Marshallplan-Jubiläumsstiftung
Austrian Marshall Plan Foundation
Fostering Transatlantic Excellence

Identification of activity patterns for individual users across multiple VGI and social media platforms

Final research report by Levente Juhász

University of Florida | Carinthia University of Applied Sciences

Fort Lauderdale, FL, USA

2018

UF UNIVERSITY of FLORIDA

FACHHOCHSCHULE KÄRNTEN

Table of Contents

Acknowledgements

Several factors contributed to the successful execution of this research project. I would like to give special credit to two individuals in particular. First, I thank my advisor at the University of Florida, Hartwig Hochmair for guiding me through the PhD process, which eventually led to this project. I owe another big gratitude to Gernot Paulus from the Carinthia University of Applied Sciences for agreeing to have me at CUAS. Thank you, Gernot for making me so welcome and providing such a nice atmosphere while conducting my research in Austria.

I am also grateful for the financial support of the Austrian Marshall Plan Foundation that allowed me to expand my horizons and reconnect with the European research community. Without the Marshall Plan Scholarship Program, this work would have not been possible.

Abstract

In recent years, online services and applications became part of our daily routine from checking weather forecasts to sharing photos on social media. These activities generate massive amounts of data on the Internet that allows researchers to answer a variety of research questions. To name a few, researchers used Volunteered Geographic Information to better understand human mobility through geotagged social media messages, or to provide ground information for first responders during natural or man-made crises. However, most studies were conducted using data from one single platform. People, on the other hand, tend to use multiple services simultaneously (e.g. Tweeting, but still posting photos on Instagram). A combination of these data sources for individuals therefore may give a more accurate picture of a user's online behavior. This phenomenon is almost completely missing from the GIScience literature.

This research project was dedicated to the idea of analyzing social media and Volunteered Geographic Information data from multiple sources on the individual level, and eventually to extend the literature by providing a first description of cross-platform user activity. The project reviewed and tested various methods that can be used to accurately extract activity spaces from social media data, as well as to mathematically compare similarity between activities. Since understanding user behavior in online services is a data intensive problem, a web application was also developed that can be used to collect geocoded activities from individuals in ten different online sources. The data collector application was deployed and tested in a real world scenario, which resulted in a database that contains geocoded user activities for 53 individuals.

1. Introduction

In recent years, online services and applications have become part of our everyday lives. For example, we use navigation apps to plan our routes to places we are unfamiliar with, photo sharing services to share vacation memories, and social media services to keep connected with people we know. During these activities, people inevitably generate massive amounts of data that can be analyzed and used to answer a number of questions. This is the so-called user generated content (UGC) that has been extensively studied in recent years. A large portion of this user generated data contains a geographic component, and is often referred to as Volunteered Geographic Information (VGI). However, VGI terminology is not standardized, and the term is often used as an umbrella for user generated geographic data. For example, VGI may be created explicitly for the purpose of producing geographic data (e.g. collectively edited maps), or be generated involuntarily, that is, unintentionally, by online users (e.g. geocoded social media posts) (See et al. 2016).

1.1. Problem statement and motivation

Understanding contribution patterns is a major challenge, and it is important in the context of spatial data quality of VGI (Budhathoki and Haythornthwaite 2013). Therefore, this research deals with various aspects of cross-platform user behavior and aims to advance our understanding on how individual users use different social media and VGI platform simultaneously.

Contribution behavior for individual crowd sourcing applications has already been extensively analyzed in the literature. However, it is less understood if and how users participate in several crowd sourcing activities. A certain percentage of Volunteered Geographic Information (VGI) (Goodchild 2007) users are typically more motivated in contributing to new, previously unmapped areas, rather than refining and updating data in already mapped areas, leading to a stagnation in data quality and lack of updates in mapped areas. Also, oftentimes a lack of commitment to long-term contributions can be observed (Neis and Zielstra 2014). It is also a common practice for people to use multiple social media applications simultaneously during their everyday activities. This include for example using navigational services, writing restaurant reviews or posting social media photos.

This research project extends the literature by analyzing activities of individual users in multiple VGI and social media platforms and provides a first description using real data as to how this phenomenon is taking place.

1.2. Document structure

This report is divided into three main sections (Section 2-4), each dealing with a certain aspect of cross-platform user behavior analysis.

The spatial dimension of user activities can be described by so called activity spaces. Even though no formal, quantifiable definition exists, these activity spaces represent the area within which the majority of social activities are carried out. In Section 2, a case study is presented that reviews and tests various methods to extract individual activity spaces from different social media applications. While doing so, this section applies concepts from wildlife ecology and applies them on the social media domain to provide more detailed representations of activity spaces than traditional methods. The case study also presents methods to quantify the spatial similarity of these activities. The latter is important to understand how users interact with the space while using different social media services

Data collection is crucial for the success of this research since any analysis aiming to analyze user contributions across different services needs to reliably identify the same individuals across multiple data sources. However, this is not a straightforward task. An earlier study that cross-checked mapping activities of the same individuals between OpenStreetMap and Mapillary relied on the matching of usernames between these two platforms (Juhász and Hochmair 2016a). A limitation of this approach is that it is not able to identify users who use different usernames in different services and therefore potentially excludes a lot of users from the analysis. To overcome this limitation, a different approach is used in this project. Section 3 presents the development of a web application that collects data from users with their consent. Since users are asked to contribute their user profiles, the approach eliminates the uncertainty originating from user name matching and other approaches and provides reliable data for further analysis.

Section 4 presents a summary of the collected data in terms of user numbers, data volume and geographic coverage. Finally, Section 5 summaries the research project and provides directions for future work.

2. Cross-checking individual activity spaces in multiple geo-social media platforms

Most geo-social media platforms are location based services that map and geocode user activities. Such geocoded activities provide the basis for the analysis of special activity pattern of these users. This study will analyze the co-location of contributions of 10 users to two prominent social media platforms, characterize their activity space, and compare the similarity of contributions to 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, Juhasz, and Hochmair 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, Juhász, and Cvetojevic 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, Juhasz, and Hochmair 2016), leading to position inaccuracies. There is evidence in the literature of the same individual contributing geo-data to multiple volunteered geographic information platforms, such as OpenStreetMap and Mapillary simultaneously (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-range 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 (dis)similarity between social media footprints in different platforms. A review of SPC methods and associated issues are given in the literature (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, Goodchild, and Longley 2015). Therefore, both grid-based and scale independent methods are presented here

2.1. Materials and methods

2.1.1. 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

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

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

authorize applications to access their data. Guidelines for developing such applications, including the authorization process, are provided in the literature (Juhász, Rousell, and Jokar Arsanjani 2016). Analysis was limited to a city for each user where they have previously lived at some point. Table 1 lists the number of data points from users for both platforms 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*

2.1.2. 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 home-range MCPs exclude points furthest from the centroid. 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.

2.1.3. Overlap and similarity 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:

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

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 entirely contains the overlap between the two. 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 an IG or Swarm. Practically, UD is a KDE output surface. UDOI is calculated as

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

where $A_{A,B}$ is the overlap area of overlap 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 now 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 that quantify the similarity between point sets that is 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

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

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

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

where M_A is the number of grid cells with only platform A events, M_B is the number of cells with only platform B events, 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 normalizes the values by dividing events in a cell by the total number of media or check-ins respectively. The Pearson-correlation coefficient between these two variables measures the spatial similarity of IG and Swarm usage.

In the computer vision domain (Coen, Ansari, and Fillmore 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 good 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

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

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 is does not result in gain (i.e. point sets are so different that $d_{KW} = d_{NT}$).

2.2. Results

2.2.1. 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. Estimating IG core areas were not successful for users 6, 10 and 11 due to the low 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 often correspond to the road network, and therefore, social media users tend to use the space accordingly. Therefore, the CHull is not an adequate method to estimate social-media user activity spaces.

Figure 1: Activity space estimation with the CHull method

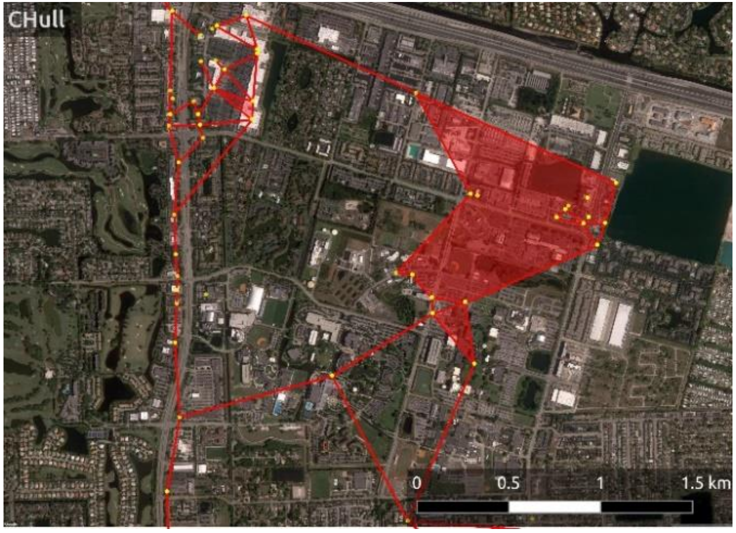
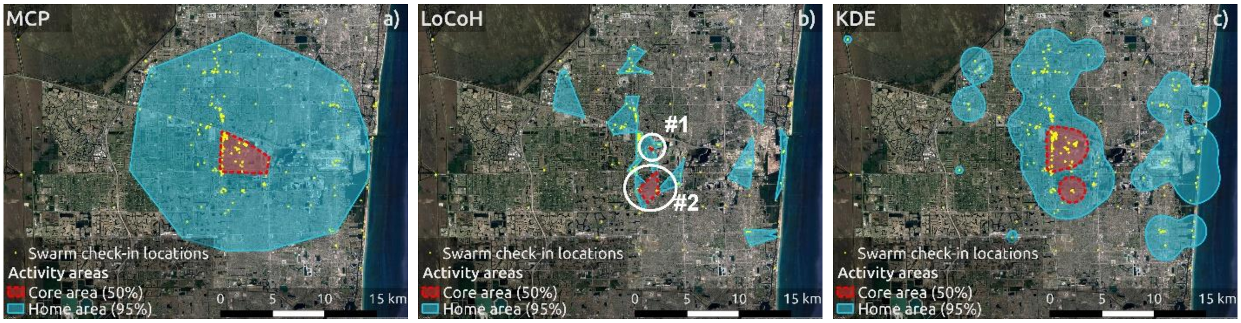


Figure 2 illustrates the results of home and core area estimation for the remaining methods. The major drawback of MCP (Figure 2a) is that it always results in convex shapes. In addition, excluding points furthest from the centroid is not adequate if activity is not uniformly distributed (e.g. when major 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.

Visual inspection of results suggest 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.

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



2.2.2. 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 example interpretation of user 2 is given. Figure 3 shows the extracted Swarm and IG activity areas. 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 in an overlap between IG and Swarm. The low UDOI value for core areas confirms that the user uses the space differently in these two platforms.

To compare user activity directly without the need to construct 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 with 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 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

Figure 3: Comparison of IG and Swarm activity areas

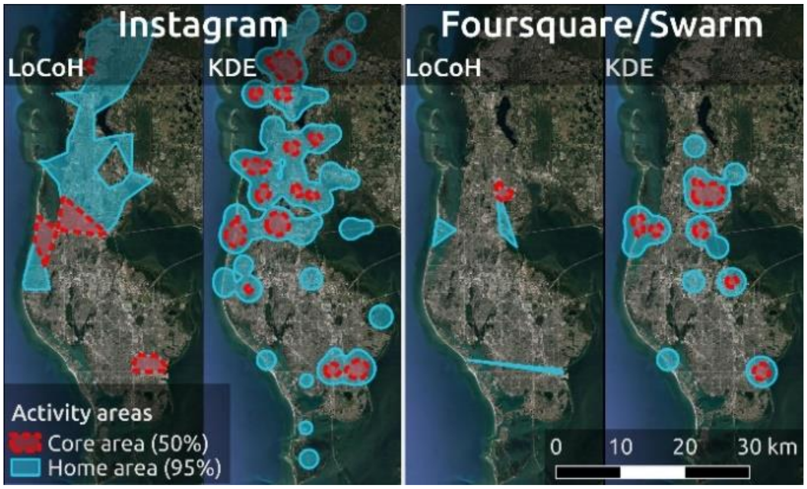


Table 3: Global similarity metrics

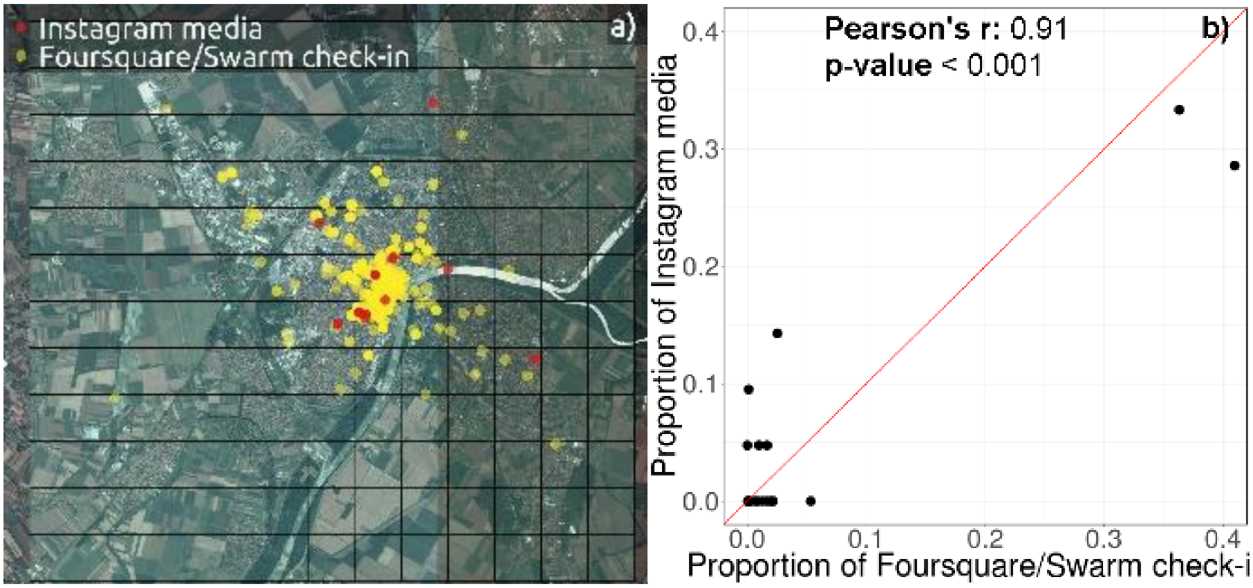
User ID	Jaccard-index (J)		Grid-correlation (GC)		RGI(s,i)	d_s
	1km	2km	1km	2km		
1	0.19	0.30	**0.41**	**0.62**	-0.25	0.49
2	0.14	0.23	0.01	**0.28**	-0.20	0.56
4	0.15	0.22	**0.58**	**0.75**	0.07	0.37
6	0.22	0.30	**0.63**	**0.84**	0.22	0.45
7	0.33	0.33	**0.78**	**0.88**	0.06	0.32
8	0.13	0.19	**0.44**	**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 value indicates that most IG posts and Swarm check-ins do not co-occur in space. However, under consideration of intensity, the Pearson-correlation coefficient yields a high level of 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 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)

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



2.3. Summary

This section 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 also include space-time geography information to automatically detect the contributions of an individual user to several platforms.

3. Developing a data collector tool

Further analysis of cross-platform user activity requires a reference database containing activities of the same individual in multiple services. As mentioned before, identifying the same user across different platforms is not a straightforward task since different services are usually not connected to each other. Users can also opt for using different usernames throughout their online activities, which makes it harder to decide whether two users from different online services are the same person or not. User data, especially data that contains high resolution location and temporal information (i.e. the whereabouts of a person) is also considered to be sensitive information, and is protected by the host service in some cases. Practically, it means that unauthorized data collection is not allowed from some social media sites. An example is accessing the location history of a user on the Foursquare/Swarm platform, which is not allowed except for the user, and his or her "friends".

Although this technical limitation prevents us to mine user contributions across multiple platforms on a larger scale, a workaround to build at least a small reference database was to develop a web application that asks users to voluntarily share their online profiles and contributions. Users were invited to visit the site at https://research.jlevente.com where the purpose of the study was explained in an introductory screen. Next the users had the option to connect (i.e. log-in) with some of their online accounts to the site, which made their online activities in these platforms visible to us. Participation was entirely voluntary and users did not receive any monetary award from participating. The application uses the standard OAuth flow implemented by the third party services. Practically, it means that users can explicitly authorize the data collector application to extract their social media activities.

The remainder of this section describes technological aspects of the data collector application and provides details about the nature of the collected data.

3.1. Technological overview

3.1.1. General considerations

It is well known that trust and transparency increases the willingness of consumers to share their personal data (Morey, Forbath, and Schoop 2015). To earn the trust of potential contributors, the data collector application was designed in a transparent and open way. Most importantly, the entire source code of the application was open sourced and made available[3] for anyone to review. To increase security, user activities were anonymized, and separated from the user database. This means that the extracted data from users (i.e. their social media locations) were not stored along with information that might be used to reverse engineer their identity. In addition, communication between the host server and users were encrypted with a secure HTTPS certificate[4].

3.1.2. Django framework

The core application was developed using Django[5], which is a popular web development framework that allows the quick creation and deployment of database driven modern websites. One example of a website developed with Django is The Washington Times[6]. Django can be customized with external packages that add extra functionality to the application. The application features the basic Django user model that allows the creation of local user profiles on the website. These users can be distinguished

[3] https://github.com/jlevente/social

[4] https://en.wikipedia.org/wiki/HTTPS

[5] https://docs.djangoproject.com/en/2.1/

[6] https://www.washingtontimes.com/

by an internal user ID. User profiles are stored in a PostgreSQL database running on the host machine of the application. Outside access to the database was restricted to prevent unauthorized use of user data.

3.1.3. User interface

The user interface consists of "views" that are basically HTML pages rendered by the Django engine from templates. Templates are rendered to pages on runtime using data defined with placemarkers to personalize the experience (e.g. programmatically print out a username). Django templates can be defined in a standardized way and can be reused across multiple pages. The home page of the application is shown in Figure 5. As a general guideline, the application was populated with content, such as information about the study, detailed technical description and explanation about participation. We found it important to provide as much information as possible for potential users to review. The main sub pages that can be selected at all times are as follows:

- **Home** (Figure 5)

  Landing page of the application with quick introduction about the research and directions for further actions
- **About the research** (Figure 6)

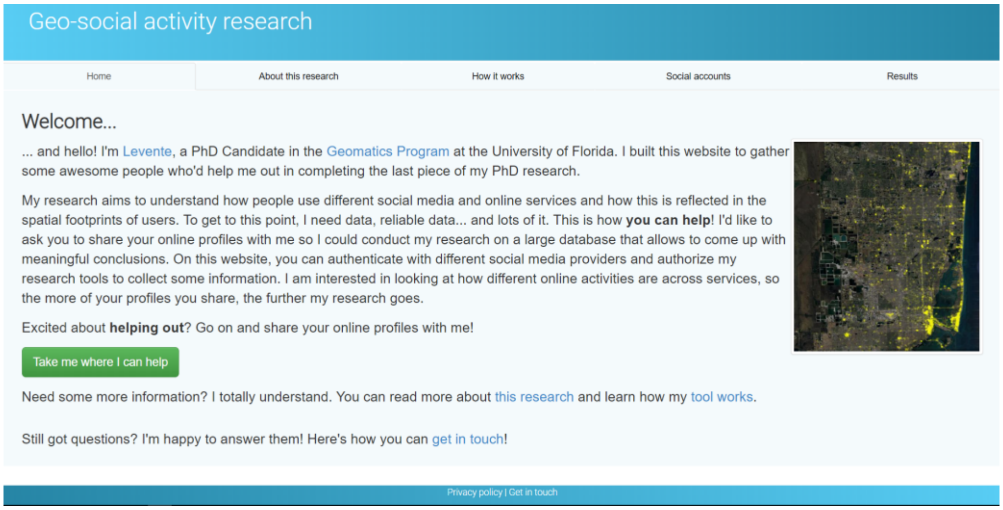
  Page dedicated to providing background information about the research, including references to already published work.
- **How it works**

  Page dedicated to the technical description of the application. This page explains in detail the technologies used, security measures taken to protect personal data, nature of data that is collected, and provides directions for users who wish to opt out from the research.
- **Social accounts** (Figure 7)

  Page where users can connect their online profiles and authorize the data collector tool to extract their online activities from the third party service. The layout of the page changes once a user is logged in. (Figure 8).
- **Results**

  Page dedicated to the presenting results to users once the study is completed

Figure 5: "Home page" of the data collector application as seen at https://research.jlevente.com



Welcome...

... and hello! I'm Levente, a PhD Candidate in the Geomatics Program at the University of Florida. I built this website to gather some awesome people who'd help me out in completing the last piece of my PhD research.

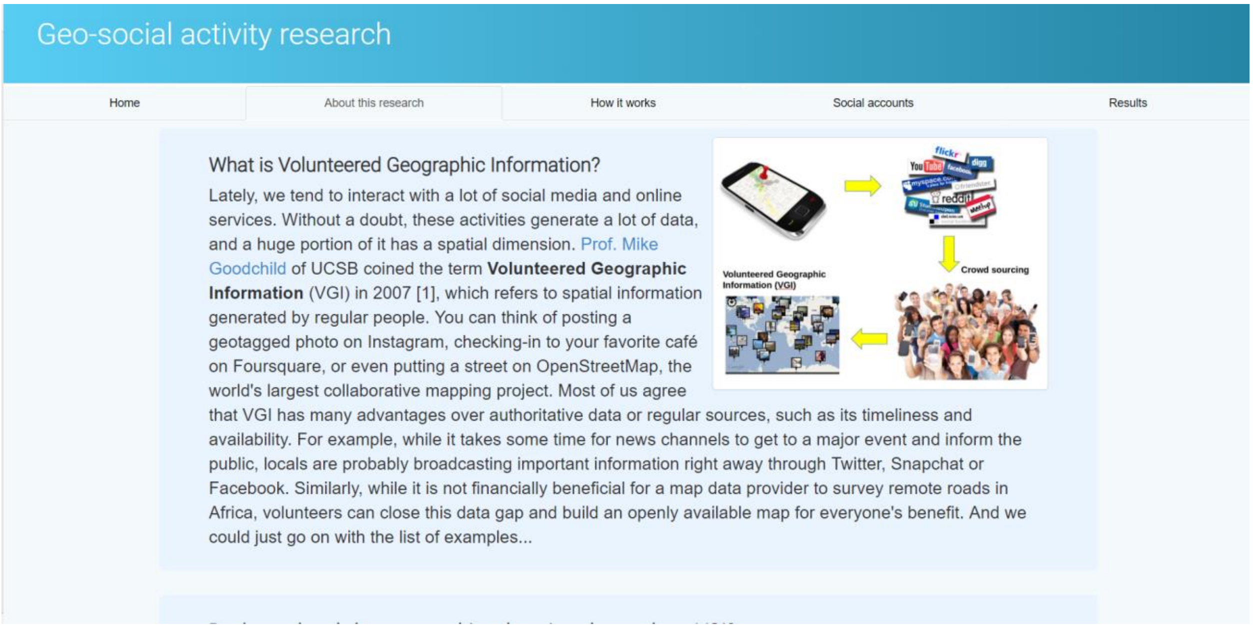
My research aims to understand how people use different social media and online services and how this is reflected in the spatial footprints of users. To get to this point, I need data, reliable data... and lots of it. This is how **you can help**! I'd like to ask you to share your online profiles with me so I could conduct my research on a large database that allows to come up with meaningful conclusions. On this website, you can authenticate with different social media providers and authorize my research tools to collect some information. I am interested in looking at how different online activities are across services, so the more of your profiles you share, the further my research goes.

Excited about **helping out**? Go on and share your online profiles with me!

Need some more information? I totally understand. You can read more about this research and learn how my tool works.

Still got questions? I'm happy to answer them! Here's how you can get in touch!

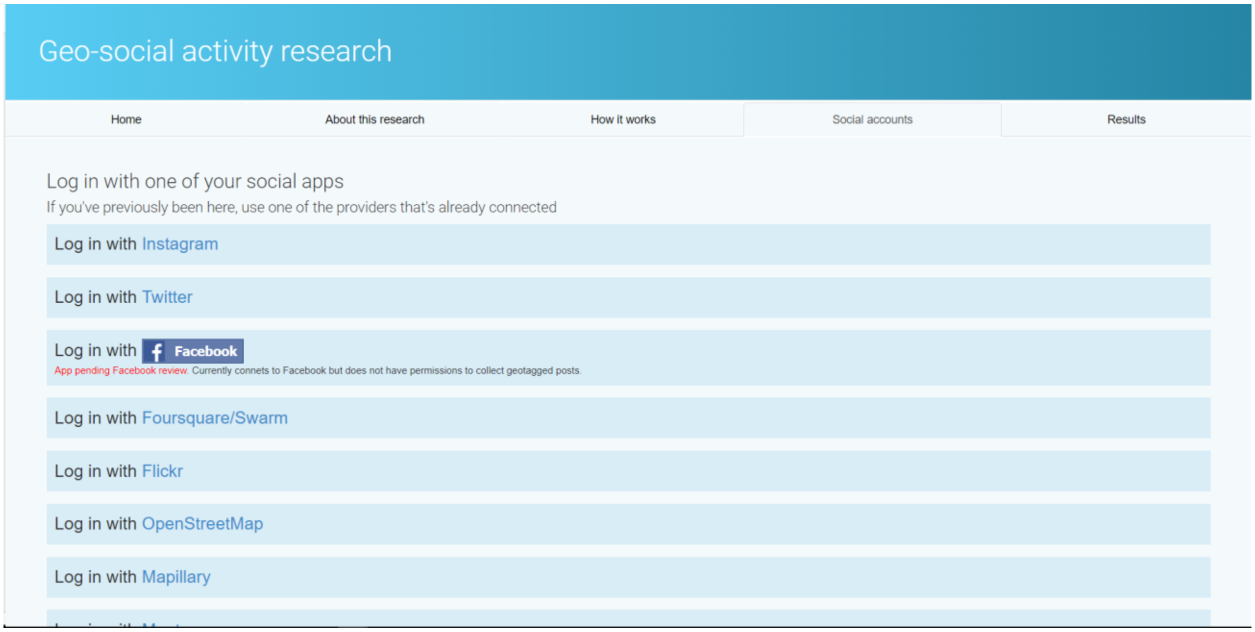
Figure 6: “About research” page describing background information about the research as seen on https://research.jlevente.com



What is Volunteered Geographic Information?

Lately, we tend to interact with a lot of social media and online services. Without a doubt, these activities generate a lot of data, and a huge portion of it has a spatial dimension. Prof. Mike Goodchild of UCSB coined the term **Volunteered Geographic Information** (VGI) in 2007 [1], which refers to spatial information generated by regular people. You can think of posting a geotagged photo on Instagram, checking-in to your favorite café on Foursquare, or even putting a street on OpenStreetMap, the world's largest collaborative mapping project. Most of us agree that VGI has many advantages over authoritative data or regular sources, such as its timeliness and availability. For example, while it takes some time for news channels to get to a major event and inform the public, locals are probably broadcasting important information right away through Twitter, Snapchat or Facebook. Similarly, while it is not financially beneficial for a map data provider to survey remote roads in Africa, volunteers can close this data gap and build an openly available map for everyone's benefit. And we could just go on with the list of examples...

Figure 7: “Social accounts” page before signing up for the research as seen on https://research.jlevente.com



Log in with one of your social apps

If you've previously been here, use one of the providers that's already connected

Log in with Instagram

Log in with Twitter

Log in with Facebook

App pending Facebook review. Currently connets to Facebook but does not have permissions to collect geotagged posts.

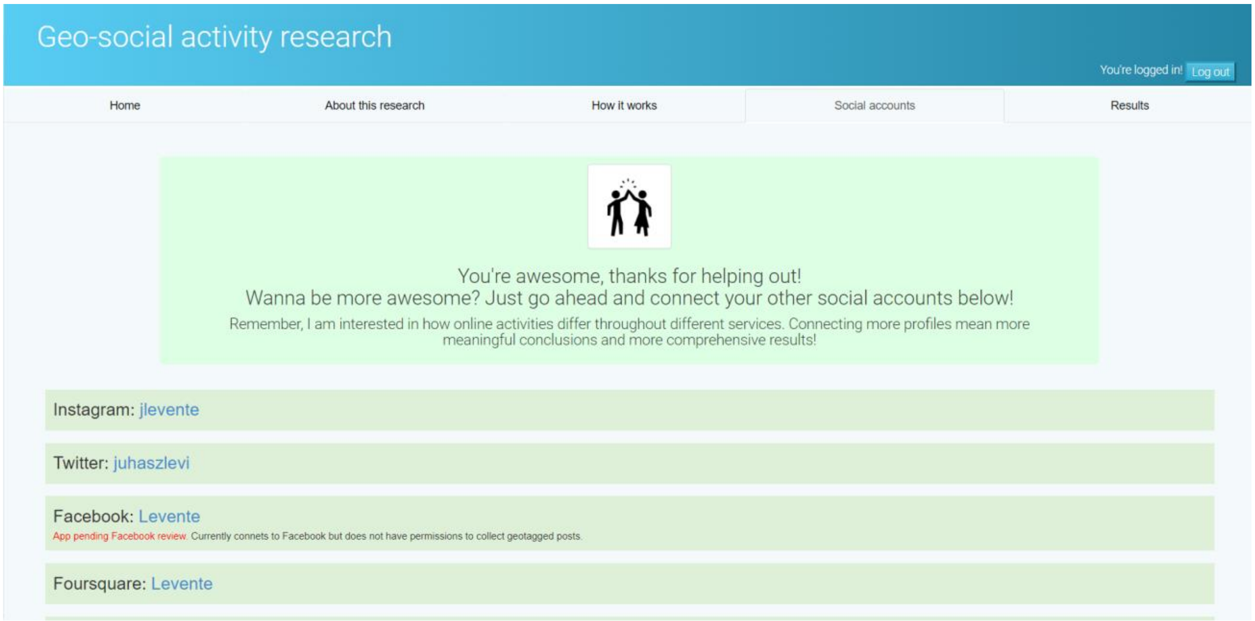
Log in with Foursquare/Swarm

Log in with Flickr

Log in with OpenStreetMap

Log in with Mapillary

Figure 8: Changed layout of the “Social accounts” page once a user is signed in as seen on https://research.jlevente.com



You're awesome, thanks for helping out!
Wanna be more awesome? Just go ahead and connect your other social accounts below!
Remember, I am interested in how online activities differ throughout different services. Connecting more profiles mean more meaningful conclusions and more comprehensive results!

Instagram: jlevente

Twitter: juhaszlevi

Facebook: Levente
App pending Facebook review. Currently connets to Facebook but does not have permissions to collect geotagged posts.

Foursquare: Levente

3.1.4. Cloud database

To increase security, spatial data of user activities were stored in a different PostgreSQL database separate from the one storing user data. This other database was running on the Amazon Web Services platform[7]. The choice of a cloud database was also made to allow for greater scalability in case of increased data volume. Scalability in this regard means an upgrade for more storage and resources within the same cloud architecture eventually allowing more processing power and the storage of more data. Spatial data was stored with the PostGIS[8] extension of PostgreSQL. Data was organized into different tables corresponding to each data platform (platforms and data are explained in Section 3.2) .To ensure the identification for the same individual users across tables, the unique identifier of each user from the Django user model was stored along with the data.

3.1.5. Authentication and data extraction process

Spatial data from different social media and VGI platforms are extracted through their Application Programming Interfaces (API) (Juhász, Rousell, and Jokar Arsanjani 2016). An API standardizes and defines the ways of interaction between software components. In the context of this research, it is a well-defined request-response system where servers (e.g. Twitter) respond to client application requests. In the case of Twitter, the response contains the tweeting history of a user, including geocoded messages.

To programmatically do this, users need to authorize the data collector application to perform some actions on their behalf (e.g. request tweeting history). This can be achieved through an authorization process in which users are redirected to a provider website (e.g. to Twitter), enter their credentials, and explicitly authorize the application running at https://research.jlevente.com to extract this information. Upon acceptance, users are redirected to the data collector application, which also receives so called authentication tokens that can be used for data extraction. This process is described in Juhász, Rousell,

[7] https://aws.amazon.com/rds/postgresql/
[8] https://postgis.net/

and Jokar Arsanjani (2016). The workflow follows the OAuth 2.0 standard[9] (except for OpenStreetMap which uses OAuth 1.0[10]) and was implemented using django-allauth, a third party package extending the django base user model with social accounts. A social account in this context refers to a user's online profiles and its authorization tokens.

Once the authorization tokens are obtained, the data collector can request user activities from social media and VGI platforms on behalf of the user using predefined API endpoints. Once the data is retrieved, it is inserted into the cloud storage described in Section 3.1.4. A description of data from each platform along with the API endpoints is provided in Section 3.2.

3.2. Considered platforms and data

Table 4 summarizes the data sources considered in this research. For each platform, a short description is given and a reference to the API endpoint from where the data can be extracted. It was expected that a comprehensive list of services need to be included in the study in order to build a meaningful database, since users 1) may not use most of these services, or 2) may decide to not share data from a service. To ensure that the final database contains a sufficient number of cross-platform activities (i.e. users actively use multiple services), nine services were chosen that cover a wide range of online activities. Mapillary and OpenStreetMap can be considered VGI platforms where users focus on generating geodata. Instagram, Twitter, Foursquare, Flickr, Strava, Meetup can be considered social media applications with a focus on the interaction between users (e.g. "liking" each other's pictures). iNaturalist, a citizen science application, contains features of both platform types since users generate useful data on purpose (i.e. recording observations for scientific research) while interacting with each other (i.e. helping each other identify what species are shown on photos).

An interesting aspect of looking at the online activities of users is how it matches with real world locations. Therefore, another data source was also considered. Google Location Services constantly track people if the functionality is enabled on their smartphones. This results in a detailed (both spatially and temporally) dataset of the user's whereabouts. An activity space extracted from the true locations of a person can be matched with social media activity spaces to see how well they align. However, Google Location History[11] cannot be obtained programmatically. Therefore, in a manual step users were presented with the option of downloading their own history from Google and sending it as an email attachment.

[9] https://oauth.net/2/

[10] https://oauth.net/1/

[11] https://www.google.com/maps/timeline?pb

Table 4: List of VGI and social media applications included in the research

Platform	Description of service	Data	Geometry
Instagram	*Photo and short video sharing social media platform*	**media/recent**[12] *List of photos/videos posted by the user*	point
Twitter	*Social media site allowing users to post short messages*	**statuses/user_timeline[13]** List of tweets submitted by the user	point/ polygon
Foursquare	*Restaurant/business review social media site with a check-in tracker*	**checkins[14]** List of check-ins (i.e. places visited) by the user	point
Flickr	*Photo sharing social media application focusing more on quality photos*	**people.getPhotos[15]** List of photos submitted by the user	point
OpenStreetMap	*VGI platform providing a worldwide database of map data*	**changesets**[16] List of changesets (grouped edis) made by the user	polygon
Mapillary	*VGI platform crowdsourcing street level photographs*	**sequences**[17] List of photo locations grouped into sequences	polyline
Meetup	*Social media application for scheduling, organizing and RSVPing events*	**events**[18] List of events the user attended to	point
Strava	*Activity and workout tracker social media application*	**activities**[19] List of workouts (bike rides, runs) by the user	polyline
iNaturalist	*Citizen science platform collecting observations of flora and fauna*	**observations**[20] List of observations made by the user	point
Google Location History	*Provides access to a user's location history if enabled within Google. Not a VGI or social media platform but can be used as a reference data.*	No API. Obtained directly from users through email	point

[12] https://www.instagram.com/developer/endpoints/users/
[13] https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline.html
[14] https://developer.foursquare.com/docs/api/users/checkins
[15] https://www.flickr.com/services/api/flickr.people.getPhotos.html
[16] https://wiki.openstreetmap.org/wiki/API_v0.6#Read:_GET_.2Fapi.2F0.6.2Fchangeset.2F.23id
[17] https://www.mapillary.com/developer/api-documentation
[18] https://www.meetup.com/meetup_api/docs/2/events/
[19] http://developers.strava.com/docs/reference/#api-Activities-getLoggedInAthleteActivities
[20] https://www.inaturalist.org/pages/api+reference#get-observations-username

4. Data description

The data collection tool described in Section 3 was made available at https://research.jlevente.com after which a promotion campaign began. The goal of this promotion was to reach as many people as possible in the hope of building a suitably large database to further explore cross-platform user behavior. The research project and site was promoted on several online outlets, such as social media sites (Twitter, Instagram, Facebook), mailing lists, and within community groups organized around mapping and geospatial technologies. Guest blog posts were also published in some outlets (e.g. the Mapillary Blog[21]) to increase the visibility of the research. In addition, several colleagues at different universities worldwide were asked to distribute a call within their networks and among their students. The campaign was actively conducted for about 3 weeks in May 2018.

Social media services are popular among young people, therefore, it was expected that the student population would be a major contributor to the data collection campaign. However, monitoring sign up numbers after each call that was sent out did not confirm this presumption. Instead, contributors of geospatial communities (e.g. OpenStreetMap and Mapillary) seemed to be the most active group in voluntarily sharing their online activities. It has to be noted that the data was anonymized, and therefore user identities were not backtraced.

4.1. Number of users and most common platforms

A total of 70 individual users engaged in the project by connecting a social account to the data collector website. The number of platforms included in the data collector is only a subset of available social media and VGI platforms. Therefore some users who contribute across other social media platforms might have been interested in this study, but were not able to actively engage in it. This effect can potentially be measured by analyzing server logs and extracting unique website visits.

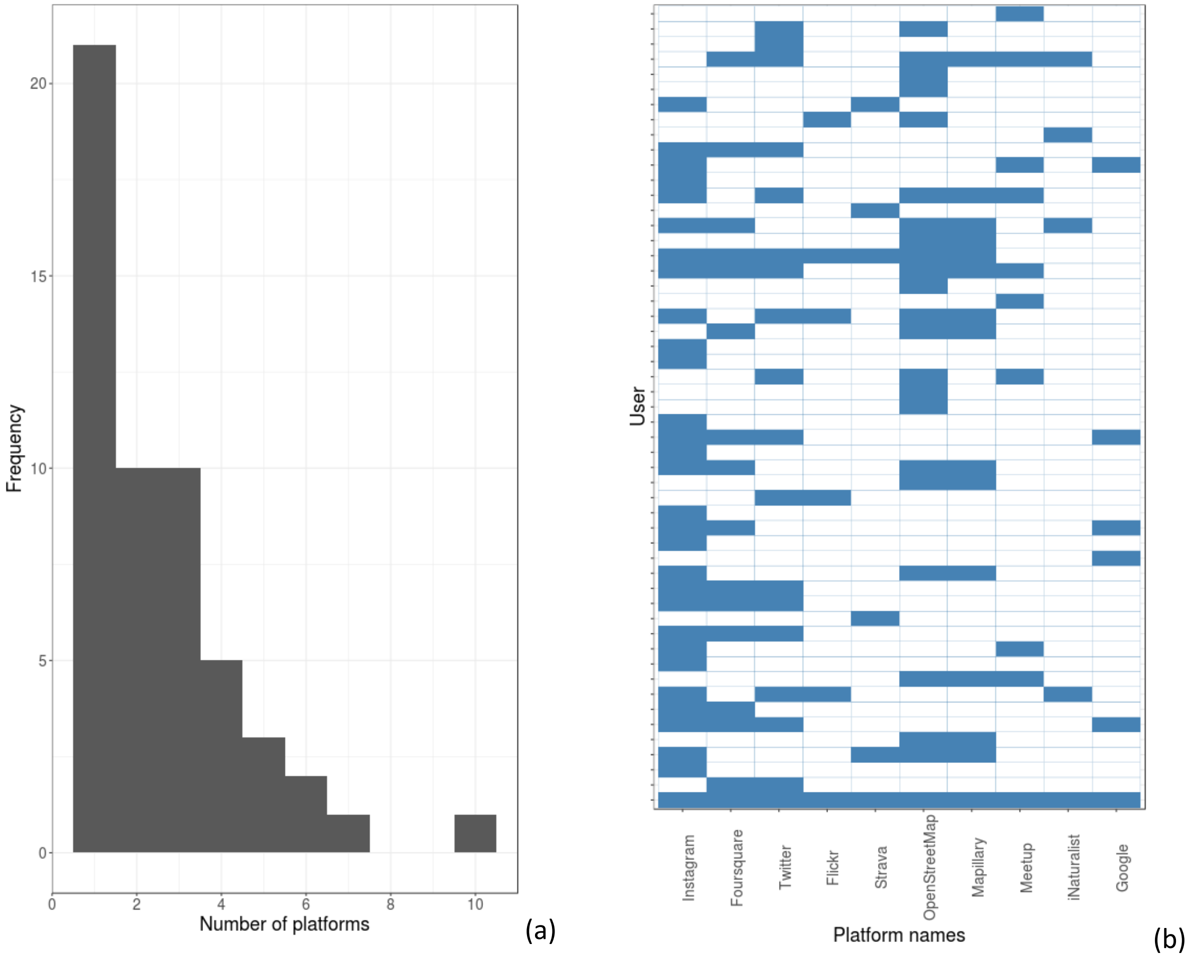
Another issue that lowered the usable data volume is the lack of geocoded activity. For 17 users, no spatial information was found on their online profiles. These cases include for example active Instagram users who do not share geolocation along with their photos. These users were excluded from further analysis. The histogram in Figure 9a shows the number of platforms with geospatial contributions for the remaining 53 users. The figure reveals that the activity 21 users can ofnly be found in one platform, which renders these users unsuitable for cross-platform activity analysis. The remaining 32 users shared activity information in 3.6 platforms on average (median: 3), with one user sharing activities in all 10 of the provided options.

Figure 9b shows for each of the 53 users with spatial activity (vertical axis) the platforms they shared. The most commonly shared platform was Instagram (30), followed by OpenStreetMap (23), Twitter (18) and Foursquare (16). The least commonly shared service was iNaturalist with only five users providing information about their citizen science observations. However, the data does not allow to distinguish between users not willing to share their activities and those not being active in this service. This ranking remains the same after excluding those 21 users with activity in only one platform. The most common platform pair was found to be between Mapillary and OpenStreetMap with all 15 Mapillary users contributing to OpenStreetMap as well, but not all OpenStreetMap contributors take Mapillary photos. This high overlap between the two user base can be explained by the similar nature of these services. Both Mapillary and OpenStreetMap contributors work towards improving openly accessible mapping data, either by editing map features (OpenStreetMap) or taking photographs that can be used to edit map

[21] https://blog.mapillary.com/update/2018/05/15/geo-social-media-mapping-research.html

features (Mapillary). Another instance of high overlap between two platforms can be observed within a subset of the social media domain. Five out of six Flickr contributors participating in this research were found to be active Twitter users as well.

Figure 9: Histogram of user numbers by platforms shared (a), and platform combinations shared by individual users (b)



4.2. Data volume

Table 4 describes that each site is used for different purposes which suggests that the data volume extracted from different platforms is not directly comparable. In addition, the geometry model also differs in many cases. For example, photo locations are represented as points in Flickr and Instagram but due to the large number of points taken automatically with Mapillary, polyline representation was chosen in that case. OpenStreetMap and Twitter (in some instances) represent the geographic coverage of a contribution with polygon geometries (i.e. the bounding box containing edits, or a place in Twitter). Nevertheless, looking at the data volume of user activities (keeping in mind different characteristics) provides useful insights into how these services are used by our sample users.

Table 5: Data volume of extracted features from social media and VGI platforms

	Instagram	**Foursquare**	**Twitter[1]**	**Flickr**	**Strava[2]**	**OpenStreetMap[3]**	**Mapillary[4]**	**Meetup**	**iNaturalist**	**Google** (x 1000**)**
Min	2	250	1	2	3	4	4	2	2	73
Max	653	8224	316	4502	837	7471	12231	113	1371	1,542
Mean	47	2378	102	1118	175	1342	2158	24	287	1132
Median	49	2124	45	380	58	329	801	14	24	1173
SD	164	1878	117	1753	352	2117	3647	33	606	306.5
# of users	30	16	18	6	6	23	15	10	5	6

[1] *Twitter: some tweets are represented with polygon geometries*

[2] *Strava: activities are represented as polylines*

[3] *OpenStreetMap: changesets are represented as polygons*

[4] *Mapillary: numbers in this table represent photo locations (nodes extracted from sequence polylines)*

Table 5 shows summary statistics about the data volume for each platform. The table reveals that users with different activity levels were reached which is illustrated by the relatively large variation of user activities within platforms. For example, the number of submitted OpenStreetMap changesets ranges from 4 to 7471. Similar patterns can be observed for all platforms included in this study

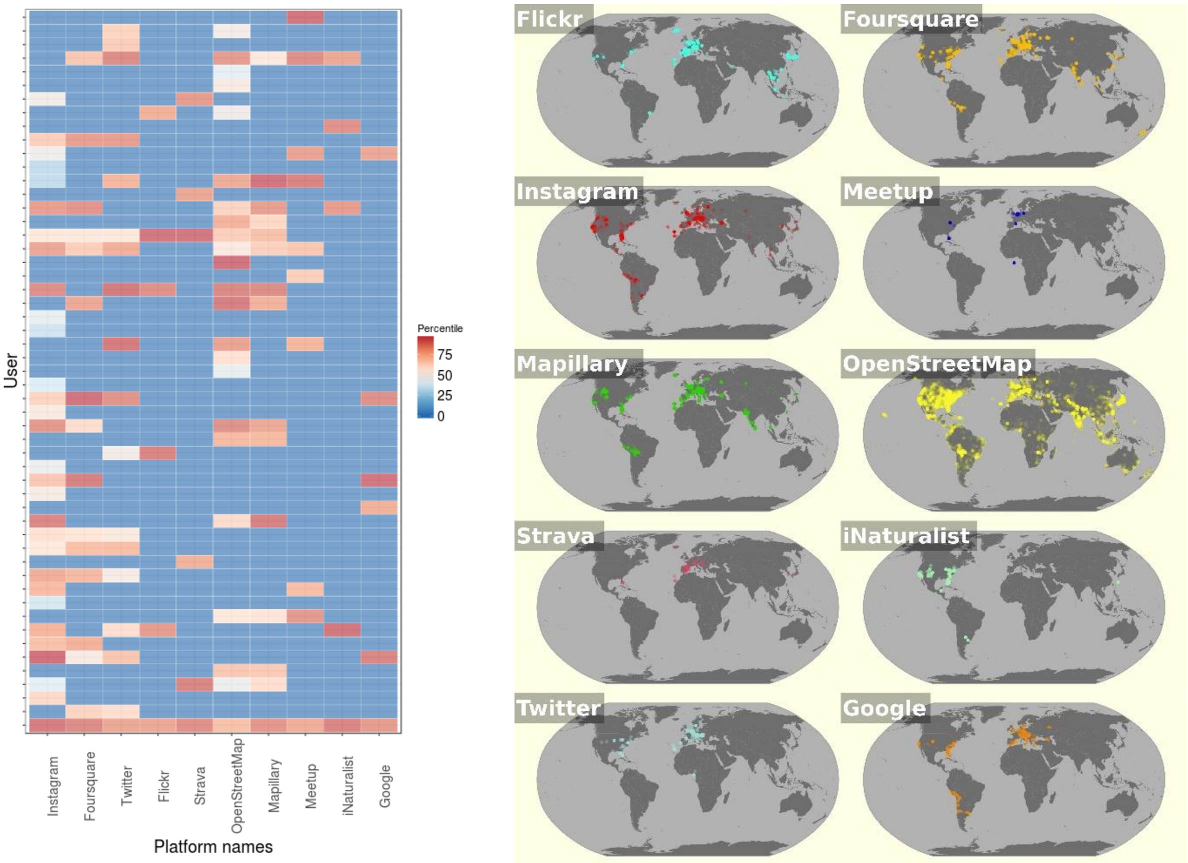
It is also expected that individual users do not contribute with the same rate to different services. To explore this, the rank of users was calculated for each platform. The rank of 1 denotes the most active user in that specific service. Figure 10a shows to what percentile this rank belongs to. Darker red colors mean that the user is one of the most active users for that platform. For horizontal lines (i.e. users), more red cells suggest that the user has an active online presence in many platforms. For example, the user situated at the bottom of Figure 10a can be considered a highly active user for all platforms. As opposed to this, most users did not make it to be among the top contributors in many platforms, which was also suggested by Figure 9.

4.3. Spatial distribution

By promoting the research in multiple online outlets and within several different groups, we aimed to engage a global group of users with activities from all over the world. This will help reveal if there are local differences in the way users use online services. For example, North American, European and Asian users might prefer to use different services. It will also allow us to explore the connection between global travel patterns and online activities, for example by looking at whether tourists prefer to use online services different from their regular activities when on vacation.

Figure 10b illustrates the spatial distribution of user activities for users in this study. It reveals that mainly users from North America and Europe were reached as activities seems to be more frequently found in these places for most platforms. The figure also suggests that local differences exists in the way people interact with the space in online services. This is most prominent in the case of OpenStreetMap, that allows for “remote mapping”. This means that OpenStreetMap users can edit virtually any areas in the world (e.g. by tracing roads from satellite imagery). This results in a more scattered pattern of activity where contributions can be found all around the world. Most other platforms however require physical presence, therefore these activities more correctly represent a user’s true activity area.

Figure 10: Percentiles of user activity within each platform (a), and spatial distribution of user activities for each platform (b)



(a)

(b)

5. Summary and future work

The rapid technological changes in recent years also transform the way people interact with online services. This study project tackled research questions in a relatively new research area within GIScience. Namely, it explored how user activities of the same individual can be analyzed across multiple services. The research focused on Volunteered Geographic Information and social media platforms. This final research report was logically organized into three main parts. Section 2 presented a case study that explored various ways to extract social media activity spaces from Instagram and Foursquare activities. The Section 3 presented the development process of a data collector application that can be used to build a reference database containing user activities for multiple platforms. In this research, nine platforms were used, namely Instagram, Foursquare, Twitter, OpenStreetMap, Mapillary, Flickr, iNaturalist, Strava and Meetup, which all have a geospatial dimension. An optional reference dataset, the Google Location History of users was also included which allows to compare the spatial footprint of social media activities to the real footprint of users. A description of the database in terms of user numbers, data volume and spatial distribution is also given in another section.

Section 2 described an early case study that analyzed the social media activities of 10 individual users to Instagram and Foursquare. Since existing ellipse based methods to estimate human activity spaces cannot capture the complexity of shapes, different methods were reviewed. It was found that borrowing the concepts of home ranges and utilization distributions from wildlife ecology is beneficial for the extraction of social media activity spaces. This case study was therefore the first attempt to apply these methods on the social media domain. We found that the choice of the range extraction method has a strong effect on mapped home and core regions. Some methods, such as Local Convex Hull based estimations perform well and can be applied on social media activity data. Section 2 also presented various ways in which the similarity of activities between different platforms can be mathematically quantifiable, which will be essential for larger scale quantitative studies.

Section 3 discussed the development of a data collector web application that can be used to build a reference database containing user activities for multiple platforms. In this research, nine platforms were used, namely Instagram, Foursquare, Twitter, OpenStreetMap, Mapillary, Flickr, iNaturalist, Strava and Meetup, which all have a geospatial dimension. An optional reference dataset, the Google Location History of users was also included which allows to compare the spatial footprint of social media activities to the real footprint of users. The developed tool was open sourced at https://github.com/jlevente/social, which has two benefits. First, it helps gaining the trust of potential users willing to "donate" their personal information for this research by allowing them to study the code. Furthermore, the code can be freely re-used and build upon by other researchers planning to conduct similar studies. The developed application was deployed on a remote server along with the implementation of several security related measures. This is an important aspect, since this research works with geocoded user activities that can be considered as sensitive information. A short description of data sources is also given in this section.

The last main part of this document, the description of the dataset collected with the tool explained above is presented in Section 4. After an active promotion campaign, several users were reached and asked to "donate" their VGI and social media activities to this research. Out of 70 users signing up with at least one of their online accounts 53 users were retained with at least some geocoded activity among all activities. Further, since this research project aims to analyze how the same individual uses multiple services simultaneously, 21 users with geocoded activities in only one services were excluded from further analysis. The remaining users provided their spatial footprints in 3.6 platforms on average, which allows to extend the case study presented in Section 1 to a larger scale. The most active user were active in all 10 platforms included in this research. The spatial distribution of activities suggest that mostly users from North America and Europe were reached, with some exceptions. For example, local activities in Asia are clearly visible in some cases.

Future work can be largely based on the newly built reference dataset. A possibility is to extend the analysis presented in Section 2 with the new dataset that contains more users and more platforms. For future work we also plan to include space-time geography information to automatically detect the contributions of an individual user to several platforms, which would further increase the number of users whose activities can be potentially analyzed. Even though the current reference dataset is an improvement, a larger database will be needed in order to draw generalized conclusions. For the analysis part, the temporal information attached to each spatial activity will also be utilized. This will help us to understand whether user preference of social media and VGI applications change over time or not, for example, with the introduction of a new service.

6. References

Budhathoki, N. R., & Haythornthwaite, C. 2013. "Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap." *American Behavioral Scientist* 57 (5):548-575. doi: 10.1177/0002764212469364.

Coen, M. H., Ansari, M. H., & Fillmore, N. 2011. "Learning from Spatial Overlap". Paper presented at the Proceedeings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Fransisco, CA.

Cvetojevic, S., Juhasz, L., & Hochmair, 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., & Longley, P. A. 2015. *Geospatial Analysis (5th ed.)*. 5 ed. Leicester: Matador.

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

Fieberg, J., & 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., & Wilmers, C. C. 2007. "LoCoH: nonparameteric kernel methods for constructing home ranges and utilization distributions." *PLOS ONE* 2 (2):e207.

Goodchild, M. F. 2007. "Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0 (Editorial)." *International Journal of Spatial Data Infrastructures Research (IJSDIR)* 2:24-32.

Hochmair, H. H. 2005. "Towards a Classification of Route Selection Criteria for Route Planning Tools." In *Developments in Spatial Data Handling*, edited by P.F. Fisher, 481-492. Berlin: Springer. doi.

Hochmair, H. H., Juhász, L., & Cvetojevic, S. 2018. "Data Quality of Points of Interest in Selected Mapping and Social Media Platforms." In *Progress in Location Based Services 2018. LBS 2018 (Lecture Notes in Geoinformation and Cartography)*, edited by Peter Kiefer, Haosheng Huang, Nico Van de Weghe and Martin Raubal, 293-313. Springer. doi.

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

Juhász, L., & Hochmair, H. H. 2016a. "Cross-Linkage between Mapillary Street Level Photos and OSM Edits." In *Geospatial Data in a Changing World: Selected papers of the 19th AGILE Conference on Geographic Information Science (Lecture Notes in Geoinformation and Cartography)*, edited by Tapani Sarjakoski, Maribel Yasmina Santos and L Tina Sarjakoski, 141-156. Berlin: Springer. doi: 10.1007/978-3-319-33783-8_9.

Juhász, L., & 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. doi: 10.1111/tgis.12190

Juhász, L., Rousell, A., & Jokar Arsanjani, J. 2016. "Technical Guidelines to Extract and Analyze VGI from Different Platforms." *Data* 1 (3):15. doi: 10.3390/data1030015.

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

Long, J., & 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.

Morey, T., Forbath, T., & Schoop, A. 2015. "Customer data: Designing for transparency and trust." *Harvard Business Review* 93 (5):96-105.

Neis, P., & Zielstra, D. 2014. "Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap." *Future Internet* 6 (1):76-106. doi: 10.3390/fi6010076.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., & Laakso, M. 2016. "Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information." *ISPRS International Journal of Geo-Information* 5 (5):55. doi: 10.3390/ijgi5050055.

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., & 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.