SEMANTIC VALIDATION OF SOCIAL MEDIA GEOGRAPHIC INFORMATION: A CASE STUDY ON INSTAGRAM DATA FOR EXPO MILANO 2015

F. Migliaccio [1] *, D. Carrion [1], F. Ferrario [2]

[1] DICA – Geodesy and Geomatics Section, Politecnico di Milano – (federica.migliaccio, daniela.carrion)@polimi.it
[2] DICA – Geodesy and Geomatics Section, Politecnico di Milano – francesco4.ferrario@mail.polimi.it

Commission IV, WG IV/10

KEY WORDS: Social media, geolocation, spatial analysis, geosocial analytics, Instagram

ABSTRACT:
Social media data, such as Instagram posts, can be associated with spatial positions. This information can be exploited to perform spatial analyses, such as identifying distribution patterns of points representing the positions of social media users during an emergency or while attending a specific event or exhibition. However, the geolocation provided by Social Media Geographic Information (SMGI) needs to be validated, in order for the spatial data to be used in a meaningful way in subsequent spatial analyses or mapping procedures. In this paper, a case study is presented based on Instagram data collected during the first two months of the Expo Milano 2015 exhibition, where the spatial data have been validated by exploiting the semantic component of the posts.

1. INTRODUCTION: INSTAGRAM AND SMGI

The analysis of the participation of the public to large/international events has been carried on for a long time with survey methodologies based on direct approaches such as ratings of the participants and on indirect indicators of an economic and statistical nature.

Recently, however, new types of analysis have emerged, based on the use of data known as Social Media Geographic Information (SMGI), (Goodchild et al., 2007; Roick et al., 2013; Steiger et al., 2016). Often, this is (un)voluntary information provided by the users themselves through their activity on social networks, such as Facebook, Twitter (Hahmann et al., 2014) or Instagram (Silva et al., 2013). In this way, a significant contribution in the production of spatial and geolocated data, called Volunteered Geographic Information (VGI), can be provided through the use of GPS or other localization systems, over devices such as smartphones, when contributing to social media. Platforms such as Facebook, Instagram or Twitter are in fact accompanied by a geolocation service that allows users to accompany their own post with geographical information, see e.g. (Carrion et al., 2017).

The information collected from social networks, considering both the spatial and the temporal content, is fundamental to study important situations related to the territory, by exploiting the posts directly shared by the users themselves. The multiplicity of applications and their implications suggests the growing importance of these data, which have become objects of interest for technologies in continuous and full evolution.

Geographical information in posts shared through social platforms can be obtained either by:
(i) "geotagging" (which extracts the spatial contents from the Exchangeable image file format – Exif – data attached to the image, or allows the user to associate a pair of geographic coordinates to the image),
(ii) exploiting the geosocial networking (which is the simple sharing of the user's position without associated multimedia data).

In order for the spatial data to be subsequently used in a meaningful way in spatial analysis or mapping procedures, the geolocation provided in this way needs to be validated. In the case study presented here, considering that the spatial patterns of the distribution of people in the area of Expo 2015 could be derived from the geotags of the Instagram posts, a simple validation procedure is proposed where the semantic component is used to validate the spatial position of the Instagram posts.

The idea behind this work is that, since data from SMGI are supplied with both spatial and temporal information, they can be very helpful or even vital in identifying patterns of distribution of people. This can be done not only in the case of a crowd attending large events such as the Expo 2015 exhibition in Milano, but also, and most importantly, in situations that occur during and after natural disasters or crisis events. So, the Expo 2015 case presented here is to be considered as an example of a general procedure which can be applied to validate SMGI data in a simple yet effective way.

Finally, when dealing with "big data" it is often assumed that lots of "low quality" data can provide good information in the same way as less high quality data. However, it is not always true that "big" data can be gathered from SMGI e.g. when producing maps for emergency situations in real time, in particular considering that the georeferenced posts correspond to a small percentage of the total (Carrion et al., 2017).

2. A DATASET OF INSTAGRAM POSTS ON EXPO 2015

The dataset used for the validation analysis were acquired from the social network Instagram through the use of the Spatext tool implemented in Phyton 2.7, which allows to download Instagram data (either characterized by geo-referencing or not) based on their spatial or textual location (hashtag) (Migliaccio

* Corresponding author

https://doi.org/10.5194/isprs-archives-XLII-2-W13-1321-2019 | © Authors 2019. CC BY 4.0 License.

et al., 2018). The test point dataset was downloaded in the Esri shapefile format; each point representing a post published by Instagram users in the period between May 1, 2015 (the exhibition opening date) and June 22, 2015, and covering an area surrounding the Expo 2015 exhibition site. In total, 102,908 posts were found to be referring to the Expo 2015 event and geolocated over the Expo 2015 area (see Figure 1 and Figure 2).

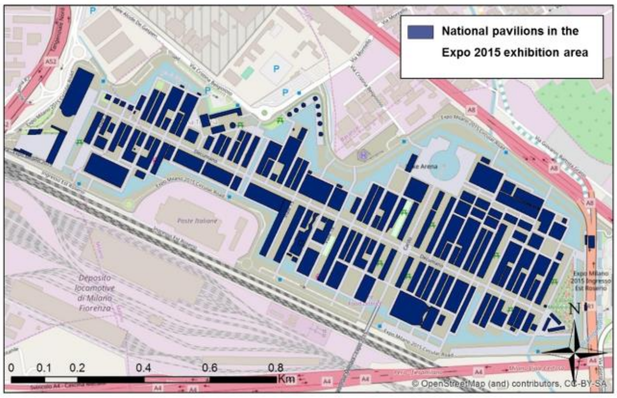
National pavilions in the Expo 2015 exhibition area
0
0.1
0.2
0.4
0.6
0.8
Km
N
© OpenStreetMap (and) contributors, CC-BY-SA

Figure 1. National pavilions inside the Expo 2015 area

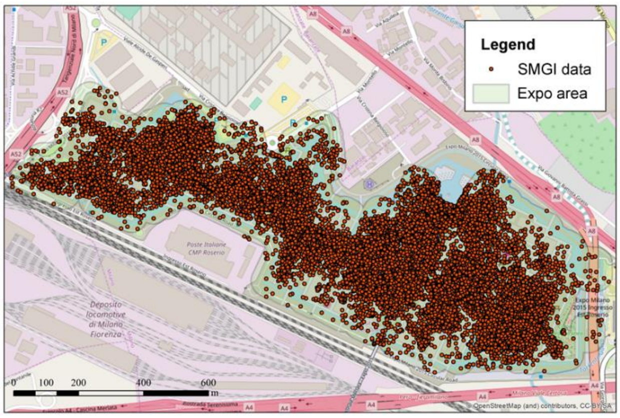

Figure 2. Dataset of points corresponding to Instagram posts referring to the Expo 2015 event, inside a 50 m buffer of the exhibition area

Each point was identified and characterized by ten different attributes, from a spatial, temporal, semantic and thematic point of view:

- the spatial component of the data is defined by the *Shape* field which, in addition to specifying the geometric type of the data, also provides its position in space;
- the temporal information depends on the *DATA* field, containing the time and date of creation of the post, and from the *TIME_UNIX* field, which identifies the moment of publication of the post by the number of seconds elapsed from the conventional midnight date of January 1, 1970;
- the fields that define this semantic aspects of the post are *PLACE* (the name of the geotag chosen by the user), *TAGS* (the set of *hastags* chosen by the user to describe their post) and *CAPTION* (the text accompanying a post); these represent in fact information provided by the users of the social platform themselves, expressing judgments, criticisms, observations and considerations on their own experience; they are therefore very useful indications in the context of semantic analysis and validation of the data;
- the thematic component of the data is represented by the Instagram user identifier field (*USER_anony*) and by the two fields *LIKE* and *COMMENT*, which are used for the interaction of the users of the social platform with the published post.

Data were preliminary filtered drawing a 50 m buffer around the actual Expo 2015 site, and excluding all the posts outside the Expo + buffer area. This left 74,669 posts for the case study purposes.

3. SEMANTIC VALIDATION OF THE POST GEOLOCATION

3.1 Validation concept for the posts geolocation

The semantic content of the data provides important information in the context of the validation of the location of the post, allowing to identify discrepancies between the geographic and the semantic content. The problem is to find out if positions attached to SMGI data, such as Instagram posts, can be really considered as representative of the spatial position at which the social media users are while publishing the post.

This of course opens up another quite important issue, namely the possible discrepancy between the time at which the picture has been taken and the time at which it has been published. For the purposes of this study, a simplifying hypothesis has been made, considering the two epochs as coincident.

In the present study, the idea was to use the semantic component of the posts to check the geolocation provided by the social platform or by the mobile device against the presence of selected keywords in the posts. Here, a very important point is the definition of a list of appropriate keywords. In the Expo 2015 case study, the keywords list contained the names (in different languages) of the Nations participating to the exhibition with their national pavilions.

For the validation procedure, the idea was to filter the position of each post based on two conditions: (i) the coordinates are included in a 50 m buffer around a specific Nation’s pavilion, and (ii) the field ‘*PLACE*’ must contain the name of the Nation (Figure 3).

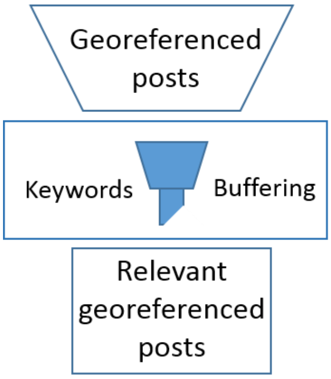

Figure 3. Validation procedure

The conditions have been separately applied in two different steps, and the procedure has been repeated by exchanging the order of the steps. In the following sub-sections it will be described how the procedure was applied and some specific cases will be presented.

3.2 ‘*PLACE*’ – Buffer validation procedure

In this case the filtering was firstly applied on the semantic content of the post (‘*PLACE*’) and secondly on its position within the buffer. In this way, it was possible to check if a post which refers to a certain pavilion, is also geolocated within that same pavilion, thus confirming somehow its geolocation.

Pavilion	N. of posts inside buffer	N. of validated posts	% of validated posts
Angola	146	137	94%
Argentina	131	129	98%
Austria	85	57	67%
Azerbaijan	225	225	100%
Belgium	65	56	86%
Brazil	721	690	96%
China	408	380	93%
Colombia	71	68	96%
Czech Republic	65	22	34%
Ecuador	53	49	92%
Estonia	165	165	100%
France	134	129	96%
Germany	201	197	98%
Israel	89	89	100%
Italy	420	272	65%
Japan	635	622	98%
Kazakhstan	190	190	100%
Malesia	85	84	99%
Mexico	115	107	93%
Monaco	55	40	73%
Morocco	50	20	40%
Netherlands	181	176	97%
Qatar	622	569	91%
Russia	827	820	99%
Slovakia	70	70	100%
South Korea	811	465	57%
Spain	245	239	98%
Switzerland	58	54	93%
Thailand	347	341	98%
Turkey	92	91	99%
UAE	296	281	95%
UK	704	693	98%
USA	199	101	51%
Vietnam	70	47	67%

Table 1. Results of the ‘PLACE’ – Buffer validation procedure

The procedure was performed for 34 pavilions (corresponding to as many Nations participating to the Expo 2015 exhibition).

In the first step, the keywords used to select the *‘PLACE’* values included the name of the relevant Nation, often combined with the term “padiglione” (“pavilion” in Italian) or with “Expo” or “Expo 2015”. The selection was done manually. In the second step (geographic validation), the selected posts not included in a 50 m buffer around the Nation’s pavilion were rejected. On average, 87% of the posts selected on the basis of the ‘PLACE’ were thus validated.

Table 1 shows the results of this validation procedure for all the considered pavilions. Figure 4 and Figure 5 show two examples of the distribution of the filtered posts at the end of the validation procedure, for two different pavilions.

In some cases (less than 25%), the second step of the validation procedure filtered out a percentage of the posts as high as 43%. All these cases were examined and they either regarded pavilions for which the number of posts was very small (i.e. less than 50), or pavilions for which the coordinates positioned them in a cell outside the 50 m buffer.

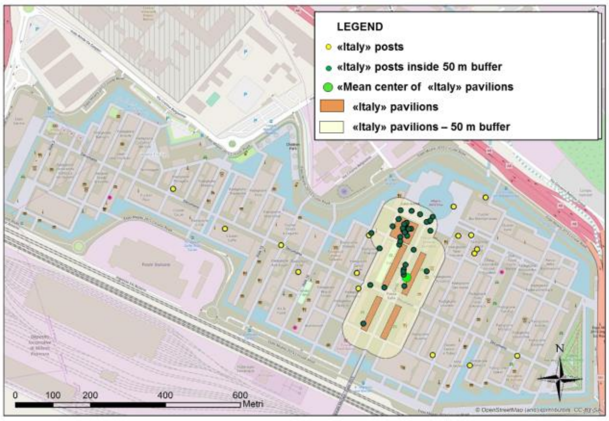
LEGEND
«Italy» posts
«Italy» posts inside 50 m buffer
«Mean center of «Italy» pavilions
«Italy» pavilions
«Italy» pavilions – 50 m buffer
N
0 100 200 400 600
Metri
© OpenStreetMap (and) contributors, CC-BY-SA

Figure 4. Example of results of the two-step ‘PLACE’-Buffer procedure for the semantic validation of the posts (case: Italy pavilions); green dots represent the validated posts

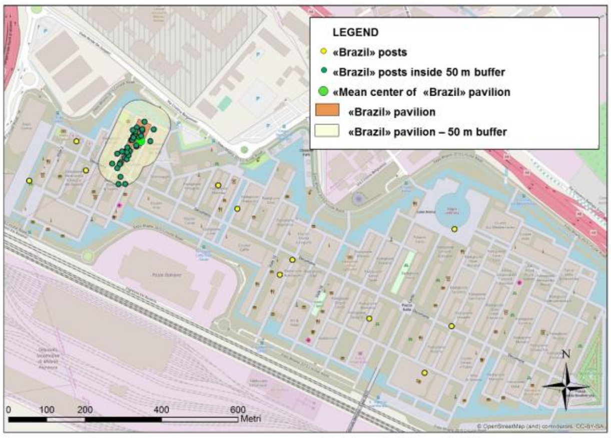
LEGEND
«Brazil» posts
«Brazil» posts inside 50 m buffer
«Mean center of «Brazil» pavilion
«Brazil» pavilion
«Brazil» pavilion – 50 m buffer
N
0 100 200 400 600
Metri
© OpenStreetMap (and) contributors, CC-BY-SA

Figure 5. Example of results of the two-step ‘PLACE’-Buffer procedure for the semantic validation of the posts (case: Brazil pavilion); green dots represent the validated posts

3.3 Buffer - '*PLACE*' validation procedure

A way to cross-check the results of the validation procedure is to reverse the order of the steps: so, accordingly, the filtering was applied again on the input dataset, but firstly on the position of the post and then on its semantic content ('*PLACE*'), for the 34 pavilions considered. Table 2 shows the results of the Buffer - '*PLACE*' validation procedure. As it could be expected, in this case the second step (semantic validation) was the very selective one, validating on average only 6% of the posts, which had been included into the 50 m buffer.

Pavilion	N. of posts inside buffer	N. of validated posts	% of validated posts
Angola	4544	137	3%
Argentina	1747	129	7%
Austria	2395	57	2%
Azerbaijan	2639	225	9%
Belgium	2904	56	2%
Brazil	4625	690	15%
China	3111	380	12%
Colombia	2846	68	2%
Czech Republic	718	22	3%
Ecuador	2780	49	2%
Estonia	2114	165	8%
France	5748	129	2%
Germany	3713	197	5%
Israel	1910	89	5%
Italy	10166	272	3%
Japan	4158	622	15%
Kazakhstan	24218	190	1%
Malesia	1101	84	8%
Mexico	4557	107	2%
Monaco	2538	40	2%
Morocco	1524	20	1%
Netherlands	4911	176	4%
Qatar	2600	569	22%
Russia	3329	820	25%
Slovakia	3132	70	2%
South Korea	5137	465	9%
Spain	5247	239	5%
Switzerland	2900	54	2%
Thailand	3652	341	9%
Turkey	2484	91	4%
UAE	25367	281	1%
UK	5251	693	13%
USA	2705	101	4%
Vietnam	3104	47	2%

Table 2. Results of the Buffer - 'PLACE' validation procedure

The results were always consistent with those obtained with the first procedure, confirming the need of taking into account the semantic content of the posts to assess the validity of the geographic location provided by the social platform.

The results can also be seen in Figure 6 and Figure 7, which show two examples of the distribution of the filtered posts at the end of Buffer - '*PLACE*' procedure for two specific pavilions.

3.4 Check of the validation procedure

Finally, in order to control the results of the validation procedure, the posts were also checked in two different ways. Firstly, the information content of 'TAGS' and 'CAPTION' for the filtered posts was examined to verify that the relevant keywords referring to the correct pavilion were present also in those fields.
Then, as a final independent check, the mean center of the point positions of the filtered posts for each pavilion was compared with the geometric center of the corresponding pavilion polygon. For most pavilions, the mean center decidedly moved towards the geometric center of the pavilion when the validated posts were used, thus confirming the increasing reliability of the filtered dataset. An example is presented in Figure 8.

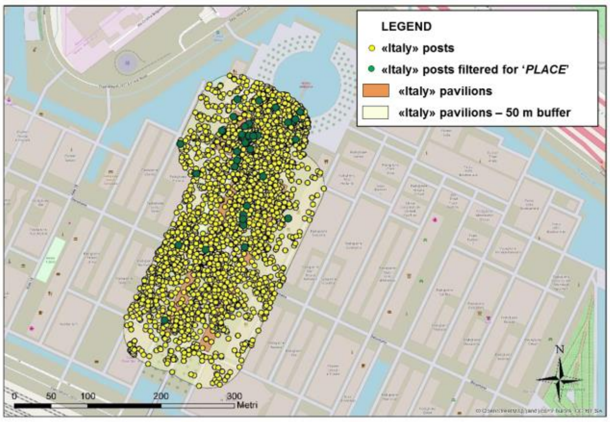
LEGEND
● «Italy» posts
● «Italy» posts filtered for '*PLACE*'
«Italy» pavilions
«Italy» pavilions – 50 m buffer
N
0 50 100 200 300
Metri

Figure 6. Example of results of the two-step Buffer-'PLACE' procedure for the semantic validation of the posts (case: Italy pavilions); green dots represent the validated posts

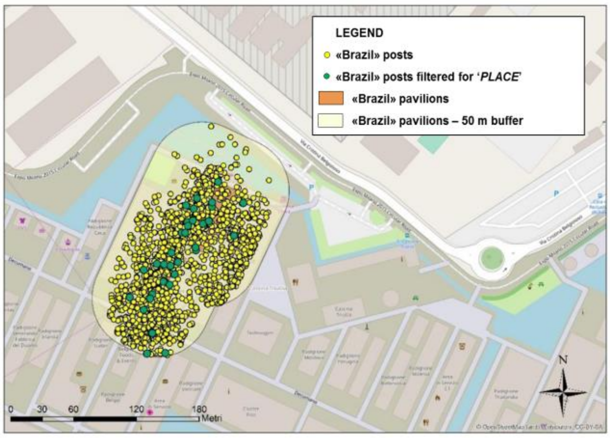
LEGEND
● «Brazil» posts
● «Brazil» posts filtered for '*PLACE*'
«Brazil» pavilions
«Brazil» pavilions – 50 m buffer
N
0 30 60 120 180
Metri

Figure 7. Example of results of the two-step Buffer-'PLACE' procedure for the semantic validation of the posts (case: Brazil pavilion); green dots represent the validated posts

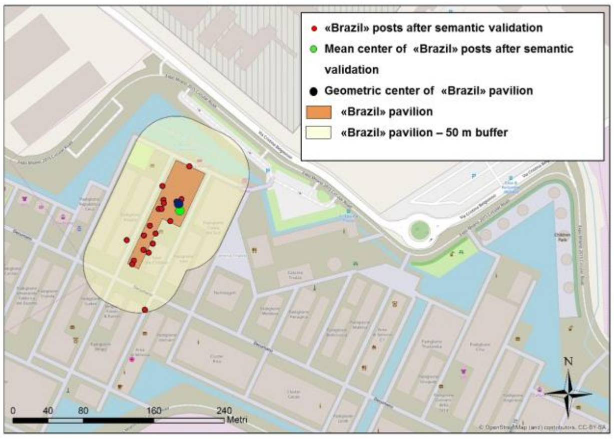
● «Brazil» posts after semantic validation
● Mean center of «Brazil» posts after semantic validation
● Geometric center of «Brazil» pavilion
«Brazil» pavilion
«Brazil» pavilion – 50 m buffer
N
0
40
80
160
240
Metri
© OpenStreetMap (and) contributors, CC-BY-SA

Figure 8. Example of results of the check of the validation procedure (case: Brazil pavilion); black dot represents the geometric center of Brazil pavilion polygon; green dot represents the mean center of validated posts

4. RESULTS

The validation procedure presented in this paper and applied to Instagram posts allowed to acquire a dataset of semantically validated data which were associated to coordinates more reliably referring to the content of the posts themselves.

The procedure is based on a two-step filtering strategy. It has been tested on a study case and cross-checked by changing the order of the filtering steps. The results proved that after the validation, the remaining points were better representative of the spatial position of the geometric features to which the posts were referring.

In particular, this procedure appears to be very effective in the cases when the geographic location is not specified by the author of the post, but is provided by the social platform based on generic terms (in this case, such terms as "Expo" or "Expo 2015"). Since these terms are not referring to a specific location inside a large area, the corresponding spatial position is then generally defined as coincident with the "barycentre" of the area. Locations close to the barycentre are thus erroneously picked, while the semantic data clearly show that those posts were not related to such locations.

In particular, the results of the semantic filtering showed that an average 93.7 % posts geotagged inside each specific pavilion of the Expo 2015 exhibition were filtered out, meaning that the semantic content of SMGI data is a highly selective tool, always leading to a consistent reduction of the available dataset. The cases of pavilions with the highest percentages of posts rejected after semantic validation of the data were explained by the fact that those posts were geotagged by a general text such as "Expo 2015" or similar, so that the associated coordinates just pointed to a "barycenter" point in the Expo 2015 site.

5. CONCLUSIONS AND LESSONS LEARNT

The geolocation provided by Social Media Geographic Information needs to be validated, in order for the spatial data to be used in a meaningful way in subsequent spatial analyses or mapping procedures. In this paper, a case study has been presented based on Instagram data collected during the first two months of the Expo Milano 2015 exhibition, where the spatial data have been validated by exploiting the semantic component of the posts.

The proposed procedure could be generalized and applied not only to similar popular events to which large crowds of people participate (such as concerts or sport events), but also, and most importantly, to emergency situations and crisis management, where SMGI is used to support the production of emergency maps and crisis events (Di Martino et al., 2017), (Earle et al., 2012).

Finally, it is important to highlight the lessons learnt while working with the semantic content of SMGI. As a matter of fact, there is a number of difficulties connected with the analysis of the semantic content of posts. First of all, the procedures that may be applied for the study of this information component generally appear to be of complex application. Indeed, manual operations are particularly burdensome in terms of time, as they entail the screening of the individual posts in order to achieve the semantic validation. Besides, the semantic interpretation can also be very complex, and that is why, now, it does not seem to be feasible in a completely automatic way. One must in fact interpret social and interactive motivations of the social platform users, which are at times not clear or unambiguous, given the presence of colloquial expressions and the absence of a formal language. Thus, the choice of the keywords to be considered to perform the data filtering is essential to be able to effectively validate the posts.

So, we can conclude with the following consideration: the interpretation of the semantic content of the posts is a complex undertaking and cannot be exclusively performed in an automated way, however it is fundamental in order to obtain more reliable samples of spatial data from SMGI.

ACKNOWLEDGEMENTS

The authors wish to thank Prof. Michele Campagna (University of Cagliari, Italy) for providing the test data set used in this study.

REFERENCES

Carrion, D., Migliaccio, F., Pagliari, D., 2017. Exploring geolocation issues in social media analytics - A case study with Tweet messages". *Proc. of the 6th Information and Communication Technologies International Conference (ICTIC)*, Vol. 6, Issue 1.

Di Martino, S., Romano, S., Bertolotto, M., Kanhabua, N., Mazzeo, A., Nejdl, W., Towards Exploiting Social Networks for Detecting Epidemic Outbreaks. Global Journal of Flexible Systems Management: 1-11, 2017

Earle, P.S., Bowden, D.C., and Guy M., Twitter earthquake detection: earthquake monitoring in a social world. Annals of Geophysics 54.6, 2012

Goodchild, M. F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69.4, pp. 211-221.

Hahmann, S., Purves, R.S., Burghardt, D., 2014. Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, pp. 1-36.

Migliaccio, F., Pagliari, D., Carrion, D., Gaspari F., 2018. Geostatistical and temporal analysis of Instagram data. The EXPO Milano 2015 case study. *Proc. of the 7th Information and*

Communication Technologies International Conference (ICTIC), Vol. 7, Issue 1, pp. 101 - 105.

Roick, O., Heuser, S., 2013. Location based social networks–definition, current state of the art and research Agenda. *Transactions in GIS*, Vol. 17, Issue 5, pp. 763-784.

Sakaki, T, Makoto O., and Yutaka M., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web. ACM*.

Silva, T. H., Melo, P. O., Almeida, J. M., Salles, J., & Loureiro, A.A., 2013. A picture of Instagram is worth more than a thousand words: Workload characterization and application. *2013 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 123-132.

Steiger, E, Westerholt, R and Zipf, A., 2016. Research on social media feeds – A GIScience perspective. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F, Purves, R. (eds.) European Handbook of Crowdsourced Geographic Information, pp. 237–254. London: Ubiquity Press.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W13-1321-2019 | © Authors 2019. CC BY 4.0 License.