, Dipak Gupta, Jean Mark Gawron, Jennifer Smith, Jiue-An Yang, et al. Mapping ideas from cyberspace to realspace: visualizing the spatial context of keywords from web page search results. *International Journal of Digital Earth*, 7(4):316–335, 2014.

27 Ming-Hsiang Tsou and Michael Leitner. Visualization of social media: seeing a mirage or a message?, 2013.

28 Jessica JunLin Wang and Sameer Singh. Video analysis of human dynamics—a survey. *Real-time imaging*, 9(5):321–346, 2003.

29 John K Wright. A method of mapping densities of population: With cape cod as an example. *Geographical Review*, 26(1):103–110, 1936.

30 Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. Detecting spam in a twitter network. *First Monday*, 15(1), 2010.