Applied Geography 78 (2017) 78–93

Contents lists available at ScienceDirect

Applied Geography

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

Coverage inequality and quality of volunteered geographic features in Chinese cities: Analyzing the associated local characteristics using geographically weighted regression

Shiliang Su [a, b, c], Chaoran Lei [a], Anyun Li [d], Jianhua Pi [a], Zhongliang Cai [a, e, *]

[a] School of Resource and Environmental Sciences, Wuhan University, Wuhan, China
[b] Beijing Key Laboratory of Urban Spatial Information Engineering, Beijing Institute of Surveying and Mapping, Beijing, China
[c] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China
[d] Network and Information Center, Changjiang Water Resources Commission, Wuhan, China
[e] Key Laboratory of Geographical Information Systems, Ministry of Education, Wuhan University, Wuhan, China

ARTICLE INFO

Article history:
Received 26 October 2015
Received in revised form
31 October 2016
Accepted 28 November 2016
Available online 8 December 2016

Keywords:
Geographically weighted regression
Volunteered geographic information
Demographic indicators
Socioeconomic indicators
Spatial inequality
OSM

ABSTRACT

The volunteered geographic information (VGI) gains increasing popularity with the general public and scientific community. However, the optimism about the VGI has been tempered by two critical issues: inequality in data coverage (social justice) and data quality. It therefore requires a better understanding of the mechanism driving VGI contributions and content quality. With a case of China, this paper demonstrates one potential avenue, examining the associations between VGI coverage/quality and local demographic and socioeconomic characteristics. In particular, VGI data are harvested from the OpenStreetMap for 333 cities in China. VGI coverage is measured by the total volume of different geographic features (point, line and polygon); and VGI quality is described from two aspects: completeness and accuracy. Geographically weighted regression (GWR) shows that both demographic and socioeconomic factors have statistically significant influences on VGI coverage and quality. More specifically, densely populous cities with more young, educated and non-agricultural people enjoy higher VGI coverage and quality. Cities with lower VGI coverage and quality are primarily located in the western and southwestern regions where the ethnic minorities concentrate. High VGI coverage and quality are possibly observed in economically developed cities with high marketization degree. Besides, possibility of high VGI coverage and quality occurs in cities with more labor in scientific research and greater percentage of employers in the tertiary industry. The GWR also demonstrates that the strength and nature of the obtained relationships vary across the 333 cities. The spatial non-stationary relationships may partially answer for the controversial empirical conclusions in earlier case studies at different scales. Quantitative analysis (Gini index, Lorenz curve and Moran's I index) further evidences the great inequality in VGI coverage and quality. It can be safely inferred that the differences in engagement and use of VGI, as a new digital divide, can raise troubling concerns on the social justice implications.

1. Introduction

Due to the advent of collaborative web-based technologies, crowd-sourcing has gradually emerged as a popular approach to the production and acquisition of geographic information (Bimonte, Boucelma, Machabert, & Sellami, 2014; Elwood, 2008; Hudson-Smith, Batty, Crooks, & Milton, 2009; Spinsanti & Ostermann, 2013). As a version of crowd-sourcing, the volunteered geographic information (VGI; Goodchild, 2007) allows the general public to create, disseminate and assemble information (location and property) for geographic features (point, line or polygon) (Longley et al., 2011). The VGI, such as the OpenStreetMap (OSM), gains increasing popularity with the general public, given that it meets the growing demand to access geographic data for daily life (Goodchild & Glennon, 2010; Haklay, 2010; Mullen et al., 2015). The VGI has been evidenced to be a low cost tool to acquire and produce geographic information effectively and timely. Consequently, the

* Corresponding author. No.129 Luoyu Rd, Wuhan, Hubei Province, China.
E-mail addresses: shiliangsu@163.com, shiliangsu@whu.edu.cn (S. Su), zlcai@whu.edu.cn (Z. Cai).

http://dx.doi.org/10.1016/j.apgeog.2016.11.002
0143-6228/

scientific community shows growing interest in the application of VGI tools into various fields (Comber et al., 2013; Foster & Dunham, 2014; Goodchild & Glennon, 2010; Mooney, Corcoran, & Ciepluch, 2013; Roche, Propeck-Zimmermann, & Mericskay, 2011; Schade et al., 2013; Upton, Ryan, O'Donoghue, & Dhubhain, 2015).

It is expected that the volume of VGI available for geographical analysis will grow exponentially given the development of web 2.0 technologies (Elwood, Goodchild, & Sui, 2012). However, the optimism about the VGI has been tempered by two critical issues: inequality in data coverage (social justice; Crutcher & Zook, 2009; Foster & Dunham, 2014) and data quality (completeness and accuracy; Bimonte et al., 2014; Goodchild & Li, 2012; Mullen et al., 2015). These challenges are largely resulted from the mechanism governing the information contribution process (Mullen et al., 2015) and the fact that the volunteers lack of professional training (Mooney, Corcoran, & Winstanley, 2010). It therefore highlights the necessity of expanded research into the associations between the coverage/quality of VGI data and local characteristics (socioeconomic and demographic variations).

Several studies have associated VGI data quality with various demographic indicators, including population density, race, age, gender, class and economic status (Arsanjani & Bakillah, 2015; Crutcher & Zook, 2009; Elwood, 2008; Tulloch, 2008; Zielstra & Zipf, 2010; Zook & Graham, 2007). Haklay (2010) discovered that VGI data for the densely populous urban areas presented higher quality. Girres and Touya (2010) pointed that VGI data with higher quality were observed in areas with more young people and higher income. No significant correlations were found between census tract level demographic characteristics and VGI data quality in Denver (Mullen et al., 2015). Foster and Dunham (2014) conceptualized the coverage inequality of VGI, whose preliminary results revealed that blocks with higher population density, white population and home value had higher level of VGI coverage of urban forest data. Earlier studies greatly contribute to the understanding of the topic under investigation; however, two issues remain to be further solved. For one thing, more quantitative analysis should be conducted, as Mullen et al. (2015) pointed that prior case studies primarily performed the analysis using qualitative measurements. For another, the coverage inequality issue receives insufficient attention and more efforts are needed for this new frontier of VGI applications (Foster & Dunham, 2014).

Geographical data, depending on the record and measurement level, generally presents autocorrelation in space to some extent (Su, Xiao, & Zhang, 2012). One pioneer study reveals that VGI data are spatially autocorrelated in Denver (Mullen et al., 2015). In this case, the efficiency and reliability of analysis would be absolutely reduced, when applying the global regression (ordinary least squares regression, OLS) to analyze the associations between VGI coverage/quality and local characteristics. In particular, significant associations may be identified when it is not the case, because the randomly distributed assumption of OLS is violated (LeSage, 2001). Besides, VGI data present great spatially heterogeneous patterns (Foster & Dunham, 2014; Mullen et al., 2015). The OLS is therefore limited to explain the actual phenomena, since the homoscedasticity assumption is violated and the spatial non-stationarity problem emerges (Wheeler & Páez, 2010). Specifically, the OLS generates the global relationships, which remain constant over space, and the local-specific relationships may possibly be hidden (Fotheringham & Brunsdon, 2002). In this context, some more sophisticated techniques should be employed to capture the spatially heterogeneous relationships.

Geographically weighted regression (GWR), which overcomes the limitations of OLS, has been widely applied to quantify the local-specific relationships (LeSage, 2001; Su, Li, Hu, Xiao, & Zhang, 2014). GWR can produce a series of spatially varying coefficients both in strength and in nature to help deepen the understanding of the phenomena under investigation (Brunsdon, Fotheringhm, & Charlton, 1998). Growing literature has taken the strong advantage of GWR to solve problems in many fields of applied geography (Ansong, Ansong, Ampomah, & Adjabeng, 2015; Cardozo, García-Palomares, & Gutiérrez, 2012; Cockx & Canters, 2015; Hu, Yang, Li, Zhang, & Xu, 2016; Ivajnšič, Kaligarič, & Žiberna, 2014; Lee & Schuett, 2014; Qian & Ukkusuri, 2015; Su et al., 2016, 2012; Xiao et al., 2013). This paper attempts to quantify the spatial non-stationary associations between VGI coverage/quality and local characteristics in China, by using the GWR and data collected from the OSM. The specific objectives are to: (1) characterize the spatial patterns of VGI coverage and quality across China; (2) quantify the inequality of VGI coverage and quality; and (3) examine the associated local demographic and socioeconomic characteristics of VGI coverage and quality.

2. Data and method

2.1. VGI data processing

The OSM is a crowd-sourcing project that volunteers worldwide cooperate to create a detailed and free geospatial database map (Mooney et al., 2013). A variety of approaches are available for contributors to create and share the spatial data with OSM, including records from portable GPS devices, local knowledge sharing, importing free spatial data and interpreting polygons or lines from remotely sensed images (Mooney et al., 2013). Features in OSM can be divided into three categories (point, line and polygon) and represent a great variety of geographic entities, including land cover, lake and river network, transportation routes, buildings, public facilities and park. All the geographic features in OSM can be downloaded freely.

We accessed to the OSM dataset on September 1st, 2015 and downloaded all the data for the entire China. Hongkong, Macao and Taiwai are excluded into analysis, given that statistical data for these regions are missing. The features with no attributes are also excluded. Since the attributes for the geographic features are recorded in a variety of formats, we first reclassify them into different types. Specifically, we propose a two-level classification system: the first level is based on the data structure (point, line and polygon) and the second level is based on the data function. According to the data function, the point features are categorized into three types (commercial sites, profitable public facilities and non-profitable public facilities), the line features are divided into three types (railway, road and waterway), and the polygon features are classified into two types (natural and artificial land cover). Then, we calculate the total volume of the geographic features for the 333 prefecture-level cities and autonomous regions. The prefecture-level city is selected as the basic unit for analysis, because: (1) it is the basic unit for urban management in China; and (2) the socioeconomic and demographic data are accessible, and those for finer level (e.g., county) are unavailable. The VGI volume is described by the total number (point features), density (length/total land area; line features) and coverage rate (area/total land area; polygon features).

It is rather difficult to obtain the high quality reference data for the entire China. Given data availability, we only test the data quality for two types of VGI data: the higher education institution including university and college (one category of the non-profitable public facilities; point) and railway (line). The reference data (1: 50000) is obtained from the National Administration of Surveying, Mapping and Geoinformation. We assess the VGI quality in reference to the methodology demonstrated in Jackson et al. (2013) for point and Haklay (2010) for line. In particular, VGI quality is