Identification of disaster-affected areas using exploratory visual analysis of georeferenced Tweets: application to a flood event

Valentina Cerutti
ITC Faculty of Geo-Information Science and Earth Observation
University of Twente
Enschede, The Netherlands
v.cerutti@utwente.nl

Georg Fuchs
Fraunhofer IAIS
Schloss Birlinghover
Sankt Augustin, Germany
Georg.Fuchs@iais.fraunhofer.de

Gennady Andrienko
Fraunhofer IAIS
Schloss Birlinghover
Sankt Augustin, Germany
Gennady.Andrienko@iais.fraunhofer.de

Natalia Andrienko
Fraunhofer IAIS
Schloss Birlinghover
Sankt Augustin, Germany
Natalia.Andrienko@iais.fraunhofer.de

Frank Ostermann
ITC Faculty of Geo-Information Science and Earth Observation
University of Twente
Enschede, The Netherlands
f.o.ostermann@utwente.nl

Abstract

To enable decision makers to conduct a rapid assessment of the situation during the disaster response phase and improve situational awareness, we propose an approach to identify affected areas using geo-spatial footprints. These geo-spatial footprints summarize information and threats and are derived from georeferenced social media messages and authoritative data sources. The combination of data mining techniques for data pre-processing and exploratory visual analysis is a promising approach for dealing with heterogeneous data under time pressure. This paper presents the first steps towards this objective by using georeferenced Tweets to define the geospatial footprint of a flood event that occurred in Italy in 2013. After cleaning the data, density-based clustering, distance-bounded spatio-temporal event clustering and data-driven territory tessellation techniques were applied; visual analysis was used to define the best parameters combination. A comparison between the results and ground-truth data was performed. The proposed methods showed positive results in the identification of areas affected by the flood at regional scale. The combination of data mining with visual analysis for parameters setting proved to be an intuitive and fast procedure that could help decision makers deal with geosocial media data and assist them with rapid assessment of the situation.

Keywords: Twitter, geospatial footprint, data mining, visual analytics, disaster management.

1 Introduction

This paper addresses a crucial challenge in disaster response: the allocation and prioritization of relief resources. The effects of many natural or man-made disasters are difficult to localize and subsequently assess. The geographic and temporal extent of disaster effects rarely matches precisely the visible boundaries of damaged infrastructure, burned vegetation, or flooded areas. Although improved remote sensing tools and methods produce ever more imagery at ever finer resolutions, the detection of, and response to many disaster-related effects is still an unsolved challenge. In-situ measurements from sensors and reports from the affected population can complement the remote sensing perspective, improve situational awareness, contribute to a common operational picture, and help decision makers to assess the spatio-temporal footprint of a disaster, which may extend well beyond the originally affected area. To enable decision makers to rapidly assess the situation, we propose to assist them in (i) the detection of relevant information regarding affected areas through data mining and visual analysis, and (ii) the contextualization with relevant information by generating a comprehensive spatio-temporal footprint of a disaster event.

This work represents the first steps towards an improved conceptual management of crisis information retrieved from social media together with authoritative data, in order to define the geospatial footprint and assess the temporal dynamics of a disaster in near real-time. Such a spatio-temporal footprint would also improve the use of social media such as Twitter as a two-way communication channel – a platform for information retrieval and alerts broadcasting.

Twitter is often considered as a "social sensor" and represents a valuable data source for emergency and disaster management [1]. It can provide important up-to-date information especially in the first minutes and hours after a disaster, when other data are not available.

However, an overreliance on social media can lead to the most heavily affected areas being overlooked because important disaster-related information might be hidden in the data shadow [2] of more densely populated but less affected

areas, which generate a higher baseline of Tweets compared to smaller and remote affected areas. Another challenge is to conduct such investigations and make decisions under time constraints, with large data, of which the characteristics are mostly unknown. To mitigate this effect, any social media data has to be contextualized with other available data sources, such as socio-economic data or critical infrastructure data. Such contextualization can rely on semantic linkage [3] or geographic contextualization [4]. Given the relatively high entry-barriers for using semantic technologies, and our target group of decision makers and relief workers who are neither data scientists nor technology specialists, we focus first on geographic contextualization as a more intuitive approach, and as a prerequisite for it, the identification of affected areas.

The guiding research question of this paper is how the combination of well-known quantitative data mining methods and qualitative visual analysis methods can help to identify potentially affected areas in a real-world scenario.

The scientific contribution of this work lies in the use of data mining techniques for data pre-processing in combination with exploratory visual analysis of Twitter messages to identify the areas affected by a flood. Different studies focused on visual social media analytics for situational awareness [5], [6]. We used these techniques as a starting point to generate a geospatial disaster footprint.

Unlike conventional spatiotemporal data, social media data is dynamic, massive, unevenly distributed in space and time, noisy, incomplete, biased in terms of population, and represented in stream of unstructured media (e.g. texts and photos), which pose fundamental challenges for representation and computation to conventional spatio-temporal analysis.

During emergencies, practitioners need to obtain a clear picture of the situation and up-to-date information. Data mining can help to synthetize and summarize data retrieved from different social network platforms and visual analytics techniques can help practitioners to refine and understand the results of such analyses.

Our objective is to identify the area affected by a flooding event, so single Tweets – here considered as spatial events (points) with an associated time reference - need to be aggregated in space and time to obtain areas that may have been or will be affected by the flood.

2 Methods

A typical work flow of social media requires extensive pre-processing of the input: removal of geographically unrelated data (e.g., a typical bounding box search criterion will include unrelated data), followed by a pattern-based removal of machine-generated data (e.g. bots or automated but unrelated status messages (e.g., FourSquare check-ins), and finally, natural language processing to mine the data and classify and annotate its content. For this initial study, we employed straightforward methods of filtering via point-in-polygon analysis to retain Tweets originating from Italy, and keyword-list filtering for topicality, which yield sufficient results (cf. Section 3.1). However, text processing of Tweets is an active field of research, and this step can principally be extended using more advanced text mining approaches.

After pre-processing, the prepared data set contains relevant data with spatial coordinates and timestamps, i.e., a set of spatio-temporal points. To identify areas within these point data sets, we applied several complementary clustering techniques to understand how affected places are represented. Since all cluster algorithms need parameterization, it is important that the analyst is able to observe and reason about the impact of parameter value selection in space and time, and to adjust these iteratively until a good cluster (in the context of the given situation) has been found. This requires tight integration between computational and visualization functionality; in our case, for all clustering operations, we used the V-Analytics toolkit (http://geoanalytics.net/, see also [7]) We distinguish three groups of clustering algorithms, which we describe next; for their concrete application and appropriate parameter settings see 3.2.

1) Density-based clustering detects densely populated regions in space-time with arbitrary shape; therefore, density-based clusters may indicate the spatial and temporal extent of a flood-related event. The number of clusters is not pre-determined and isolated points are optionally discarded as noise, therefore this method is suited for an initial overview and detection of (significant) event candidates. We used the OPTICS [8] implementation integrated in V-Analytics.

2) Distance-bounded spatio-temporal event clustering [9] is similar in that it, too, has a notion of point densities defined around any given point's neighbourhood. Unlike density-based clustering, it can be applied to time-dynamic data sets (data streams) and thus can detect emerging spatio-temporal clusters and track their evolution in real-time. This method additionally reconstructs trajectories of clusters, i.e. the evolution of the centre of a cluster's spatial footprint over time. This can help to reason about how a given complex situation evolved and, potentially, predict its future development.

3) Data-driven territory tessellation divides a territory into convex polygons of approximately equal sizes on the basis of point distribution [10]. The algorithm looks for spatial clusters of points that can be enclosed by circles with a user-chosen radius. A concentration of points having a larger size and/or complex shape will be divided into several clusters. The cluster centroids are then used to generate Voronoi polygons. In data about people's activities, cluster centroids most often indicate the foci of people's attention. Points can be aggregated in space by the tessellation and in time by user-selected time intervals (e.g. days or months). Resulting time series can be analyzed by means of, for example, partition-based clustering. From this, spatial time series are derived for each Voronoi cell to reason about temporal dynamics in different regions of geographic territory.

For the aforementioned visual inspection of intermediate results and parameter impact, various visual analysis techniques have been applied to interpret the results: space-time cube, frequency histograms, time graphs, qualitative colouring, and animated maps.

3 Case study

The case study we selected to test our methodology is a flood that took place in the Sardinia region (Italy) on 18 and 19 November 2013.

3.1 Data description and preparation

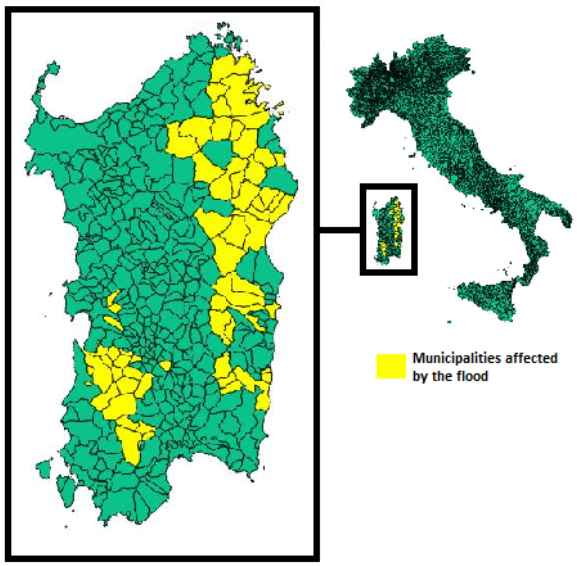
We used a set of all the georeferenced Tweets that were posted between December 2012 and April 2014 within a bounding box containing Italy, collected from the Twitter streaming API.

To extract potentially relevant flood-related Tweets, we applied a keywords-based filter to this dataset. We built a lexicon containing flood-related keywords in Italian language by generating a word cloud using a subset of the data posted during the flood occurrence, then we refined the extracted flood-related terms by reviewing and generalizing (stemming) the terms, - e.g. alluvion# to retrieve Tweets containing alluvione, alluvionato, or alluvionati – and by querying the whole dataset. The final dataset consisted of 3,000 Tweets for the month of November 2013, of which 897 tweets generated between 18/11/2013 and 20/11/2013.

Demographic data was then used to normalize the results and filter out big clusters in large cities and to better compare the analysis results. Census data give information on residential population, but do not consider movement of people in the different hours of the day, when the number of people can vary considerably and consequently also the number of Tweets. However, census data provide a sensible baseline for expected number of Tweets.

For evaluation of our results, we relied on ground truth information from official reports. The quarterly report "Relazione trimestrale 18/11/2013 – 17/02/2014" available on the website of Sardinia region[1] and "Ordinanza n. 3 del 22/11/2013"[2] contain information about the municipalities affected by the flood on 18-19 November 2013 (see Figure 1).

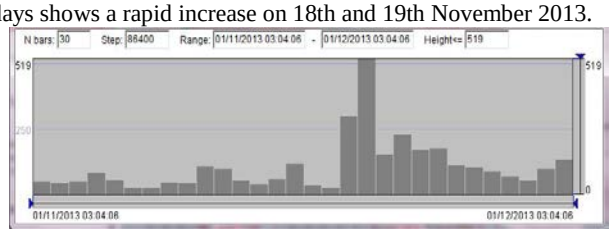
Figure 1: Map of municipalities affected by the flood event on 18-19 November 2013, elaborated from the report 'Ordinanza n 3 del 22/11/2013'.



3.2 Data exploration, visual analysis of the results and evaluation

The flood event under investigation was one of the major flooding events that occurred in Italy that year, resulting in fatalities and widespread damage across the region. However, using a random sample containing 5% of all georeferenced Tweets dataset generated during the month of November 2013, it was not possible to detect the flood event visually, because no significant increase or decrease in the number of Tweets occurred in the temporal proximity of the flood. Focusing on Tweets georeferenced in Sardinia, it was possible to note a slight increase on the number of Tweets on 18th and 19th November 2013 compared to the rest of the month. However, text cloud analysis revealed that in those days the most frequent terms contained in the messages were related to the flood, while flood-related terms were not frequent for the rest of the month. Analyses at local/regional scale are important to detect and define the scale of the event. After applying the flood-related keyword-based filter, we could visually detect a rapid increase in the Tweets frequencies during the 18th and 19th of November 2013 (Figure 2).

Figure 2: The frequency histogram of flood-related tweets by days shows a rapid increase on 18th and 19th November 2013.



We performed the clustering described in section 2. For OPTICS, after using visual analysis to iteratively adjust the parameters, we chose 10 km as distance threshold, 1 day as temporal threshold and 3 as minimum neighborhood distance threshold, obtaining 115 clusters, with 31.9% of the points classified as noise. In this analysis, clusters of Tweets appear in different areas of Sardinia region only after the 18th of November, while in other cities like Milan and Rome clusters have a temporal distribution over the entire period of interest (November 2013) (Figure 3). The appearance of clusters in some areas of Sardinia region may indicate that the event occurred in those areas (as indicated in Figure 4). Smaller clusters appear also in other Italian cities on November 18th: Tweets' content analysis revealed they relate to the weather event occurring in Sardinia.

Also for the distance-bounded event clustering we used visualization techniques to iteratively adjust the parameters and we chose 10 km as maximum event spatial distance, 6 h as maximum time distance between events, and 5 as minimum number of event in a group and real-time simulation. Results indicate some areas in Sardinia region as affected by the flood event (Figure 5).

1 http://www.regione.sardegna.it/documenti/1_231_20140403083225.pdf

2 http://www.regione.sardegna.it/documenti/1_406_20131123142053.pdf

Figure 3: Results of spatio-temporal density-based clustering using OPTICS visualized in a space-time cube.[3]

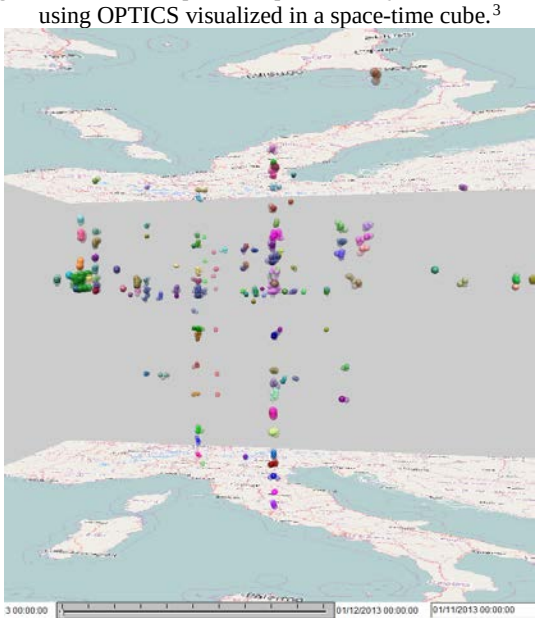
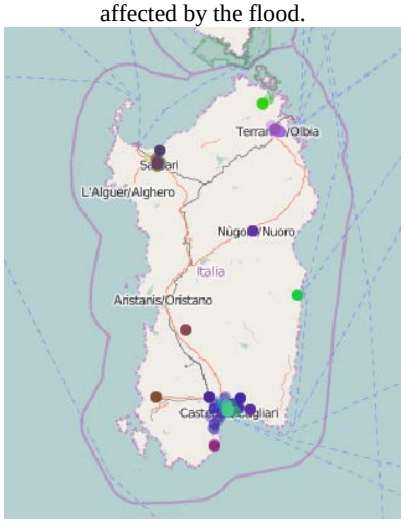


Figure 4: Results of spatio-temporal density-based clustering using OPTICS for Sardinia. Clusters identify potential areas affected by the flood.



Comparing Figures 4 and 5, it is visible that both clustering techniques resulted in identifying similar areas potentially affected by the flood.

In addition, we aggregated the data in space using the data-driven territory tessellation - obtaining Voronoi polygons around groups of points using maximus group radius of 25 km – and in time using 1-day time interval. From the time series graph of tweets frequencies, it was possible to identify which areas (corresponding to the Voronoi polygons) were potentially affected (Figure 6). This method found less possibly affected areas compared with the previous two.

Figure 5: Results of distance-bounded event clustering for the time period 18-20 November 2013: clusters identify areas potentially affected by the flood.

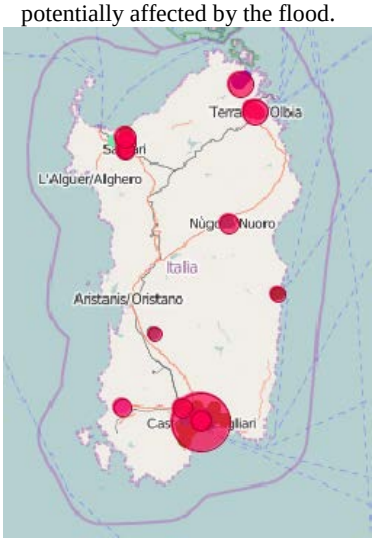
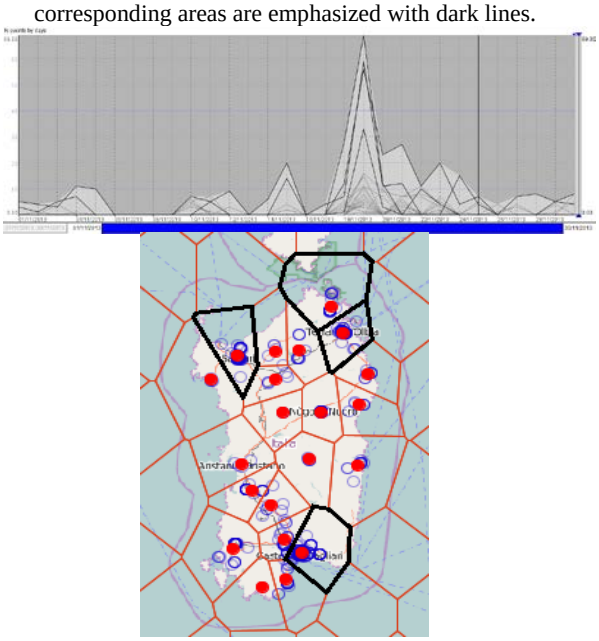


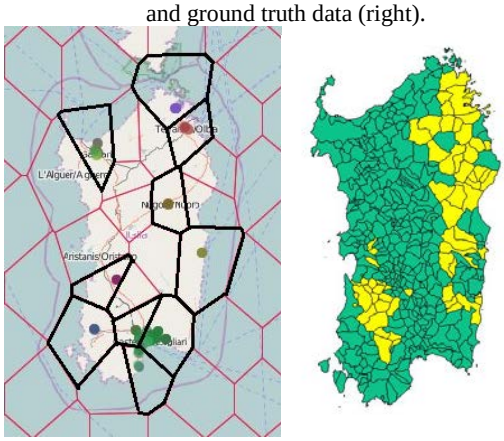
Figure 6: Time series graph of Tweet frequencies and results of the data-driven territory tessellation. Each line represents the number of tweets per daily interval for one Voronoi cell of the spatial tessellation. The highest frequencies and their corresponding areas are emphasized with dark lines.



To better establish affected areas we combined the results of clustering methods with the spatial tessellation, and considered as affected those areas corresponding to Voronoi polygons in which clusters reside (see Figure *7*, left image).

3 Figures 3 to 7: Basemap OpenStreetMap © OpenStreetMap contributors, http://www.openstreetmap.org/copyright

Figure 7: Comparison of results of clustering methods (left) and ground truth data (right).



The comparison between the results of the clustering methods and the ground truth data shows how our analysis performed well in the identification of affected areas, resulting in few false negative. False positives appear in areas nearby Cagliari (south of Sardinia) and Sassari (north-west of Sardinia), classified as affected while they were not. Analysts can explain or remove false positive by contextualizing the results. According to official reports (see Section 3.1), these cities played an active role in the emergency management during the flood, so this can explain the biases, and points at the benefits of future work for better contextualizing the data and the analysis.

4 Conclusions and future work

In this paper, we presented the first steps towards an approach to define the geospatial footprint of a flood event using georeferenced Tweets. We used data mining techniques for data pre-processing in combination with exploratory visual analysis of Twitter messages to identify the areas affected by a flood. The proposed methods showed good results in the identification of areas affected by the flood at regional scale and the combination of data mining with visual analysis for parameter setting proved to be an intuitive and fast procedure that could help decision makers deal with Twitter data and assist them for rapid assessment of the situation. This analysis took on average 4 iterations to set optimal parameters and was performed in less than two hours; it reveals to be promising for meeting users' needs, but a users' evaluation is required. With a properly tested and documented procedure, the analysis time can be reduced significantly.

To obtain more precise footprints, further analyses are required. False positives are an issue that will be addressed in future research. Future work will also take into account the extraction of locations (and other potentially useful information) mentioned in (non-georeferenced) social media and will compare locations of origin (defined by the geotag) and content, in order to obtain more detailed footprints. Additional useful local information that can increase situational awareness during emergency situations will be derived from analysis of social media contents and authoritative data.

References

[1] M. F. Goodchild and J. A. Glennon. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.

[2] T. Shelton, A. Poorthuis, M. Graham, and M. Zook. Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum*, 52:167–179, 2014.

[3] L. M. Vilches-Blázquez, B. Villazón-Terrazas, O. Corcho, and A. Gómez-Pérez. Integrating geographical information in the Linked Digital Earth. *International Journal of Digital Earth*, 7(7): 554–575, 2014.

[4] L. Spinsanti and F. Ostermann. Automated geographic context analysis for volunteered information. *Applied Geography*, 43: 36–44, 2013.

[5] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl. ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions On Visualization And Computer Graphics*, 19 (12): 2022–2031, 2013.

[6] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. SensePlace2: GeoTwitter analytics support for situational awareness. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 181–190, 2011.

[7] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual Analytics of Movement*. Springer, 2013.

[8] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander. OPTICS: Ordering Points To Identify the Clustering Structure. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, 28: 49–60, 1999.

[9] N. Andrienko, G. Andrienko, G. Fuchs, and S. Rinzivillo. Detection, Tracking, and Visualization of Spatial Event Clusters for Real Time Monitoring. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015.

[10] N. Andrienko and G. Andrienko. Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics*, 17(2): 205–219, 2011.