Cross-checking user activities in multiple geo-social media networks

Levente Juhász
University of Florida
3205 College Ave
Ft. Lauderdale, FL, USA
levente.juhasz@ufl.edu

Hartwig H. Hochmair
University of Florida
3205 College Ave
Ft. Lauderdale, FL, USA
hhhochmair@ufl.edu

Abstract

Geo-social media users tend to use different services simultaneously. Whereas significant research attention was put into analyzing contribution patterns to individual social-media platforms, it is less known how the same individual user uses different services. This pilot study analyzes the spatial behavior of users in multiple geo-social media services using geotagged Instagram media objects and Foursquare check-ins for 10 individual users. It assesses the usability of several methods a) to extract and map user activity spaces and b) to quantify the similarity of social media contributions of a user in different platforms. The analysis of user behaviour across multiple platforms can help to refine travel behaviour models and also to assess the dynamics of user activity levels in different platforms over time.

Keywords: geo-social media; activity space; Instagram; Foursquare

1 Introduction

Most geo-social media platforms are location based services that map and geocode user activities. Such geocoded activities provide the basis for the spatial analysis of activity patterns of these users. This study will analyze the locations of contributions of 10 users to two prominent social media platforms. For each user it will characterize the activity space obtained from each platform, and also compare the similarity of contributions between both platforms. The two platforms used are Instagram (IG), a photo and video sharing service with 500 million daily users[1], and Foursquare with 50 million active monthly users[2]. Foursquare provides two apps, namely Foursquare City Guide, which is used to review and rate businesses (e.g. restaurants), and Foursquare Swarm, which is a check-in tracker that allows users to log visited places. Geolocation in Instagram is done by attaching a predefined location to a media object (Cvetojevic et al., 2016). Swarm, the check-in tracker of Foursquare uses a similar approach and lets users select a place from nearby venues. These predefined locations are user-generated, therefore often contain errors (Hochmair et al., 2018). IG users sometimes associate their photos with generic locations (i.e. a city or region) instead of choosing the true location of the image for increased privacy (Cvetojevic et al., 2016), leading to position inaccuracies. There is evidence in the literature that individuals do contribute geo-data to multiple volunteered geographic information platforms, such as OpenStreetMap and Mapillary (Juhász and Hochmair, 2016a).

Human activity space is defined as the area within which the majority of an individual’s day-to-day activities are carried out (Johnston et al., 2000). Traditionally, studies approximate this area with ellipse-based representations (Yuan and Raubal, 2016), however, such ellipses cannot capture the complexity of shapes associated with human activities. Wildlife ecology developed the concepts of home-ranges and utilization distributions (UD). A home-range of an animal is the area in which the animal conducts 95% of its activities (Worton, 1987). UD is the probability distribution defining an animal’s use of space (Van Winkle, 1975). Core areas are often defined by the 50% probability contour. We adapt these concepts to social media use. The first objective of this paper is therefore to adapt several methods from wildlife ecology to extract home and core areas for IG and Swarm users.

The second objective is to apply and evaluate several methods of spatial pattern comparison (SPC) to mathematically quantify the (dis)similarity between social media footprints in different platforms. A review of SPC methods and associated issues are given by Long and Robertson (2017). One of the issues associated with SPC is the modifiable areal unit problem (MAUP), which means that different spatial configurations (e.g. grid size) affect the results of statistical analysis (De Smith et al., 2015). Therefore, both grid-based and scale independent methods are presented here.

[1] http://blog.instagram.com/post/165759350412/170926-news

[2] https://foursquare.com/about

2 Materials and methods

2.1 Dataset description

Locations of IG media (photos, videos) and Swarm check-ins from 10 individuals were used to test different methods of activity space extraction and comparison. The 10 users were selected based on the criteria of using both IG and Swarm simultaneously. For privacy reasons, user-sensitive data (e.g. location history) from IG and Foursquare are not accessible to the public, therefore users need to explicitly authorize applications to access their data. Guidelines for developing such applications, including the authorization process, are provided in the literature (Juhász et al., 2016). The analysis was limited to a city for each user where he or she had previously lived at some point. Table 1 lists the number of data points from users for both platforms that were used in the study.

Table 1: Summary of the dataset

User ID	City	Instagram (geotagged)	Swarm
1	Fort Lauderdale, FL	82	1,360
2	Tampa Bay area, FL	342	230
4	Szeged, Hungary	21	589
6	Budapest, Hungary	14	56
7	Salzburg, Austria	14	193
8	Budapest, Hungary	39	1,583
9	Szeged, Hungary	21	1,743
10	Budapest, Hungary	20	6,620
11	Szeged, Hungary	9	2,620
12	Miami, FL	16	322
	Total	*578*	*15,136*

2.2 Methods for activity space extraction

The minimum convex polygon (MCP) represents the minimum area containing all observations and is a widely used home-range estimation tool (Mohr, 1947). To estimate the home-range, a certain number of points furthest from the centroid can be excluded for the generation of the MCP. For example, the area retained after excluding 50% of the furthest points can be considered the core area. While simple, MCPs by definition can only produce convex shapes, which sometimes does not correspond to a real world scenario. Characteristic hull (CHull) methods based on Delaunay triangulation overcome this limitation (Downs and Horner, 2009). An advantage of CHull based methods is that they can handle disjoint areas and do not require any input parameters. Local convex hulls (LoCoH) utilize a similar concept as MCPs, and build convex hulls from observations and their neighbors (Getz et al., 2007). Different variations exist depending on neighbour selection criteria, such as fixed-r LoCoH or adaptive. The adaptive LoCoH selects a variable number of neighbors so that the sum of distances is less than a given threshold. Hulls can be then merged together from smallest to largest to extract home-ranges. LoCoH tools provide natural looking results but are sensitive to input parameter selection.

Kernel density estimators (KDE) are also used to extract home-ranges by generating a probabilistic surface. This allows to determine the estimated proportion of observed events within a selected area. Their drawback is that estimations are affected by bandwidth selection and that they are not robust with complex shapes (Downs and Horner, 2009).

This paper illustrates the adaption of home and core ranges from wildlife ecology to the geo-social media domain.

2.3 Overlap and similiarity metrics

Two metrics from Fieberg and Kochanny (2005) are applied to the extracted activity areas explained in Section 2.2. The simplest method calculates the percent overlap between activity areas from two sources as

$$O_{A,B} = {}^{A_{A,B}}\!\big/_{A_A} \tag{1}$$

where $O_{A,B}$ is the overlap index that shows the proportion of the activity area in platform A (A_A) that overlaps with the activity area in platform B, and $A_{A,B}$ is the area of overlap between platforms A and B activity areas. The overlap index ranges from 0 to 1. 0 means no overlap, whereas 1 means that the activity area of platform A is entirely within that of platform B. Another overlap metric is the UD overlap index (UDOI), which is a function of the product of two UDs. UD in this context is the probability distribution defining a user's use of space in IG or Swarm. Practically, UD is a KDE output surface. UDOI is calculated as

$$UDOI = A_{A,B}\left(\iint \widehat{UD}_A(x,y)\ \times\ \widehat{UD}_B(x,y)\ dxdy\right) \tag{2}$$

where $A_{A,B}$ is the overlap area between platform A and B. UDs, $\widehat{UD}_A$ and $\widehat{UD}_B$ are the estimated UDs for platforms A and B, i.e. Swarm and IG. UDOI equals 0, if there is no overlap between home-ranges, and it is 1 in case of a 100% overlap (assuming that the two UDs are equally distributed). The drawback of these two overlap indices is that they depend on the extraction of activity spaces. Therefore, we present four other approaches to quantify the similarity between point sets that are independent of extracted activity spaces.

One approach is the radius of gyration (RG) which measures the spread of point locations around the mass center and can therefore be applied to individual users (Juhász and Hochmair, 2016b). A radius of gyration index (RGI) between two platforms A and B can be calculated as

$$RGI_{A,B} = \frac{RG_A - RG_B}{RG_A + RG_B} \tag{3}$$

where RG_A and RG_B are the radius of gyration values for platforms A and B, respectively. This index ranges between -1 and 1, where a positive value means that locations in platform A are more spread than in platform B, a negative value means the opposite, and zero means identical spread. The drawback is that the RGI does not provide information about the co-location of two point sets.

The Jaccard-index (J) is a normalized similarity measure that measures the co-occurrence of attributes in different object classes (Hochmair, 2005). In the context of this study, the analyzed geographic space can be subdivided into regular grid cells, and J can be calculated as

$$J = \frac{M_{AB}}{M_A + M_B + M_{AB}} \tag{4}$$

where M_A is the number of grid cells with platform A events only, M_B is the number of cells with platform B events only, and M_{AB} is the number of cells with both types of events. J ranges from 0 (no overlap) to 1 (platform A and B events occur in the same cells).

Adapted from Lenormand et al. (2014), another grid-based approach (GC – grid correlation) can be used. It aggregates the number of IG media objects and Swarm check-ins by grid cells and then normalizes these grid values by dividing them by the total number of media or check-ins, respectively. The Pearson-correlation coefficient between these two grid-based variables measures the spatial similarity of IG and Swarm usage.

In the computer vision domain Coen et al. (2011) proposed a similarity distance (d_s) between two point sets that uses the Kantorovich-Wasserstein metric (d_{KW}). The d_{KW} metric provides an optimal solution to the transportation problem which can be formulated as: “What is the optimal way to ship goods from suppliers to receivers?” and denotes the maximally cooperative way (i.e., involving communication to minimize global cost) to transport masses between sources and sinks. d_s is defined as

$$d_s(A,B) = \frac{d_{KW}(A,B)}{d_{NT}(A,B)}) \tag{5}$$

where d_{NT} is the naïve solution to the same problem, simply summing all ground distances between the point sets. d_s measures how much is gained by optimization of the transport problem. d_s equals 0 if the point sets are identical (i.e. receivers in the original problem are co-located with suppliers, therefore the optimal distance is 0). It equals 1 if the optimization does not result in gain (i.e. point sets are so different that $d_{KW} = d_{NT}$).

3 Results

3.1 Activity spaces

Home and core areas were computed for three vector-based methods (MCP, CHull, LoCoH – adaptive with half the maximum distance) and for a KDE based method (using a bivariate normal kernel) as described in Section 2.2. Estimation of IG core areas was not successful for users 6, 10 and 11 due to the low point number and the distribution of those points. The CHull method produces artificial patterns in most real world scenarios as seen in Figure 1. Thin triangles (line-like features on Figure 1) appeared in the extracted activity areas that are most prominent along roads. This is a common scenario, since businesses are typically located along the road network, and therefore, social media users tend to use the space accordingly. Hence, the CHull is not an adequate method to estimate activity spaces of social-media users.

Figure 1: Activity space estimation with the CHull method

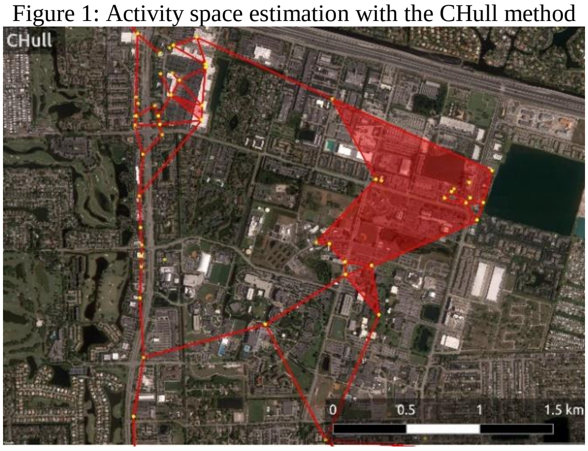
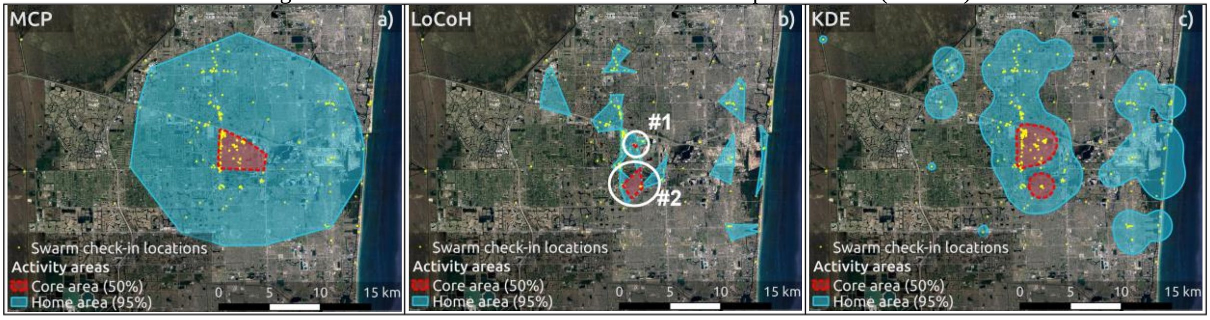


Figure 2 illustrates the results of home and core area estimation for the remaining methods. The major drawback of MCP (Figure 2a) is that it always results in convex shapes. In addition, excluding points furthest from the centroid is not adequate if the activity is not uniformly distributed (e.g. when major activity happens around two distinct locations). Both LoCoH (Figure 2b) and KDE (Figure 2c) overcome these limitations and allow concave and disjoint geometries. However, both methods depend on input parameters, such as a radius in case of LoCoH, and bandwidth and grid size in case of KDE.

Figure 2: Estimated home and core areas for Foursquare/Swarm (User #1)

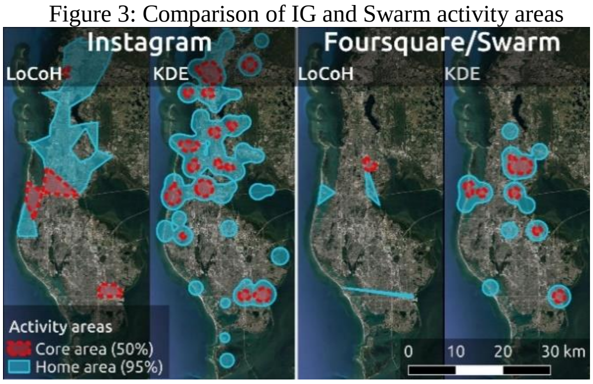


Visual inspection of results suggests that MCP and KDE overestimate both home and core areas. As opposed to this, LoCoH performed well in the core area estimation for user 1 in Figure 2b, by producing two disjoint areas, i.e. around the workplace (#1) and the usual lunch spot (#2), where most daily activities happen.

3.2 Similarity of activities

To apply and illustrate overlap metrics that depend on activity space extraction, overlap indices (O) were calculated for both home and core areas between IG and Swarm for all users, based on areas extracted with LoCoH and KDE. UDOI was calculated based on the results of KDE. Results are listed in Table 2. For clarity, an interpretation of user 2 is given as an example. Figure 3 shows the extracted Swarm and IG activity areas for this user. Moving from left to right in Table 2, an O_{si} value of 0.784 means that 78.4% of the Swarm home area is overlapped by IG. However, O_{is} shows that only 6.1% of the IG home area is overlapped by Swarm activity, suggesting that IG covers a much larger area among the two platforms. Table 2 also shows that for user 2, core areas extracted with the LoCoH method do not overlap, meaning that the IG and Swarm activities of this user are focused on different areas. Home areas extracted with a kernel based method show a similar pattern, however, with less spatial separation, which might be explained by the overestimation of KDE areas. This resulted also in an overlap between IG and Swarm core areas. The low UDOI value for core areas confirms that the user uses the space differently in these two platforms.

Figure 3: Comparison of IG and Swarm activity areas



To compare user activity directly without the generation of home or core range estimates, we test several approaches. Table 3 lists J, GC, RGI and d_s similarity statistics calculated for the 10 users. J and GC are grid based methods affected by MAUP. To elaborate on this effect, we calculate J and GC for 1km and 2km grids. J measures the spatial co-occurrence of IG and Swarm activities regardless of their intensity. To account for intensity, GC can be used. A higher correlation for users indicates that those users post IG photos primarily at those areas where they also check-in. Values in bold indicate statistical significance at a 1% significance level. The RGI quantifies spread. Values close to 0 indicate that the user uses IG and Swarm within equal range of a center location. A positive RGI in this table means that the user's Swarm check-ins are more spread out than IG posts, a negative RGI means the opposite. A higher d_s value means that the point sets of IG posts and Swarm check-ins differ whereas a d_s value closer to 0 indicates that the point sets are closer to identity.

Table 3: Global similarity metrics

User ID	Jaccard-index (J)		Grid-correlation (GC)		RGI(s,i)	d_s
	1km	2km	1km	2km		
1	0.19	0.30	**0.41**	**0.62**	-0.25	0.49
2	0.14	0.23	0.01	**0.28**	-0.20	0.56
4	0.15	0.22	**0.58**	**0.75**	0.07	0.37
6	0.22	0.30	**0.63**	**0.84**	0.22	0.45
7	0.33	0.33	**0.78**	**0.88**	0.06	0.32
8	0.13	0.19	**0.44**	**0.66**	-0.05	0.31
9	0.12	0.15	**0.91**	**0.97**	-0.27	0.49
10	0.05	0.05	0.05	0.11	0.11	0.71
11	0.06	0.13	**0.91**	**0.96**	-0.10	0.47
12	0.07	0.10	0.20	**0.49**	0.34	0.68

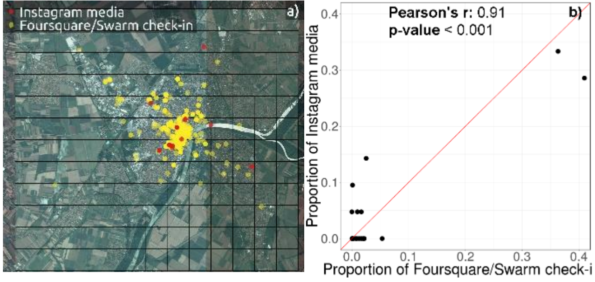
As Table 3 shows, increasing grid size results in higher J indices and stronger correlations (GC) between IG and Swarm activity. The similarity approaches can be illustrated for a sample user (user 9). Figure 4a shows the IG post and Swarm check-in locations on top of a 1km grid. The relatively low Jaccard-index values (0.12; 0.15) indicate that most IG posts and Swarm check-ins do not co-occur in space. However, under consideration of intensity, the Pearson-correlation coefficient (0,91; 0.97) yields strong agreement between IG and Swarm (Figure 4b). This is because areas with the highest number of check-in locations correspond well to the majority of IG photos, i.e., in the city center. The negative RGI value for this user indicates that check-in activity is spatially more

Table 2: Overlap indices for home and core areas calculated based on LoCoH and KDE, along with UDOI

User ID	Home-area (LoCoH)		Core-area (LoCoH)		Home-area (KDE)			Core-area (KDE)		
	O_{si}	O_{is}	O_{si}	O_{is}	O_{si}	O_{is}	UDOI	O_{si}	O_{is}	UDOI
1	0.571	0.212	0.01	0.002	0.792	0.641	0.887	0.883	0.308	0.081
2	0.784	0.061	0.000	0.000	0.739	0.245	0.821	0.356	0.143	0.135
4	0.109	0.632	0.000	0.000	0.723	0.998	1.249	0.588	0.767	0.146
6	0.239	1.000	-	-	0.449	0.997	1.337	0.752	0.991	0.256
7	0.121	0.779	0.000	0.000	0.783	0.918	1.204	0.870	0.849	0.249
8	0.545	0.779	0.200	0.061	0.601	0.845	1.153	0.511	0.502	0.170
9	0.166	0.884	0.232	0.098	0.911	0.567	0.670	1.000	0.756	0.100
10	0.05	0.866	-	-	0.427	0.862	0.348	0.000	0.000	0.004
11	0.638	0.832	-	-	0.775	0.779	0.869	0.888	0.867	0.128
12	0.076	0.659	0.000	0.000	0.312	0.797	0.880	0.477	0.465	0.193

concentrated (in the city center), which can also be confirmed visually.

Figure 4: Swarm and Instagram activity for user 9 (a) and Pearson correlation (b)



4 Summary

This study applied the concept of home-ranges and utilization distributions from wildlife ecology to Instagram and Foursquare/Swarm users to extract home and core areas. Results show that the choice of the range extraction method has a strong effect on mapped home and core regions, and that KDE methods tend to overestimate the spatial extent of events. The paper also presented methods to quantify the similarity between spatial patterns of a user's geo-social media activities. Future work will extend the analysis to additional social media platforms and a larger user base. We will also explore possibilities to obtain more reliable (i.e. true) activity areas from alternative sources, e.g. from phone data, to be able to evaluate activity areas obtained from social media platforms.

References

Coen, M. H., Ansari, M. H. and Fillmore, N. (2011) Learning from Spatial Overlap. In: *Proceedeings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, San Fransisco, CA. pp. 177-182, 2011.

Cvetojevic, S., Juhasz, L. and Hochmair, H. H. (2016) Positional Accuracy of Twitter and Instagram Images in Urban Environments. *GI_Forum 2016*, 1**,** 191-203.

De Smith, M. J., Goodchild, M. F. and Longley, P. A. (2015). *Geospatial Analysis (5th ed.).* Leicester: Matador.

Downs, J. A. and Horner, M. W. (2009) A Characteristic-Hull Based Method for Home Range Estimation. *Transactions in GIS*, 13(5- 6)**,** 527-537.

Fieberg, J. and Kochanny, C. O. (2005) Quantifying Home-Range Overlap: The Importance of the Utilization Distribution. *Journal of Wildlife Management*, 69(4)**,** 1346-1359.

Getz, W. M., Fortmann-Roe, S., Cross, P. C., Lyons, A. J., Ryan, S. J. and Wilmers, C. C. (2007) LoCoH: nonparameteric kernel methods for constructing home ranges and utilization distributions. *PLOS ONE*, 2(2)**,** e207.

Hochmair, H. H. (2005) Towards a Classification of Route Selection Criteria for Route Planning Tools. In*:* Fisher, P. F. (ed.) *Developments in Spatial Data Handling.* Berlin, Springer, pp. 481-492.

Hochmair, H. H., Juhász, L. and Cvetojevic, S. (2018) Data Quality of Points of Interest in Selected Mapping and Social Media Platforms. In*:* Kiefer, P., Huang, H., Van De Weghe, N. and Raubal, M. (eds.) *Progress in Location Based Services 2018. (Lecture Notes in Geoinformation and Cartography).* Springer, pp. 293-313.

Johnston, R. J., Gregory, D., Pratt, G. and Watts, M. (2000). *The Dictionary of Human Geography.* Oxford: Wiley.

Juhász, L. and Hochmair, H. H. (2016a) Cross-Linkage between Mapillary Street Level Photos and OSM Edits. In*:* Sarjakoski, T., Santos, M. Y. and Sarjakoski, L. T. (eds.) *Geospatial Data in a Changing World: AGILE 2016 (Lecture Notes in Geoinformation and Cartography).* Berlin, Springer, pp. 141-156.

Juhász, L. and Hochmair, H. H. (2016b) User Contribution Patterns and Completeness Evaluation of Mapillary, a Crowdsourced Street Level Photo Service. *Transactions in GIS*, 20(6)**,** 925-947.

Juhász, L., Rousell, A. and Jokar Arsanjani, J. (2016) Technical Guidelines to Extract and Analyze VGI from Different Platforms. *Data*, 1(3)**,** 15.

Lenormand, M., Picornell, M., Cantú-Ros, O. G., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frías-Martinez, E. and Ramasco, J. J. (2014) Cross-checking Different Sources of Mobility Information. *PLOS ONE*, 9(8), e105184.

Long, J. and Robertson, C. (2017) Comparing spatial patterns. *Geography Compass*, 12(e12356).

Mohr, C. O. (1947) Table of equivalent populations of North American small mammals. *The American Midland Naturalist*, 37(1)**,** 223-249.

Van Winkle, W. (1975) Comparison of Several Probabilistic Home-Range Models. *The Journal of wildlife management***,** 118-123.

Worton, B. J. (1987) A review of models of home range for animal movement. *Ecological Modelling*, 38(3-4)**,** 277-298.

Yuan, Y. and Raubal, M. (2016) Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. *International Journal of Geographical Information Science*, 30(8)**,** 1594-1621.