International Cartographic Association
Association Cartographique Internationale

Finding cultural heritage traces from modern social media

Francisco Porras-Bernardez [a, *], Georg Gartner [a], Nico Van de Weghe [b], Steven Verstockt [c]

[a] *Research Group Cartography, TU Wien, Vienna, Austria. francisco.porras.bernardez@tuwien.ac.at, Georg Gartner, georg.gartner@tuwien.ac.at*
[b] *CartoGIS Research Group, UGent, Ghent, Belgium. nico.vandeweghe@ugent.be*
[c] *Internet Technology and Data Science Lab (IDLab), UGent, Ghent, Belgium. steven.verstockt@ugent.be*

* Corresponding author

Keywords: Geo-Social Media, Flickr, KDE, Heatmap, Cultural Heritage, Location Based Social Networks

Abstract:

This work is on development within the framework of the project *Eureca: EUropean Region Enrichment in City Archives and collections* of the University of Ghent (IDLab, CartoGIS), the Technical University of Vienna (Research Group Cartography) and several city and state archives. *Eureca* focuses on revealing traces (i.e. origins or influences) of European regions that have shaped the cities in which we live today and will further develop tools to explore these traces when visiting a city. Different historical, architectural, economic, political, and cultural reasons form the base of these traces, and will be used as input to disclose cultural heritage items that can be linked to specific European regions and origins. The enriched metadata that will result in this project will be further usable to perform new fundamental research and applied studies, and to facilitate the exploitation of the collections to a broader public and attract new groups of cultural heritage consumers.

The specific focus of this work is on Geo-Social media (GSM) (Ostermann, 2015) as a source of information to identify these European traces of the past. The objective of this research is finding the *footprint* of Europeans visiting other euro-cities by determining areas of preference in a city for specific nationalities and during certain periods. The footprints represent areas of attraction for visitors in the city and the reasons for this attraction could be multiple: available services, architecture, historical/cultural hotspots, etc. Finding these modern footprints will be a base to identify the most visited cultural heritage points of interest (POIs) for specific nationalities or even cities of origin and during specific periods of the year. Finally, this will contribute to the development of location based services (LBS) that will help users to explore traces of their own region of origin in other European cities.

Social media data have been used in research widely and despite their multiple limitations, they have been proven useful for geographic research in different fields. Geotagged social media provide better insights on the spatial behaviour of their users. Some of the most used media in the literature include Foursquare, Twitter or Flickr. Foursquare is the least interesting for us because of its user base and amount of data available. Twitter provides a huge amount of geotagged text for semantic analysis but Flickr's user profile is more suited for tourist behaviour analysis. Furthermore, Flickr provides a well-developed set of Application Programming Interfaces (APIs) to enable easier access to their data.

The first phase of this research involved the data collection from Flickr via two of its APIs. There are several Flickr datasets openly available, nevertheless we opted for building our own collection to avoid problems related to accessibility, accuracy and temporal coverage. Metadata of each uploaded picture such as photo owner, uploading date, geolocation, etc. was retrieved. In a second process, another API will be used to obtain the user name, location (user manually-provided) and other attributes. This location attribute have to be processed because of the heterogeneity of the data format. If only *city* is provided, the places have to be matched to a gazetteer to determine the country.

The data retrieved covered a squared area of 68 Mill. km^2 representing a huge area around the continental Europe. In order to determine the nationality of each user the first source of information is the self-reported location included in her profile. Unfortunately, this information is often missing or can be simply false. For the majority of the users, the home location has to be inferred by some kind of method. A simple method based on previous works on home determination from user's GSM data (Li & Goodchild, 2012; Bojic et al., 2015) was developed and tested. To identify a country as user's home location, all the pictures uploaded during a year in each country were considered. If the temporal difference between last and first photo was greater than 6 months, the user was labelled as local resident in that country. For comparison purposes, a second threshold of 3 months was also applied. With both thresholds, in some cases users were labelled with double home location because of being present in both cities in the same year.

We are aware of some limitations of this approach. For instance, a user can visit two times the same city in the same year. Besides, those users uploading pictures between the end of one year and the beginning of the following one will

not be classified in that country. The nature of the Flickr user is a limitation itself; some individuals can upload one single photo and others may contribute thousands.

The method will be improved in future work by requiring a minimum of images uploaded during the chosen period. Also, it will be analysed the continuous stream of uploads during time instead of simply considering natural years. Additionally, the language of the title and tags could be used to infer the nationality. Moreover, the first information that will be taken into consideration is the self-reported home location obtained from the user's profile. This new approach will increase the number of users correctly labelled so that we can get a better differentiation between locals and tourists and between different nationalities. This will be key for our further analysis.

The uploaded photos can be visualised as points in the space given that we have their geolocation. We can generate a continuous raster surface from these points using Kernel Density Estimation (KDE) (Grothe & Schaab, 2009). These raster are heatmaps that represent areas of high concentration of pictures. These heatmaps represent a footprint of the visitors in the city. Thus, the areas more visited by tourists from a specific origin will be visible and also an analysis of the temporal evolution will be possible. The continuous surfaces built with KDE are very well suited for the task of determining vague areas open enough for further POIs identification in *Eureca*. In addition, to include areas of interest (AOIs) when dealing with open spaces like parks, squares or large buildings. Figure 1 shows examples of footprints in Vienna and Ghent.

The footprints will reveal the most preferred places for specific origins. Furthermore, all the footprints will be compared through spatial analysis. Using map algebra (Tomlin, 1990), we will obtain areas of common interest for Europeans and for instance classify the areas as high, moderate or low "Euro-visitor interest". This can be applied for aggregated groups e.g. Mediterranean nations, German-speaking countries, etc. In further steps, Flickr data from the rest of the world will be collected to apply the same approach for more groups.

Regarding the results already obtained, the final number of points retrieved was about 66 million and covered a period (2004-2018) representing Flickr photos from 62 countries. Initial research was done with a selection of 2 European cities and countries: Ghent (Belgium) and Vienna (Austria). Next steps will include all those countries fully retrieved from Flickr and the 10 European capitals with the highest amount of data available.

PHOTOS IN VIENNA		
Classification	**Photos**	**%**
Austrians 6mo	209,117	50.77%
Austrians 3mo	21,455	5.21%
Belgians 6mo	**4,279**	**1.04%**
Belgians 3mo	1,519	0.37%
Rest no locals	170,832	41.48%
Double (A/B)	4,681	1.14%
TOTAL	**411,883**	**100%**

PHOTOS IN GHENT		
Classification	**Photos**	**%**
Belgians 6mo	55,625	62.35%
Belgians 3mo	6,696	7.51%
Austrians 6mo	**302**	**0.34%**
Austrians 3m	259	0.29%
Rest no locals	24,944	27.96%
Double (B/A)	1,384	1.55%
TOTAL	**89,210**	**100%**

Table 1. Classification of pictures. Left: Vienna / Right: Ghent

Table 1 shows the classification performed by the initial method using both thresholds. 1,04% of all the pictures within Vienna were classified as uploaded by Belgians (6 months), whereas 41,48% were contributed by any other foreign nationality. Meanwhile, Austrians (6 months) uploaded 0.34% of the pictures within Ghent, whereas 27.96% came from other nationalities.

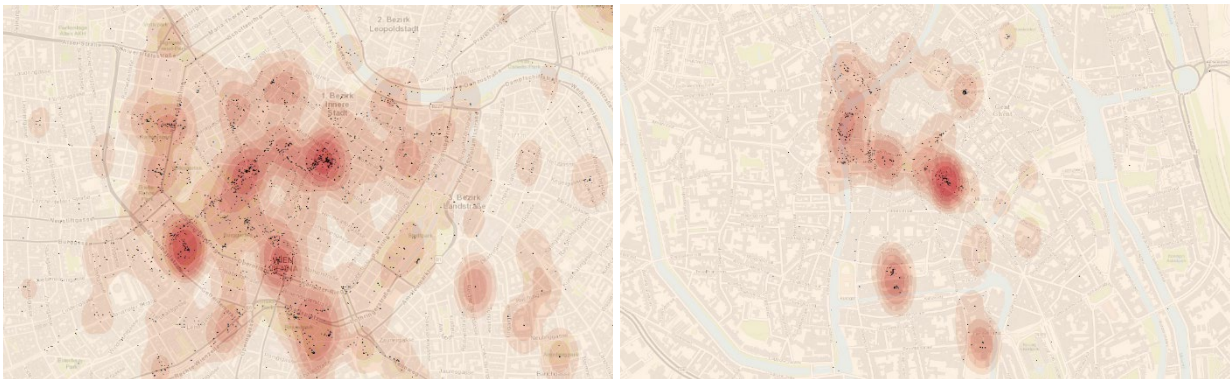

Figure 1. Visitor footprint. Left: Belgians in Vienna / Right: Austrians in Ghent (6 months threshold).

Several conclusions can be drawn from the initial results. The number of photos available for each city can vary greatly; this has to be considered in terms of relative representativeness. The inclusion of the self-reported user information should improve the theoretical accuracy of the user home location determination. It could serve as some kind of ground truth to estimate precision and recall of our own classification method. Increasing the dataset with world coverage and classifying the home location of all the global users should reduce the number of ambivalent cases by applying other strategies. In sum, further work is required but this initial approach seems to be useful for establishing GSM as a valuable modern source of information to identify cultural heritage POIs/AOIs that will reveal European traces of the past within the *Eureca* project.