

Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database

Jacinto Estima
ISEGI – Universidade Nova de
Lisboa
Campus de Campolide
Lisboa, Portugal
Jacinto.estima@gmail.com

Cidália C. Fonte
Department of Mathematics
University of Coimbra /
INESC Coimbra
Coimbra, Portugal
cfonte@mat.uc.pt

Marco Painho
ISEGI – Universidade Nova de
Lisboa
Campus de Campolide
Lisboa, Portugal
painho@isegi.unl.pt

Abstract

Volunteered Geographic Information has been increasing exponentially over the last years, capturing the attention of the scientific community. Researchers have been very active exploring a vast amount of initiatives and trying to develop methodologies and possible real applications for this new source of geographic information. Land Use/Cover production is one of the areas where this type of geographic information might be very useful. In this paper we evaluate if geo-referenced and publicly available photos from the Flickr initiative can be used as a source of geographic information to help Land Use/Cover classification. Using the Corine Land Cover nomenclature, we compare the classification obtained for selected photo locations, against the classification obtained from high resolution satellite imagery for the same locations. We conclude that this source cannot be used alone for the purpose of Land Use/Cover classification but we also believe that it might contain helpful information if combined with other sources.

Keywords: Volunteered Geographic Information, Land Use/Cover, Flickr, Accuracy assessment.

1 Introduction

The availability of Volunteered Geographic Information (VGI), a term coined by Michael Goodchild in 2007 [1], has been increasing exponentially in the last years, due to the introduction of the Web 2.0 and the increasing availability of low cost positioning equipment, among other technology improvements [2, 3].

The scientific community has been trying to explore this vast amount of information to use it in the solution of world problems. Navigation [5], crisis management and emergency response [6, 7] are just a few examples of the research that has been conducted in the last years. Although some of the main issues and concerns pointed out about this type of Geographic Information (GI) are related with their heterogeneity, quality control and metadata absence, among others, it is also agreed that their main advantages are associated with their temporal coverage and volume [8].

More recently, Fischer [4] argued that, in some cases, when VGI is used for different purposes than those for which volunteers have contributed, it can be seen as a not-so-Volunteered Geographic Information and had termed this as involuntary geographic information (iVGI).

The specific application of this type of data for Land Use/Cover (LULC) production had also been investigated recently, achieving some interesting results (e.g. [9, 10]). For example, Urban Land Use was produced using data from OpenStreetMap [11]. Photos and descriptions extracted from the Degrees of Confluence Project were used to assess the accuracy of Land Cover Maps [12, 13]. In a paper studying Flickr photos [9], the authors explored a collection of Flickr geotagged and publicly available photos in terms of their

temporal and spatial distributions over Continental Portugal and also its distribution over Land Use/Cover classes, using as a reference the European Corine Land Cover (CLC) database. They concluded that this source of VGI might be very valuable for the purpose of LULC production when combined with other sources.

In this paper we evaluate whether what is seen on Flickr photos can be used to provide a LULC class or not and, whenever it is possible to do so, we evaluate if the identified class is correct. This evaluation was made using a stratified sample of photos considering the CLC level 1 classes as strata, and comparing the classes extracted from the Flickr photos to the class assigned to the photos location using high resolution satellite imagery.

The paper is structured as follows. After the introduction we describe the methodology used, followed by the presentation and discussion of the obtained results. The paper ends drawing some conclusions and indicating future research directions.

2 Material and Methods

2.1 Description of the study area and datasets

The defined study area is the Portuguese municipality of Coimbra, covering an area of approximately 300 km².

Three datasets were used in this study: 1) the geo-referenced Flickr photos for the study area over the period ranging between 2004 and 2013, corresponding to a total of 4977 photos; 2) the CLC database, composed by the version 16 (04/2012) for the CLC2006 inventory, downloaded from the

European Environment Agency (EEA)¹; 3) the high resolution satellite imagery, with 30cm spatial resolution, available for the study area at the ArcGIS software as basemap.

Figure 1 shows the CLC map for the study area and Figure 2 shows the points corresponding to the spatial location of the photos situated in each of the three CLC classes used for this analysis, overlaid with the high resolution satellite images.

Figure 1 - CLC level 1 classes in Coimbra municipality

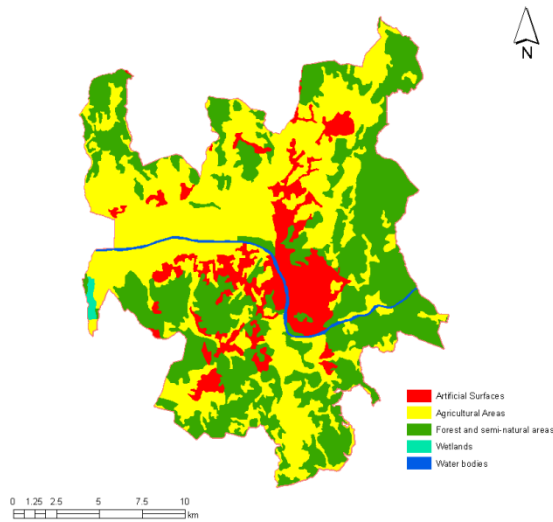
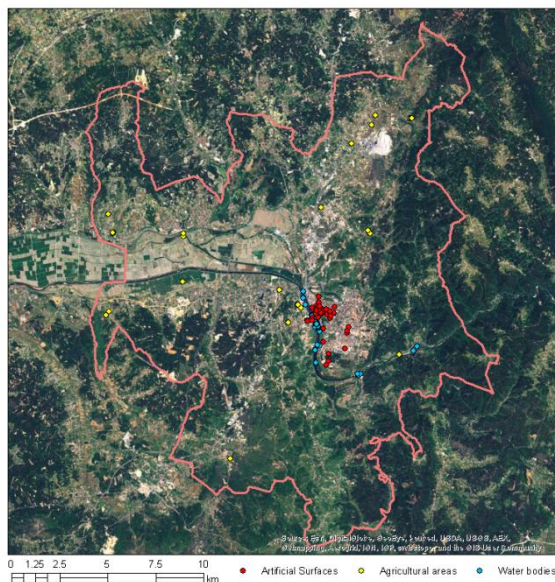


Figure 2 - Location of the sample Flickr photos used for the analysis



2.2 Data processing

For the purpose of this preliminary study, the position associated to the 4977 photos was intersected with the CLC level 1 classes and the three classes that from a user

¹ <http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-2>

perspective were more likely to have information were selected, namely classes 1, 2 and 5, respectively Artificial Surfaces (AS), Agricultural Areas (AA) and Water Bodies (WB), corresponding to a total of 4892 photos.

Table 1 summarizes the distribution of selected photos over the three CLC classes.

Table 1 - Summary of Flickr photos

CLC Classes	Flickr Photos
Class AS	4703
Class AA	64
Class WB	125
Total	4892

2.3 Methods

The methodology adopted to conduct this analysis was as follows:

1. A stratified sample of 60 photo locations was selected for each of the three classes chosen for the analysis, considering the CLC classes as strata;
2. An expert classification of Flickr photos was done, based on the image content interpretation, according to the CLC nomenclature. Using the photo assigned to each location, we first evaluate whether it was possible to attribute a class or not and, when possible, a class was then assigned to the corresponding location;
3. Flickr photo locations were overlaid with the high resolution satellite imagery, and a land cover class was assigned to each location based on the imagery interpretation.

3 Results and discussion

Following the methodology described in section 2.3, Table 2 and Table 3 show the resultant classification of the locations based on the interpretation of Flickr photos and satellite imagery respectively. Besides CLC level 1 classes, two more classes were considered: “Not Clear” and “Not Good”. The “Not clear” class refers to those photos where more than one class is present and it is not clear which one, if any, is predominant, and the “Not Good” class refers to those photos that do not show predominantly any type of landscape and therefore cannot be used in LULC classification.

Having a closer look to the spatial distribution of the photos relatively to the CLC classes (see Figure 1 and Figure 2) it is clear that for class AS they are centered in the more touristic places of the city of Coimbra and for WB most photos are located in the region of the river where touristic boats operate. A more even distribution can be seen for the class AA.

Results from the interpretation of Flickr photos are shown in Table 2. The percentage of photos considered “not good” for LULC classification is relatively high, with 41.7% for class AS, 26.7% for class AA and 21.7% for class WB. Another negative aspect is related with the percentage of photos classified as “not clear”, with 18.3%, 26.7% and 6.7% for classes AS, AA and WB respectively. These two classes together, representing photos that do not fit in any CLC class, embody a high percentage of photos with classes AS, AA and

WB getting respectively 60%, 53.4% and 28.4%. In the opposite direction, the value for locations correctly classified is very low for all the classes with the class AA getting the worst value, below 20%. Looking at the value for photos wrongly classified, we can see a good result for class AS, with 0%, while classes AA had 28.3% and class WB 15%.

Table 2 - Classification of Flickr photos

		CLC Classes containing the photo's location		
		Class AS	Class AA	Class WB
		Classification based on Flickr photos	Class AS	24
Class AA	--		11	--
Class F	--		--	3
Class W	--		6	--
Class WB	--		1	34
Not Clear	11		16	4
Not Good	25		16	13
Total of photos	60		60	60
Correct		40.0%	18.3%	56.7%
Wrong		0.0%	28.3%	15.0%
Not clear		18.3%	26.7%	6.7%
Not good		41.7%	26.7%	21.7%

Table 3 - Classification of Flickr photos' locations based on the satellite imagery

		CLC Classes containing the photo's location		
		Class AS	Class AA	Class WB
		Classification based on satellite imagery	Class AS	60
Class AA	--		39	--
Class F	--		--	7
Class W	--		--	--
Class WB	--		--	52
Not Clear	--		20	1
Total of points	60		60	60
Correct			100.0%	65.0%
Wrong		0.0%	1.7%	11.7%
Not clear		0.0%	33.3%	1.7%

During the classification process, however, some problems related to the use of the Flickr photos became apparent, contributing to increase the negative aspects of this source. Among the collection of photos analyzed, we have seen photos showing predominantly people, photos taken inside houses, photos showing small details and photos taken far from what is shown in the image reflecting a high level of zoom. This last case was particularly present for photos considered inside class WS, where although the picture shows mainly water it is easy to realize that the pictures were taken from land.

The assignment of classes to the photo locations using the satellite imagery produced the results shown in Table 3. It can be seen that 100% of the points located at the AS areas in the CLC map were actually assigned to the class AS, with values of respectively 65% and 87% for the classes AA and WB.

At some locations it is not clear to which class the point should be assigned, due to the mixture of classes observed at the vicinity of the point and to the fact that the minimum mapping unit of the CLC map is 25ha, which means that the class choice cannot be done analyzing only what exists at each point, but also looking at a larger vicinity. In Table 3 it can be seen that this difficulty occurred for 20 points. However, a closer analysis showed that only 4 of these points correspond to different locations. The other 16 points, even though corresponding to different Flickr photos, actually were assigned exactly to the same spatial location, meaning that the volunteer assigned the same coordinates to a large number of photos. Moreover, an analysis of the photos as well as the photos tags also showed that there are also other photos wrongly geotagged, since the coordinates assigned are far from the real location where the photo was taken.

4 Conclusions and future research

The concentration of Flickr photos in touristic places, leading to a spatially poor distribution of locations, was confirmed. Consequently, urban areas and touristic places are likely to have more photos than rural areas. This spatial clustering needs to be assessed in order to understand its impact and the possibilities of its use to further parameterize the validation process. Besides, not all the photos were given a CLC class either because the predominant class was not clear or the image was not showing any type of landscape at all. As an example, for classes AS and AA, more than 50% of the photos could not be classified.

Some of the issues need further research in order to find possible solutions to overcome them. This is the particular case of WB class, where some pictures were taken from land and therefore their location is not inside the water body. One possible strategy to make those photos useful would be to buffer the water body with an acceptable distance and consider also inland locations inside that buffer.

Based on what has been exposed, we might conclude that this source of VGI is not suitable for LULC classification when used alone. Nevertheless we believe that it might contain useful information that can be helpful if combined with other sources. On the other side, different classes had different results and not all the classes were explored in this study. Therefore we are planning to continue this research to explore all the classes, compare the results at different locations and create pre-processing procedures that can improve the quality of the results.

It is also planned to combine this source of VGI with other sources, such as Panoramio, OpenStreetMap, among others, to understand if some of the faced issues would be solved and/or minimized when multiple sources are combined.

Acknowledgements: This research has been partially supported by COST Action TD1202, Mapping and the Citizen Sensor. C. Fonte work was partially supported by the Portuguese Foundation for Science and Technology under project grant PEst-OE/EEI/UI308/2014.

References

- [1] Goodchild, M. (Nov. 2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*. vol. 69, no. 4. pp. 211–221.
- [2] Elwood, S., Goodchild, M. F., & Sui, D. Z. (May 2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Ann. Assoc. Am. Geogr.* vol. 102, no. 3. pp. 571–590.
- [3] Heipke, C. (Nov. 2010). Crowdsourcing geospatial data. *ISPRS J. Photogramm. Remote Sens.* vol. 65, no. 6. pp. 550–557.
- [4] Fischer, F. (2012). VGI as Big Data: A New but Delicate Geographic Data-Source. *Geoinformatics April/May*. no. May. pp. 46–47.
- [5] Holone, H., Misund, G., & Holmstedt, H. (2007). Users Are Doing It For Themselves: Pedestrian Navigation With User Generated Content. in *International Conference on Next Generation Mobile Applications, Services and Technologies*. no. Ngmast.
- [6] Goodchild, M., & Glennon, J. A. (Sep. 2010). Crowdsourcing geographic information for disaster response: a research frontier. *Int. J. Digit. Earth*. vol. 3, no. 3. pp. 231–241.
- [7] Zook, M., Graham, M., Shelton, T., & Gorman, S. (Jan. 2010). Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Med. Heal. Policy*. vol. 2, no. 2. pp. 6–32.
- [8] Leung, D., & Newsam, S. (2010). Proximate sensing: Inferring what-is-where from georeferenced photo collections. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 2955–2962.
- [9] Estima, J., & Painho, M. (2013). Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover. in *Computational Science and Its Applications – ICCSA 2013*. vol. 7974. pp. 205–220.
- [10] Estima, J., & Painho, M. (2013). Exploratory analysis of OpenStreetMap for land use classification. in *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '13*. pp. 39–46.
- [11] Arsanjani, J., Helbich, M., Bakillah, M., Hagenauer, J., & Zipf, A. (2013). Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science*, 27(12), 2264–2278. doi:10.1080/13658816.2013.800871
- [12] Iwao, K., Nishida, K., Kinoshita, T., Yamagata, Y., 2006. Validating land cover maps with Degree Confluence Project information. *Geophysical Research Letters* 33.
- [13] Foody, G.M., Boyd, D.S., 2013. Using Volunteered Data in Land Cover Map Validation: Mapping West African Forests. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6, 1305–1312.