Applied Geography 73 (2016) 77–88

Contents lists available at ScienceDirect

Applied Geography

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

Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity

Quynh C. Nguyen [a, *], Suraj Kath [b], Hsien-Wen Meng [a], Dapeng Li [c], Ken R. Smith [d], James A. VanDerslice [e], Ming Wen [f], Feifei Li [b]

[a] Department of Health Promotion and Education, College of Health, University of Utah, Salt Lake City, UT, USA
[b] School of Computing, University of Utah, USA
[c] Department of Geography, University of Utah, USA
[d] Department of Family and Consumer Studies, University of Utah, USA
[e] Division of Public Health, Department of Family and Preventive Medicine, School of Medicine, University of Utah, USA
[f] Department of Sociology, University of Utah, USA

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form
17 June 2016
Accepted 19 June 2016
Available online 1 July 2016

Keywords:
Twitter messaging
Neighborhood
Happiness
Physical activity
Diet
Food

ABSTRACT

Objectives: Using publicly available, geotagged Twitter data, we created neighborhood indicators for happiness, food and physical activity for three large counties: Salt Lake, San Francisco and New York.
Methods: We utilize 2.8 million tweets collected between February–August 2015 in our analysis. Geocoordinates of where tweets were sent allow us to spatially join them to 2010 census tract locations. We implemented quality control checks and tested associations between Twitter-derived variables and sociodemographic characteristics.
Results: For a random subset of tweets, manually labeled tweets and algorithm labeled tweets had excellent levels of agreement: 73% for happiness; 83% for food, and 85% for physical activity. Happy tweets, healthy food references, and physical activity references were less frequent in census tracts with greater economic disadvantage and higher proportions of racial/ethnic minorities and youths.
Conclusions: Social media can be leveraged to provide greater understanding of the well-being and health behaviors of communities—information that has been previously difficult and expensive to obtain consistently across geographies. More open access neighborhood data can enable better design of programs and policies addressing social determinants of health.

1. Introduction

The literature examining neighborhood effects on health has flourished in the last decade (Diez Roux, 2001). Extant research has provided evidence on associations between the neighborhood environment and mortality risk (Eames, Ben-Shlomom, & Marmot, 1993; Morris, Blane, & White, 1996; Townsend, Phillimore, & Beattie, 1988; Tyroler et al., 1993; Waitzman & Smith, 1998a, 1998b; Wing, Barnett, Casper, & Tyroler, 1992), life expectancy (Clarke et al., 2010), mental health (Truong & Ma, 2006), self-rated health (Wen, Browning, & Cagney, 2003), obesity, (Black, Macinko, Dixon, & Fryer, 2010; Heinrich et al., 2008; Mujahid et al., 2008; Smith et al., 2008), and diabetes (Grigsby-Toussaint et al., 2010; Lysy et al., 2013)—even after adjusting for individual characteristics. Poor access to healthy food (Christiansen, Qureshi, Schaible, Park, & Gittelsohn, 2013; Inagami, Cohen, & Finch, 2006; Morland, Wing, Diez Roux, & Poole, 2002; Morland, Wing, Roux, 2002; Stafford, 2007; Wang, Kim, Gonzalez, MacLeod, & Winkleby, 2007), fast food chains (Block, Scribner, & DeSalvo, 2004), the lack of recreational facilities (Brownson, Hoehner, Day, Forsyth, & Sallis, 2009; Roemmich et al., 2006), and higher crime rates (Mujahid et al., 2008; Stafford, 2007) all correlate with higher obesity rates. Community happiness levels also have been inversely related to obesity as well as other outcomes including hypertension, suicide, and life expectancy (Blanchflower & Oswald, 2008; Bray & Gunnell, 2006; Di Tella & MacCulloch, 2008; Dodds, Harris, Kloumann, Bliss, & Danforth, 2011; Oswald & Powdthavee, 2007; Tella, MacCulloch, & Oswald, 2003). Adverse neighborhood conditions concentrate in poor, minority neighborhoods (Black et al., 2010; Diez-Roux,

* Corresponding author. Department of Health Promotion and Education, University of Utah, 1901 E South Campus Drive, Annex B 2124, Salt Lake City, UT 84112, USA.
E-mail address: quynh.nguyen@health.utah.edu (Q.C. Nguyen).

http://dx.doi.org/10.1016/j.apgeog.2016.06.003
0143-6228/

1998; Duncan, Jones, & Moon, 1998; Macintyre, Maciver, & Sooman, 1993), thereby increasing health disparities. Furthermore, the epidemic rise in obesity and related chronic diseases in recent decades signal the importance of structural forces and social processes.

Nonetheless, the dearth of data on contextual factors limits the investigation of multilevel effects on health. Certain places (National Archive of Criminal Justice; Baltimore Neighborhood Indicators Alliance – The Jacob France Institute) have extensive neighborhood data collected on them, but they are the exception rather than the rule, and it is difficult to make comparisons across geographies because available measures vary greatly across them. Also patterns seen in specific places may not apply to other places. For instance, estimates and patterns seen in urban areas may not apply to rural areas. Neighborhood data collection is expensive and time consuming, and then only available for certain places or time periods and become outdated quickly (Peterson and Krivo, 2000). Moreover, while comparable neighborhood data across large areas are highly lacking, the neighborhood data we do have are typically data on compositional characteristics (e.g., percent females) and features of the built environment (e.g., number of grocery stores and health care clinics). These data do not capture the social environment, or an individual's interactions with that environment.

Social processes and networks can affect health via a myriad of mechanisms, such as 1) the maintenance of norms around healthy behaviors via informal social control; 2) the stimulation of new interests such as a new sport or exercise; 3) political advocacy for access to neighborhood amenities and protection against stressors and toxic agents; 4) emotional support; and 5) the dispersal of knowledge about health promotion practices (Ali, Amialchuk, & Heiland, 2011; Berkman & Syme, 1979; Cohen, Finch, Bower, & Sastry, 2006; Kim, Subramanian, Gortmaker, & Kawachi, 2006; Vartanian, Sokol, Herman, & Polivy, 2013). According to Social Learning Theory, learning takes place in a social context (Bandura, 1977). Behaviors are adopted by observing how the behavior is performed by others, attitudes around that behavior, and outcomes associated with that behavior. Empirically, the adoption of specific health behaviors related to food consumption, health screening, smoking, alcohol consumption, drug use, and sleep has been observed to disperse through social networks (Keating, O'Malley, Murabito, Smith, & Christakis, 2011; Mednick, Christakis, & Fowler, 2010; Pachucki, Jacques, & Christakis, 2011; Rosenquist, Murabito, Fowler, & Christakis, 2010; Roy, 2004; Smith & Christakis, 2008). Similarly, evidence suggests that emotional states such as mood (Kramer, Guillory, & Hancock, 2014), happiness (Fowler & Christakis, 2008), depression (Rosenquist, Fowler, & Christakis, 2011), and suicidality (Bearman, & Moody, 2004) can spread through social networks. The measurement of area-level happiness and subjective-well-being is a new and expanding research endeavor (Gallup-Healthways, 2013; Gill, French, Gergle, & Oberlander, 2008; Helliwell, Layard, & Sachs, 2012; Kramer, 2010; Quercia, Ellis, Capraz, & Crowcroft, 2012). For instance, in 2012, the United Nations began its annual release of a World Happiness Report (Helliwell et al., 2012). Social media may influence individuals' health behaviors but may also be a way to characterize prevalent community characteristics and patterns of behaviors.

1.1. Study aims

Given the vast literature documenting the influence of social networks on individual health behaviors and health outcomes, we believe that social media data represent an important new data resource for neighborhood researchers. Thus, using publicly available, geotagged Twitter data, we construct novel indicators of neighborhood happiness levels, healthiness of food, and physical activity. We conduct quality control activities and perform validation analysis comparing Twitter-derived neighborhood indicators to demographic and economic characteristics of the corresponding census tract. In order to test our computer algorithm for constructing neighborhood indicators, we selected three counties that display diversity in regards to geographical location, landscape, housing market, cultural characteristics, and demographic characteristics (e.g., racial/ethnic composition, age distribution, and household size). The three counties are the following: Salt Lake County, San Francisco County, and New York County.

2. Methods

2.1. Social media data collection

From February–August 2015, we utilized Twitter's Streaming Application Programming Interface (API) to continuously collect a random 1% subset of publicly available tweets with latitudes and longitude coordinates. We present in-depth analyses and findings for three counties in the United States: Salt Lake County (367,204 tweets); San Francisco County (same as San Francisco city; 653,670 tweets); and New York County (1,828,026 tweets).

2.2. Spatial join

We linked 99.8% of tweets with available GPS coordinates to their respective 2010 census tract locations. We used Python and relevant GIS libraries (Shapely and Fiona) to accomplish this task. An R-Tree was used to build a spatial index (Guttman, 1984) on census tract polygon data. R-tree indexing allows for faster spatial searches on the data because the R-tree groups data into bounding rectangles and narrows the search space. A query that does not intersect the bounding rectangle cannot intersect any of its component parts. Utilizing the latitude and longitude coordinates of where tweets were sent, spatial joins were performed on tweets to identify the corresponding census tracts. Tweets that were not assigned a census location included those with destinations bordering the United States (i.e., Mexico and Canada).

2.3. Processing tweets

We processed tweets to create variables that measure sentiment, food, and physical activity. To accomplish this task, we utilized a bag-of-words algorithm which creates a simplifying representation of tweets that disregards grammar and word order, but has the capacity to track the frequency of terms or components of tweets, and then performs computations on those components and terms. Several steps were conducted that first included dividing each tweet into tokens (Stanford Natural Language Processing Group). A tokenizer divides text into a sequence of tokens, which roughly correspond to "words." We are using a tokenizer particularly suitable for processing English text called the PTBTokenizer (aka the Stanford Tokenizer). The PTBTokenizer is an efficient, fast, and deterministic tokenizer. It can tokenize text at a rate of about 1,000,000 tokens per second on a standard personal computer. Utilizing heuristics, it can usually differentiate when single quotes are part of words and when periods do and do not imply sentence boundaries. After we obtain the tokens (i.e., individual words) from a tweet, we then search each word in our word dictionary to get its corresponding happiness score for sentiment analysis. Currently, we are ignoring words which are not present in our word dictionary. Using this algorithm, sentiment scores can be assigned to approximately 80–85% of tweets across geographies.