Applied Geography 92 (2018) 21–30

Contents lists available at ScienceDirect

Applied Geography

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

Comparing spatial patterns of crowdsourced and conventional bicycling datasets

Lindsey Conrow[a,*], Elizabeth Wentz[a], Trisalyn Nelson[a], Christopher Pettit[b]

[a] School of Geographical Sciences and Urban Planning, Arizona State University, P.O. Box 875302, Tempe, USA
[b] City Futures Research Centre, UNSW Built Environment, University of New South Wales, Sydney New South Wales 2052, Australia

ARTICLE INFO

Keywords:
Crowdsourced data
Pattern analysis
Bicycling activity

ABSTRACT

Conventional bicycling data have critical limitations related to spatial and temporal scale when analyzing bicycling as a transport mode. Novel crowdsourced data from smartphone apps have the potential to overcome those limitations by providing more detailed data. Questions remain, however, about whether crowdsourced data are representative of general bicycling behavior rather than just those cyclists who use the apps. This paper aims to explore the gap in understanding of how conventional and crowdsourced data correspond in representing bicycle ridership. Specifically, we use local indicators of spatial association to generate locations of similarity and dissimilarity based on the difference in ridership proportions between a conventional manual count and crowdsourced data from the Strava app in the Greater Sydney Australia region. Results identify where the data correspond and where they differ significantly, which has implications for using crowdsourced data in planning and infrastructure decisions. Fourteen count locations had significant low-low spatial association; similarity was found more often in areas with lower population density, greater social disadvantage, and lower ridership overall. Five locations had high-high spatial association, or were locations of dissimilar rank values indicating that they did not have a strong spatial match. Higher coefficients of variation were associated with population density, the number of bicycle journey to work trips, and percentage of residential land use for the significant locations of dissimilarity. IRSD and bicycle infrastructure density were lower than the locations that were not significantly dissimilar. For the significant locations of similarity, all coefficient of variation measures were lower than the locations that were not significant. Areas where ridership show locations of similarity are those where it may be suitable to substitute conventional data for the more detailed crowdsourced data, given further investigation into potential bias related to rider demographics.

1. Introduction

Progress in planning and research for active transportation, or non-motorized transport modes like walking and bicycling for practical purposes, is limited by a lack of data related to where and when people use those modes. Since active transport trips tend to be shorter in duration and occur at finer-scale levels of movement, they necessarily require detailed fine-resolution data for analysis at the neighborhood level where the activity, and therefore policy and planning decisions, occur (Cervero & Duncan, 2003). The lack of reliable data and knowledge about non-motorized travel has limited measures of accessibility and examinations related to human mobility using those modes (Iacono, Krizek, & El-Geneidy, 2010). Further, there is a lack of data that can be used to link non-motorized transport behavior with infrastructure and other network features that may influence it (Broach, Dill, & Gliebe, 2012). This research has two goals: first to determine how crowdsourced and conventional bicycling data correspond in representing bicycling activity and second to determine how that correspondence helps understand factors that underlie bicycling activity.

For bicycling in particular, data limitations are associated with the sampling strategies for conventional data collection. The primary conventional methods for collecting bicycling travel activity data are manual bicycle counts, automated bicycle counts, regional travel surveys, and direct questionnaires. Manual bicycle counts are one of the most common methods for gathering bicycling data; they are conducted by counting travel volumes at specific locations for all riders who pass the location. The advantages of counts are that they do not depend on user participation, though they also have significant disadvantages. No additional information is collected, so route information, cyclist demographics, and reasons for the trip are not included (Kuzmyak & Dill, 2014). Another disadvantage is the representation of the sample is limited both spatially (e.g., count locations may not be spatially

* Corresponding author.
E-mail addresses: lconrow@asu.edu (L. Conrow), wentz@asu.edu (E. Wentz), Trisalyn.Nelson@asu.edu (T. Nelson), c.pettit@unsw.edu.au (C. Pettit).

https://doi.org/10.1016/j.apgeog.2018.01.009
Received 10 October 2017; Received in revised form 29 December 2017; Accepted 20 January 2018

distributed in such a way to better understand problems) and temporally (e.g., typically conducted annually or semi-annually for a few hours on one or two days) (Kuzmyak & Dill, 2014; Ryus et al., 2014). One alternative to manual counts is automated count technologies that continuously collect travel volumes as riders pass the counter (Kuzmyak & Dill, 2014). While these counters solve the problem of temporal sampling, the spatial sampling problem and the lack of associated travel data remain unsolved. Regional travel surveys tend to be more comprehensive and include all travel modalities including automobile and public transportation. Travel surveys typically sample individual travel activities over a short period of time, such as a day or a week's worth of travel and as such, are able to gain more insight on route information and reasons for travel (Dill, 2009). Despite the additional information that is collected, travel surveys are often still limited in overall sample and detail. Since bicycling constitutes a comparatively small percentage of modal share, the activity may be missed completely, and route information may not be collected. If route information is inferred later, many surveys assume a cyclist takes the shortest path between an origin and destination (van Heeswijck et al., 2015) though this may not be the case as cyclists are willing to go out of their way to avoid traffic and stay on bicycling infrastructure (Dill, 2009). Information about cycling activity may also be gleaned from direct questionnaires to cyclists. Direct questionnaires have been used to examine how cyclists view urban design (Forsyth & Krizek, 2011), perceptions of risk while cycling (Lawson, Pakrashi, Ghosh, & Szeto, 2013; Møller & Hels, 2008), bicycle facility planning (Dill, 2009; Rybarczyk & Wu, 2010), and route-choice modeling (Broach et al., 2012; Dill & Gliebe, 2008). These questionnaires generate valuable demographic and experience or opinion based data, but they also have a tendency toward limited spatial coverage and small sample sizes.

Novel crowdsourced data from smartphone apps have the potential to improve on the resolution of conventional data collection methods. Data collected from personal mobile devices overcome limitations in spatial and temporal scope by both providing finer-scale specificity about the actual route and not depending on the timing of a survey. These smartphone-based geosocial networking apps utilize built-in GPS functionality to allow users to record activity locations and often have a social component in terms of connecting to, competing with, or sharing information among other users (Elwood, Goodchild, & Sui, 2012). The finer scale detail from near-continuous time frames is needed to generate knowledge about the neighborhood and network contexts that drive behavior associated with non-motorized modes. Detailed GPS data sets can be used to examine both individual and group level movement behaviors, which allows for broader application contexts as well (Meijles, de Bakker, Groote, & Barske, 2014). The challenges with using crowdsourced bicycling data are the inherent biases due to self-selected participation. Users who generate data using smartphone technology are limited to those who have access, have the motivation to participate, and who have the resources (e.g., money, time) required to take part (Goodchild, 2007; Heipke, 2010). This means that crowdsourced data sources may be biased and limiting in terms of extensive and generalizable representation of greater populations, despite potential benefits associated with the detail and information they may provide. Groups such as commuters, students, children, and average recreational riders could be missed completely. This is problematic because relying on biased information could lead to increased inequities in transportation planning and policy.

There are pertinent questions related to the effectiveness of these crowdsourced data for understanding the drivers of bicycling behavior because the data may be biased or of poor representative quality. Studies that have compared crowdsourced bicycling data to conventional bicycling data have found similar ridership volumes with correspondence closest when volumes were grouped categorically such as low, medium, and high volumes or according to peak hours, suggesting that spatial patterns between them may be similar (Jestico, Nelson, & Winters, 2016). For example, all riders in an urban downtown may use similar routes because of limited choices, which helps explain the relationship between crowdsourced and conventional count data. Using bicycle count and survey data, increased ridership is usually associated with increased bicycling infrastructure (e.g., bicycle lanes, separated cycle paths) (Broach et al., 2012). In the case of crowdsourced data, presence of bicycling facilities was not predictive of ridership volumes (Jestico et al., 2016). The poor association between crowdsourced ridership volume and bicycling infrastructure may be related to the manual count locations in known areas of bicycling activity. Another study using Strava data found that bicycling infrastructure was only moderately associated with the density of bicycling activity; it is possible that fitness-oriented cyclists using mobile apps such as Strava may not seek out urban areas where infrastructure is located to support commute activity (Griffin & Jiao, 2015).

Crowdsourced data also show wide potential for examinations of urban areas and urban transportation. For example, ridership volumes collected by Strava have also been used to show how changes in bicycling infrastructure influence bicycling activity over the short-term (Heesch & Langdon, 2017). Though changes in ridership volumes occurred around bicycling infrastructure improvements, conventional data were still needed to adjust volumes across the region because of variations in app use across the area (Heesch & Langdon, 2017). Since all riders in the area do not use the app, their differing preferences for particular routes or types of infrastructure may influence conclusions if volumes are not adjusted based on the larger bicycling patterns from all riders in a region. Further, Strava ridership have been associated with particular neighborhood characteristics like highly connected streets within residential areas, though the characteristics and riding environments of cyclists using the Strava app may differ from those who are not (Sun, 2017; Sun, Du, Wang, & Zhuang, 2017). Health and exposure characteristics have also been explored using these data. Researchers found differences between recreational and commuting riders in terms of where they ride and how much pollution to which they are exposed (Sun & Mobasheri, 2017). Since recreational riders tended toward the outskirts of an urban area, they were potentially exposed to less air pollution than commuters in the same region (Sun & Mobasheri, 2017). These studies help indicate where infrastructure changes could be made to positively influence cyclist health, safety, and overall ridership.

As a step toward developing a method for conflating conventional and crowdsourced bicycling data, we seek to explore the as yet understudied area of understanding how crowdsourced and conventional data correspond in representing activity. Specifically, we first ask how do manual count data compare with Strava crowdsourced data in terms of bicycling activity volume? We conduct this exploration by analyzing the spatial pattern of crowdsourced data as compared to data collected through conventional methods in the Greater Sydney area. Spatial pattern analysis is used in this context as a way to measure how the ridership in crowdsourced and conventional data correspond and show differences among the included datasets. The analysis provides greater precision of correspondence in the comparison as the approach avoids binning ridership into just low, medium, and high areas. While previous studies examined the strength of associations between manual counts and crowdsourced data, they did not examine the spatial associations and ridership patterns between locations for the differing data sources. Specifically, we compare manual bicycle count data with crowdsourced data using local Moran's I_i.

Spatial pattern analysis is a commonly used analytical tool to identify where in a study area there are highly correlated areas of activity. Since highly correlated areas of activity may give some indication about the processes that underlie them (Nelson & Boots, 2008), we then also explore socio-economic demographics and infrastructure in the area to determine their explanatory value related to the patterns of data correspondence and differences we discover. Between discovering areas of high and low correspondence and examining the contests that underline them, we will better understand the analytical potential that