ISPRS Journal of Photogrammetry and Remote Sensing 146 (2018) 1–10

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Tweets or nighttime lights: Comparison for preeminence in estimating socioeconomic factors

Naizhuo Zhao[a,b,*], Guofeng Cao[a,b], Wei Zhang[c], Eric L. Samson[d]

[a] Center for Geospatial Technology, Texas Tech University, Lubbock, TX 79409, USA
[b] Department of Geosciences, Texas Tech University, Lubbock, TX 79409, USA
[c] Department of Computer Science, Texas Tech University, Lubbock, TX 79409, USA
[d] Mayan Esteem Project, 222 Main Street, Suite 204, Farmington, CT 06032, USA

ARTICLE INFO

Keywords:
Nighttime lights imagery
Twitter
Socioeconomic factors
Location-based social media
The United States

ABSTRACT

Nighttime lights (NTL) imagery is one of the most commonly used tools to quantitatively study socioeconomic systems over large areas. In this study we aim to use location-based social media big data to challenge the primacy of NTL imagery on estimating socioeconomic factors. Geo-tagged tweets posted in the contiguous United States in 2013 were retrieved to produce a tweet image with the same spatial resolution of the NTL imagery (i.e., 0.00833° × 0.00833°). Sum tweet (the total number of tweets) and sum light (summed DN value of the NTL image) of each state or county were obtained from the tweets and the NTL images, respectively, to estimate three important socioeconomic factors: personal income, electric power consumption, and fossil fuel carbon dioxide emissions. Results show that sum tweet is a better measure of personal income and electric power consumption while carbon dioxide emissions can be more accurately estimated by sum light. We further exploited that African-Americans adults are more likely than White seniors to post geotagged tweets in the US, yet did not find any significant correlations between proportions of the subpopulations and the estimation accuracy of the socioeconomic factors. Existence of saturated pixels and blooming effects and failure to remove gas flaring reduce quality of NTL imagery in estimating socioeconomic factors, however, such problems are non-existent in the tweet images. This study reveals that the number of geo-tagged tweets has great potential to be deemed as a substitute of brightness of NTL to assess socioeconomic factors over large geographic areas.

1. Introduction

Remote sensing is not only an effective tool to observe environmental changes but also has an extensive record of successful application by creating proxies for quantitative data of human systems. Since Croft (1973) first recognized the potential of nighttime lights (NTL) satellite images to monitor emissions of waste gas from oil fields, NTL images have been used to detect human activities and their impacts on natural systems (e.g. Elvidge et al., 1997; Gallo et al., 2004; Imhoff et al., 1997a; Milesi et al., 2003). Estimating socioeconomic factors (e.g., gross domestic product, electric power consumption, and fossil fuel carbon dioxide (CO_2) emissions) is an important application aspect of NTL images (e.g. Chen and Nordhaus, 2010; Doll et al., 2000, 2006; Ghosh et al., 2009; Ghosh et al., 2010a, 2010b; Henderson et al., 2011; Letu et al., 2010; Lo, 2002; Oda and Maksyutov, 2011; Sutton et al., 2007; Zhao et al., 2011; Zhao et al., 2012, 2015). Such applications became more extensive and reliable after the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Information (NCEI) (formerly National Geophysical Data Center (NGDC)) released the Defense Meteorological Satellite Program's Operational Linescan System (DMSP-OLS) stable light time-series image products in which ephemeral lights (typically fires) were removed (Doll, 2008). However, the nature of NTL (e.g. blooming effect) and drawbacks of the current stable light image products (e.g., existence of saturated pixels) adversely affect estimation of NTL images on the socioeconomic parameters (Doll, 2008; Letu et al., 2010; Zhao et al., 2015). Despite these limitations and shortcomings, NTL imagery is still one of the most widely used tools in quantitatively evaluating socioeconomic systems over large areas given its efficiency and economy (Chen and Nordhaus, 2010).

The advent of the social media big data revolution provides great opportunities to further understand anthropogenic activities and Earth changes (Bik and Goldstein, 2013; Sloan and Morgan, 2015). Hundreds of millions of social media users share their observations on the

* Corresponding author at: Center for Geospatial Technology, Texas Tech University, Lubbock, TX 79409, USA.
E-mail address: zhao.naizhuo@gmail.com (N. Zhao).

https://doi.org/10.1016/j.isprsjprs.2018.08.018
Received 5 April 2018; Received in revised form 30 July 2018; Accepted 24 August 2018

surrounding environments and become numerous individual "citizen sensors" (Goodchild, 2007; Liu et al., 2015). Moreover, with the extensive uses of smart mobile devices, increasing social media data contain location information indicating where the data are posted. Thus, the location-based social media data can reflect not only what activities have happened but also where they occur (Sloan and Morgan, 2015).

A tweet is a brief message digitally transmitted on Twitter. As one of leading social media platforms, Twitter currently has over 336 million monthly active users publishing about 500 million tweets daily (Statista, 2018). Due to the wide uses of smart mobile devices, a massive number of the tweets are geo-tagged with longitude and latitude information showing the positions of where the Twitter users were when the tweets were posted (de Bruijn et al., 2018). Geo-tagged tweets have been used to track human mobility (e.g. Hawelka et al., 2014; Jurdak et al., 2015, Luo et al., 2016), investigate demographics of population (e.g. Luo et al., 2016; Mislove et al., 2011), analyze urban patterns (e.g. Ferrari and Mamei, 2011; Frias-Martinez et al., 2012), and surveille outbreaks of infectious diseases (e.g. Allen et al., 2016; Broniatowski et al., 2013). Furthermore, geo-tagged tweets with geographic coordinates can be flexibly converted to a space-time data cube with different resolutions (Cao et al., 2015) and easily integrated with remote sensing images (Klotz et al., 2017).

Despite the recognized potential of what has been termed as "social sensing" (Liu et al., 2015), geo-tagged tweets were rarely adopted to estimate socioeconomic factors over large areas. As brightness of NTL can be a proxy for population density and economic level (Chen and Nordhaus, 2010; Doll et al., 2000; Sutton et al., 2007), the number of active twitter users can indicate population and other human activities (Luo et al., 2016) with geo-tagged tweets accurately indicating positions where the population exists and the activities occur. Thus geo-tagged tweets should have similar capabilities of NTL images to quantitatively assess socioeconomic factors.

Accordingly, the major objective of this study is to investigate how effective the number of tweets is used as a proxy of socioeconomic factors compared to brightness of NTL in the contiguous US. In the following section we present methods converting massive individual geo-tagged tweets to a raster image making it possible to estimate personal income, electric power consumption, and CO_2 emissions just as stable lights image product can do. Section 3 exhibits results of comparing the number of tweets and brightness of NTL to estimate the socioeconomic parameters at the state and the county levels. Section 4 discusses strengths and limitations of using Twitter big data and NTL imagery to assess socioeconomic factors and the prospect of future applications of geo-tagged tweets with the new generation NTL image products, namely the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) image composites, to estimate and map human socioeconomic activities at finer spatial and temporal resolutions.

2. Data and methods

2.1. Collecting NTL imagery and official socioeconomic data

A version 4 DMSP-OLS stable lights image composite for 2013 was obtained from NOAA's NCEI (NCEI, 2018). This image product is a composite of all the available cloud-free DMSP-OLS NTL data for the 2013 calendar year in the NCEI's digital archive with a spatial resolution of 0.00833° × 0.00833° (~1 km × 1 km). Ephemeral lights such as fire, lightning, and other background noise have been removed from the image product. Digital number (DN) values in the annual stable light image composite vary from 0 to 63 and represent average brightness of NTL for the year except 63. The maximal DN value, 63, indicates saturated pixels and is not to scale of lower DN values. The saturated pixels usually exist in urban core areas of the US while gas flaring of petroleum production can also result in a certain number of saturated pixels in rural areas (see Fig. 1a).

A comprehensive dataset on personal income was retrieved by the Interactive Data Application developed by the US Department of Commerce's Bureau of Economic Analysis (BEA) (BEA, 2018). The dataset contains three variables: population, per capita personal income, and total personal income for 49 states including the District of Columbia and 3057 counties in the contiguous US.

Electric power consumption data were taken from the US Energy Information Administration (EIA) (The U.S. EIA, 2018), at the state-level with county-level data unavailable. County-level electricity data were obtained from the Energy Consumption Data Management System (ECDMS) of the California Energy Commission (CEC) (CEC, 2018), but the data are only for the counties in California. State-level CO_2 emissions data were also retrieved from the US EIA. However, we did not find any official CO_2 emissions datasets publicly released and reported at the county level. Thus, in this study we only compared the capabilities of geotagged tweets and NTL imagery estimating the amount of CO_2 emissions at the state level.

A demographic dataset was retrieved from the US Census Bureau by the use of a R package namely "censusapi" (Recht, 2018). This demographic dataset includes each county's subpopulations by races (i.e. White, Black, Hispanic, and Asian) and age groups (i.e. under 18 years, 18–64 years, and 65 years and over) and was utilized in this study to examine whether some subpopulations tend to or not to post geo-tagged tweets, resulting in predictable errors when the geo-tagged tweets were employed to estimate socioeconomic factors.

2.2. Producing Twitter imagery

Geo-tagged tweets posted between January 1, 2013 and December 31, 2013 were collected by the public Twitter Stream API and archived to a NoSQL database. To eliminate spam, bot, and cyborg tweets (data noises) to be used in estimating the socioeconomic factors (Tsou et al., 2017), we adopted a simple procedure to filter out the Twitter users if their geo-tagged tweets were posted in two (or more) far-away (> 500 km) places within a very short period (< 1 h) (Luo et al., 2016). Next, an empty frame with an extent of 66.00002–125.00000°W and 24.00003°N–50.00003°N was established to cover the entire contiguous US and northern regions of Mexico. All geo-tagged tweets located in this frame were selected from the archive and the frame was further divided into 22,089,600 cells (3120 rows × 7080 columns) with uniform size of 0.00833° × 0.00833°. We counted the tweets located in each of the cells and valued the cell with the count. When NTL imagery was used to estimate or spatially disaggregate the socioeconomic factors, a region's brightness of NTL was deemed as a proxy of population density or economic level of the area while lit area was used to indicate the extent that human activities reached (Zhao et al., 2017a, 2017b). As the uses of NTL imagery in estimating the socioeconomic factors, we processed geo-tagged tweets from two aspects: 1) the number of tweets was used to indicate population and the number of socioeconomic activities and thus, we counted multiple tweets reported in one cell by one ID user as one tweet; 2) locations of the tweets were used to delineate extents of occurrence of these socioeconomic activities and so multiple tweets reported in different cells by one ID user were counted multiple times for the individual cells. The valued grid was then cropped by the contiguous US boundary and exported as a 32-bit tweet image. The choice of 32-bit ensures that the tweet image has a sufficiently large quantization range to avoid saturation on pixel values. The tweet image was ultimately populated by 58,272,064 individual tweets after repetitive tweets reported in one cell by one ID user were removed (see Fig. 1b).

2.3. Estimating the socioeconomic factors and evaluating the estimate results

To explore whether the number of tweets has similar capabilities to brightness of NTL as an indicator of the three socioeconomic factors (i.e. personal income, electric power consumption, and CO_2 emissions),