Consistency Across Geosocial Media Platforms

Carsten Keßler*, Grant D. McKenzie**

* Department of Planning, Aalborg University Copenhagen, Denmark
kessler@plan.aau.dk
** Department of Geography, McGill University, Montreal, Canada
grant.mckenzie@mcgill.ca

Abstract. The increasing use of geosocial media in research to draw quantitative and qualitative conclusions about urban environments bears questions about the consistency of the data across the different platforms. This paper therefore presents an initial comparative analysis of data from six different geosocial media platforms (Facebook, Twitter, Google, Foursquare, Flickr, and Instagram) for Washington, D.C., using population and zoning data for reference. We find that there is little consistency between the different platforms at small spatial units and even semantically rich datasets have severe limitations when predicting functional zones in a city. The results show that researchers need to carefully evaluate which platform they can use for a particular study, and that more work is needed to better understand the differences between the different platforms.

Keywords. Geosocial Media, Location-Based Social Networks

1. Introduction

The abundance of data from a large number of users, easily accessible through APIs, has led to numerous studies based on geosocial media. Researchers have used geosocial check-ins or geotagged tweets and photos to study online communities (Yin et al., 2016), event detection (Sakaki et al., 2010), urban structure (Hollenstein and Purves, 2010) and its dynamics (McKenzie et al., 2015), gazetteers (Keßler et al., 2009), and functional regions (Gao et al., 2017), to name but a few examples. Population mapping (Patel et al., 2017; Aubrecht et al., 2011) and population mobility (Noulas et al., 2011) have drawn particular interest, following the assumption that users of geosocial media can be used as a representative sample of the overall population in a city. Some of this work has produced interesting, meaningful,

Published in "Adjunct Proceedings of the 15th International Conference on Location Based Services (LBS 2019)", edited by Georg Gartner and Haosheng Huang, LBS 2019, 11–13 November 2019, Vienna, Austria.

This contribution underwent double-blind peer review based on the paper.
https://doi.org/10.34726/lbs2019.57 | © Authors 2019. CC BY 4.0 License.

and broadly cited results. Systematic comparisons across different sources of geosocial media are still scarce in the literature, though, with the few examples focusing on the complementarity of different geosocial data sources (Lee et al., 2004) or differences in the ability to track individual users (Silva et al., 2013; Wang et al., 2013].

The goal of this research is therefore a systematic quantitative and semantic cross-comparison of data from six widely used (geo-)social networks. This paper presents initial results for Washington, D.C., comparing the datasets to each other and to population and zoning in the city as reference data.

2. Data

The data used in this paper consists of 8 different datasets for the area of Washington, D.C., and is summarized in Table 1.

Source	Data points	Acquisition period
Facebook[a]	2,409 places	December 2018
Twitter[a]	118 places	April 2016
Google[a]	6,978 places	April 2016
Foursquare[a]	24,428 venues	September 2017
Flickr[a]	6,945 geotagged photos	December 2018
Instagram[a]	3,130 geotagged photos	April 2016
Population[b]	179 census tracts; 6,507 census blocks	July 2019
DC Zoning[c]	885 zones with 149 classes	July 2019

Table 1. Overview of datasets used, obtained from the respective API (*a*), from the 2010 census (*b*), and from https://opendata.dc.gov/datasets/zoning-regulations-of-2016 (*c*).

3. Consistency analysis

A visual comparison of density across the datasets (see Figure 1) shows little consistency, particularly when compared to the distribution of population across the city. While this may be a result of Washington, D.C. being a major tourist destination – the National Mall and government districts in the center of D.C.'s diamond shape are bare of any population, but show the highest densities for photo-based platforms and Foursquare POIs –, this raises serious concerns about the use of geosocial media to enhance population mapping.

In order to quantify the degree of consistency across these datasets, each of them has been aggregated to the containing census tract and block, respectively. The corresponding numbers for the 179 tracts and 6507 blocks were then tested for correlation; results are summarized in the pair plot shown in

Figure 2. There are still reasonable correlations for the fairly large census tracts, with the maximum values R = 0.85 between number of Foursquare venues and population, and R = 0.81 between number of Foursquare venues and number of Instagram places (see upper right half of Figure 2). When going to the block level, however, the maximum correlation values obtained are much lower, with R = 0.46 between the number of Foursquare venues and the number of Foursquare checkins, and R = 0.4 between number of Foursquare venues and number of Instagram places (see lower left half of Figure 2). Again, these findings do not support the use of geosocial media data for fine-grained population mapping.

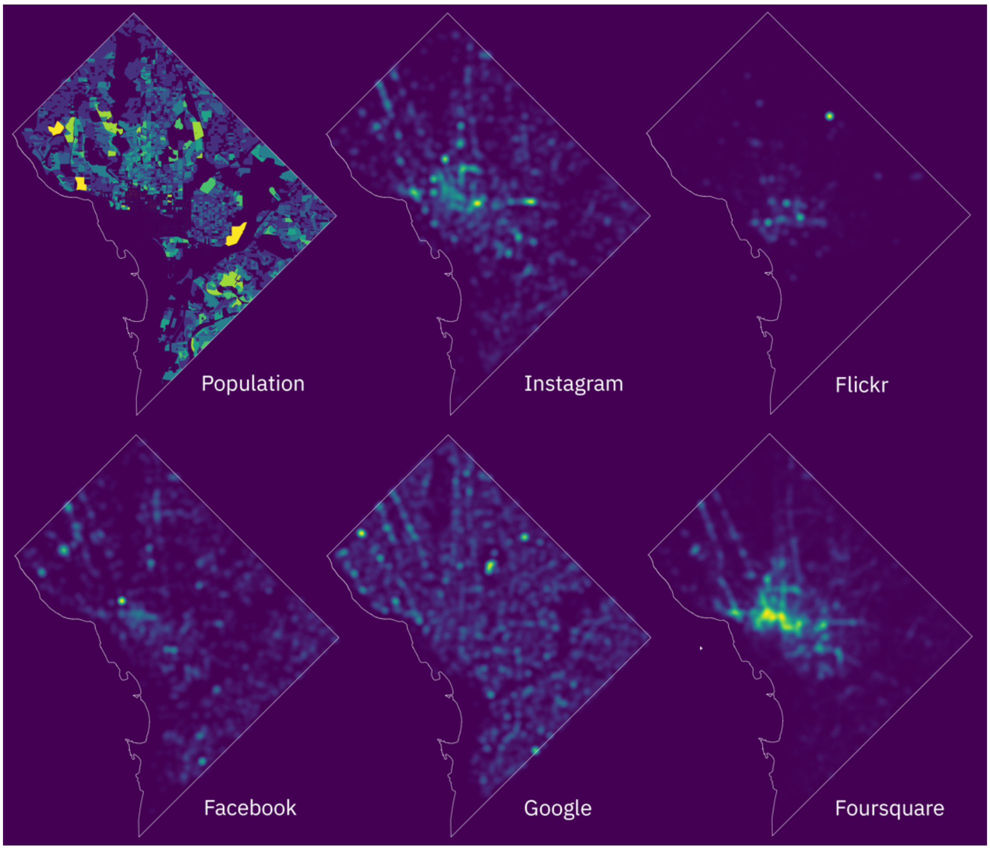

Figure 1. Density maps of data from the different platforms, with population per census block as reference; Twitter places are not shown due to the small number places in the dataset.

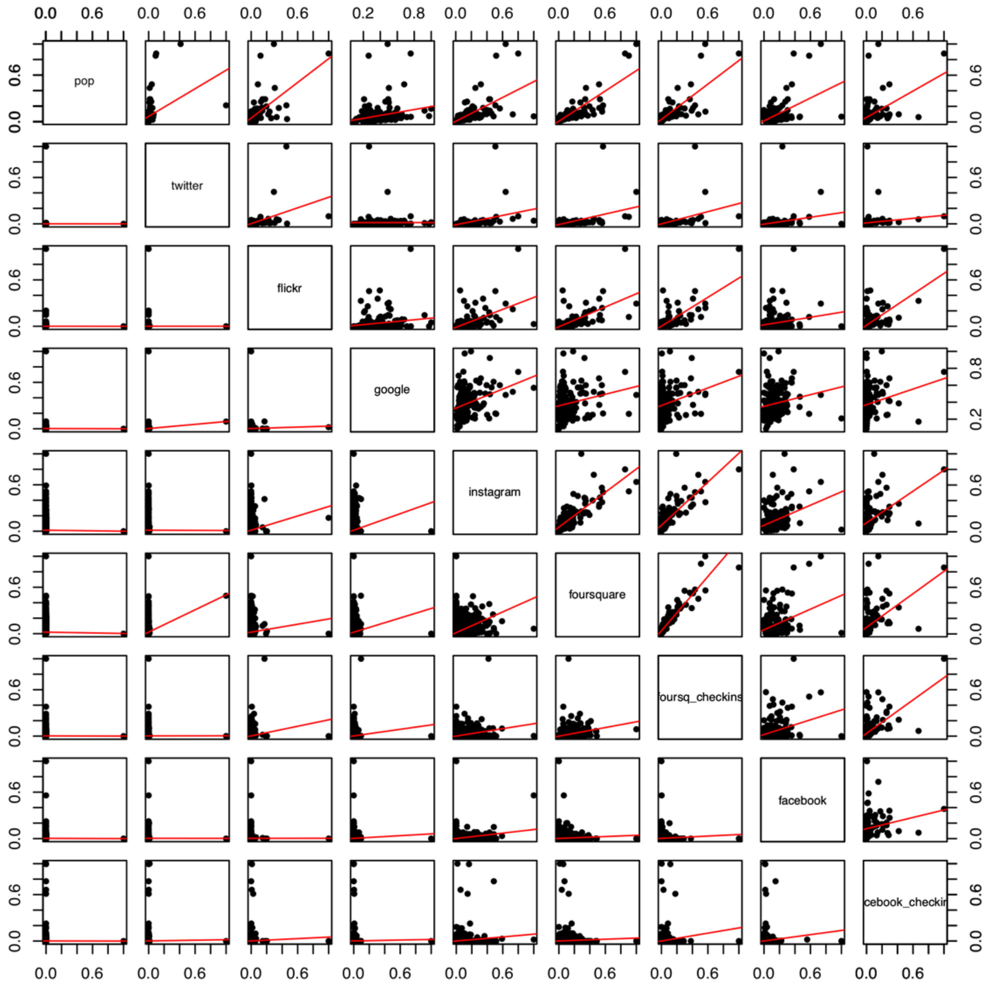

Figure 2. Pair plot showing the relationships between all combinations of data sources per census tract (upper right) and per census block (lower left).

After this quantitative analysis, we wanted to see whether the semantic information in geosocial media can reveal any insights about functional areas of a city. We used the Foursquare dataset for this purpose, as it contains both the largest number of data points (24,428 venues) and is also semantically the richest, with a total of 10 top-level categories (e.g., *Food*, *Outdoors/Recreation*) and 449 second-level categories (e.g., *Latin American Restaurant*, *Athletics/Sports*). The 885 zones defined in Washington, D.C.'s zoning regulations were used as functional areas in this analysis. Each zone belongs to one of 149 classes, grouped into six zoning groups (*Downtown*; *Mixed Use*; *Production, Distribution, and Repair*; *Residential*; and *Special Purpose Zones*). Since the distribution of different kinds of POIs across zones should be indicative of the zones' functions, we aggregated each first-level and second-level category of Foursquare POIs to the 885 zones and used the corresponding vector to train a random forest model to classify each zone into the correct zoning group. When training on the top-level Foursquare categories, the random forest classifier obtains an out-of-bag estimate of error rate of 39.8%; using the much richer second-level categories only lowers the out-of-bag estimate of error rate to 35.2%. This indicates that even an extensive and semantically rich dataset such as the one used here is not sufficient to reflect functional zones in urban environments.

4. Conclusions and Future Work

Our analysis of Washington, D.C. geosocial media data has shown that there is little consistency across the different platforms at small spatial units. At a semantic level, an initial analysis has shown that the use of the data to predict functional zones in the city was limited, even when using rich semantic annotations. Our results indicate that more care needs to be taken when using such data to draw conclusions about urban areas both at a quantitative and at a qualitative level. Researchers need to be much more aware of the kind of platform they choose for their research. Further research is needed to gain a better understanding of the nature of the differences between the datasets. These may be related to the socio-economic groups that use the different platforms, but also to data collection practices (active check-ins and posts vs. passive observation of user presence, as in the case of Google places, for example), as well as the classification systems that produce the semantic information on the platforms. Future work will therefore focus on an assessment of the semantic consistency between the platforms, and a replication in further major cities.

References

Aubrecht, C., J. Ungar, and S. Freire. "Exploring the potential of volunteered geographic information for modeling spatio-temporal characteristics of urban population: a case study for Lisbon Metro using foursquare check-in data." *International Conference Virtual City and Territory (7è: 2011: Lisboa)*, Lisbon, 2011.

Gao, S., K. Janowicz, and H. Couclelis. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, 21(3):446– 467, 2017.

Hollenstein, L. and R. Purves. Exploring place through user-generated content: Using flickr tags to describe city cores. *Journal of Spatial Information Science*, 2010(1):21–48, 2010.

Keßler, C., P. Maué, J. T. Heuer, and T. Bartoschek. Bottom-up gazetteers: Learning from the implicit semantics of geotags. In *International Conference on GeoSpatial Semantics*, pages 83–102. Springer, 2009.

Lee, K., R. K. Ganti, M. Srivatsa, and L. Liu. When twitter meets foursquare: tweet location prediction using foursquare. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 198–207, 2014.

McKenzie, G., K. Janowicz, S. Gao, J.-A. Yang, and Y. Hu. POI Pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 50(2):71–85, 2015.

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011, July). An empirical study of geographic user activity patterns in foursquare. In *Fifth international AAAI conference on weblogs and social media*.

Patel, N. N. , F. R. Stevens, Z. Huang, A. E. Gaughan, I. Elyazar, and A. J. Tatem. Improving large area population mapping using geotweet densities. *Transactions in GIS*, 21(2):317–331, 2017.

Sakaki, T., M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

Silva, T. H., P. O. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. Loureiro. A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. In *Proceedings of the 2nd ACM SIGKDD workshop on urban computing*, page 4. ACM, 2013.

Wang, P., W. He, and J. Zhao. A tale of three social networks: User activity comparisons across facebook, twitter, and foursquare. *IEEE Internet Computing*, 18(2):10–15, 2013.

Yin, H., Z. Hu, X. Zhou, H. Wang, K. Zheng, Q. V. H. Nguyen, and S. Sadiq. Discovering interpretable geo-social communities for user behavior prediction. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 942–953. IEEE, 2016.