Applied Geography 43 (2013) 147–158

Contents lists available at SciVerse ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Web data mining for automatic inventory of geohazards at national scale☆

Alessandro Battistini*, Samuele Segoni [1], Goffredo Manzo [1], Filippo Catani [2], Nicola Casagli [3]

Department of Earth Sciences, University of Firenze, Via La Pira n°4, 50121 Firenze, Italy

Keywords:
Geotagging
News
Geodatabase
Geohazards
Inventories
WebGIS

A B S T R A C T

In this study, we present a fully automated procedure to analyze online news using data mining techniques. It is then used to compile and continually update a geohazard database. The procedure is based on new technologies that publish news on the internet, i.e., the news is analyzed, georeferenced and attributed to a category of geohazards (the current categories are landslides, floods and earthquakes). A continuous flow of georeferenced events is established to populate and update the geodatabase automatically and in near-real time. We tested the procedure for 2 years at a national scale, creating a geodatabase containing more than 20,000 news items concerning geohazards that occurred in Italy. This procedure enables continuous feedback from events in the real world, such that information about geohazards can be fully exploited rapidly (compared to traditional techniques based on remote sensing, field surveys and historical inventories).

Introduction

Geohazards, including earthquakes, floods and landslides, are a common cause of casualties and economical losses (Petley, 2012; Wirtz, Kron, Löw, & Steuer, 2012); therefore, significant efforts are always ongoing to mitigate related risks. When studying geohazards, especially at a regional or a national scale, the availability of archives and geodatabases (Papagiannaki, Lagouvardos, & Kotroni, 2013; Wirz et al., 2012), which can provide information about past and recent events (including magnitude, timing and location), is of primary importance.

In particular, the availability of updated and complete inventories is of paramount importance for hazard and risk assessment (Hilker, Badoux, & Hegg, 2009; Van Westen, Castellanos, & Kuriakose, 2009) and for the creation of early warning models (Brázdil, Kundzewicz, & Benito, 2006; Martelloni, Segoni, Fanti, & Catani, 2012; Montesarchio, Ridolfi, Russo, & Napolitano, 2011; Rossi, Catani, Leoni, Segoni, & Tofani, 2013). Unfortunately, one of the primary limitations of the existing archives and databases (especially for landslides and floods) is their update rate and methodology. They are usually compiled manually using field surveys (Brunsden, 1985; Santangelo, Cardinali, Rossi, Mondini, & Guzzetti, 2010) and, sometimes, remote sensing (Lu, Casagli, Catani, & Tofani, 2012; McKean & Roering, 2003; Segoni et al., 2009; Soeters & Van Westen, 1996). Systems using automated or real-time updates are still uncommon and only used for a few types of geohazards.

An earthquake is the geohazard that can rely on the most effective and fast geolocalization and characterization methods. A worldwide network of sensors and processing stations exists that is able to record and localize the occurrence of major and medium events, even at the global scale in near-real time (Wald et al., 2008). In addition, several national agencies provide the same information at a national scale, including minor events, in real time. Global and national scale information is easily accessible on the Internet or via specific applications that also provide notification services (Satriano, Wu, Zollo, & Kanamori, 2011; Wald et al., 2008; Wetmiller, Adams, & Woodgold, 2007).

* Corresponding author. Tel.: +39 055 275 7548; fax: +39 055 275 6296.
E-mail addresses: alessandro.battistini@gmail.com (A. Battistini), samuele.segoni@unifi.it (S. Segoni), goffredo.manzo@gmail.com (G. Manzo), filippo.catani@unifi.it (F. Catani), nicola.casagli@unifi.it (N. Casagli).
[1] Tel.: +39 055 275 7551; fax: +39 055 275 6296.
[2] Tel.: +39 055 275 7559; fax: +39 055 275 6296.
[3] Tel.: +39 055 275 7523; fax: +39 055 275 6296.

0143-6228/$ – see front matter © 2013 The Authors. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.apgeog.2013.06.012

A flood is a geohazard with an overwhelming state of the art (e.g., hydraulic studies performed on the Arno river (Italy) date back at least to the 12th century, as reported by Morelli, Segoni, Manzo, Ermini, & Catani, 2012). In addition, the location of the primary elements at risk and sources of hazards are usually well-known and documented (Cummings, Todhunter, & Rundquist, 2012; Hilker et al., 2009; Morelli et al., 2012; Salvati, Bianchi, Rossi, & Guzzetti, 2010; Varis, Kummu, & Salmivaara, 2012). Despite that, the study of floods and hydraulic risk in general requires the use of long time series (Ahilan, O'Sullivan, & Bruen, 2012; Gräler et al., 2013; Ricci, Piacentini, Thual, Le Pape, & Jonville, 2011). Most countries rely on a number of gage stations able to monitor water levels and fluvial discharges with high accuracy. Many national or regional hydrological services have been tracking these data for decades, or even centuries, providing the data for scientific study (Alfieri et al., 2013; Böhm & Wetzel, 2006; Brázdil et al., 2006; Schmocker-Fackel and Naef, 2010; Silva, Portela, & Naghettini, 2012; Zhou, Persaud, & Wang, 2006). In many countries, it is possible to retrieve and use detailed descriptions of flood events (at the basin or even on a national scale) that occurred prior to the establishment of runoff measurements (Barriendos et al., 2003; Barriendos & Rodrigo, 2006; Schmocker-Fackel & Naef, 2010; Thorndycraft et al., 2005).

Creating complete and updated databases is more difficult for landslides (Galli, Ardizzone, Cardinali, Guzzetti, & Reichenbach, 2008; Roth, 1983; Santangelo et al., 2010). For this geohazard, efforts are required for the development of models and their application, and efforts are also required for the collection of complete data sets to be used in the calibration and validation process. Therefore, the detection and mapping of landslides can often be the primary goal of a single research study (Tofani, Segoni, Agostini, Catani, & Casagli, 2013, and references therein). However, several global landslide databases are currently operative (GSC, 2011; Kirschbaum et al., 2009; Petley, 2012; USGS, 2011). Although they are useful tools for estimating the hazard and its impact on society, they are characterized by a significant degree of incompleteness because they primarily contain major events with catastrophic effects. At a national scale, several existing archives and geodatabases may have more detail and thus a higher degree of completeness and homogeneity (Dikau, Cavallin, & Jager, 1996; Hilker et al., 2009; Salvati et al., 2010; Trigila, Iadanza, & Spizzichino, 2010; Van Den Eeckhaut & Hervás, 2012). These instruments, although very useful for susceptibility assessments (Van Den Eeckhaut & Hervás, 2012), present some drawbacks, which prevent wider applications in landslide studies, i.e., they are not continually updated and seldom provide systematic information about the landslide time of occurrence (thus they cannot be useful for the temporal calibration/validation of forecasting models). The collection of recent landslide data can be a very demanding task, even at detailed scales, because it is usually accomplished using field surveys (Brunsden, 1985; Santangelo et al., 2010), remote sensing techniques (Lu et al., 2012; McKean & Roering, 2003; Segoni et al., 2009; Soeters & Van Westen, 1996) or manual retrieval of information from newspapers and technical reports (Brunetti et al., 2010; Hilker et al., 2009; Lagomarsino, Segoni, Fanti, & Catani, 2013; Rosi, Segoni, Catani, & Casagli, 2012), thus demanding a significant amount of time and human resources.

The primary goal of this study is to address this need with a data mining technique that automatically retrieves geohazard events from the Internet and organizes them in a continually self-updating geodatabase. The proposed approach is based on the concept that whenever a geohazard event produces a noticeable effect, the news is reported on the Internet. Therefore, Internet news can be considered a continuous stream of information, and appropriate data mining techniques can be set up to separate non-relevant information from pertinent items. Once the latter are identified by an automated procedure, each single event can be analyzed and stored in a geodatabase, together with related information (including a reference location and the time of occurrence). The proposed data mining procedure retrieves news in really simple syndication (RSS) format and analyzes it to identify the geohazard typology and its time of occurrence. In addition, the comparison with a geodatabase of toponyms is used to assess the spatial location of the events. The data mining procedure uses algorithms that are specifically calibrated for a single geohazard typology.

We tested the proposed data mining technique at the national scale, using Italy (approximately 300,000 km^2) as the test site and landslide as the primary target geohazard. After a test period of 2 years (November 2009–November 2011), a national scale geodatabase of approximately 5500 dated and georeferenced landslides was created automatically. These data can be used to update existing geodatabases or to calibrate or validate landslide forecasting models. This study focused on landslides because they represent the geohazard for which the creation of a model is most needed; our results could thus be more immediately and directly exploited. We also include the results of preliminary applications to earthquakes and floods to show that the proposed data mining technique can be applied (with appropriate specific adjustments) to almost any type of geohazard and to all news categories.

Materials and methods

Most of the news distributed via telematic channels is currently structured in RSS format (RSS Advisor Board, 2009) or in its analog, Atom (Nottingham & Sayre, 2005). RSS and Atom are XML-based formats that establish the distribution of documents via the web. They categorize the news and ensure that appropriate software can utilize news from different sources. This software, called a feed aggregator or feed reader, distributes news in a standardized format and in a unique environment. Actually, a feed is a summary of web news content along with links to their full versions. By subscribing to a particular web feed, it is possible to receive constantly updated summaries of the news collected from the site and, filtering them, to obtain only topics of interest (e.g., news about landslides and floods).

Unfortunately, the information structured in the RSS or Atom format does not have a native geographical location; the news itself is not associated with any structured geo-location. The RSS format was expanded in 2007, in its georeferenced version, adding a suitable geotag, GeoRSS (Geographically encoded objects for RSS feeds). According to Reed (2006), GeoRSS is a simple method that attaches a location, or "geotag", to an RSS feed. GeoRSS standardizes the way in which "where" is encoded with enough simplicity and descriptive power to satisfy the majority of needs and to describe the location of the Web content. With the geotag, a news item gains a reference to a specific location (area or geographical point) defined in the WGS84 coordinate system. News in this format is easily collected to be used in our database. Unfortunately, the GeoRSS format is not established as a standard, so networks rarely provide such information. In particular, in Italy and in the Italian language (in which this work was performed), it is not possible to find a news bulletin able to distribute news in the GeoRSS format.

The only way to obtain geolocalized news is to use an RSS-to-GeoRSS conversion service. This service, using data-mining procedures, compares the words in the news item to a database of locations to search for terms that may have a geographical location. The primary application (GeoNames, http://www.geonames.org) currently available online uses a limited set of locations (a few thousand for the entire country of Italy), resulting in poor localization accuracy.

Table 1
News bulletin categories.

Categories	Description
Politics	Topic on politics, law, religion, economy and finance
Entertainment	Topic on entertainment, TV, cinema, music and gossip
Culture	Topic on culture, art, books, universities
Trend	Topic on trend and fashion for women, men, youth, children
Technology	Topic on technology, science, energy, IT and other specialized areas
Environment	Topic on environment, nature, animals and ecology
Health	Topic on health, nutrition, wellness and gastronomy
Hobby	Topic on hobbies, travel, leisure, collecting
House	Topic on construction, architecture
Sports	Topic on sports, cars, racings, boats
Territory	Topic on territory
Meteo	Topic on meteorology
Foreign	Topic on facts happened abroad
None	Not classified

News should be researched using the local language; otherwise, only major events of international relevance would be acquired. Because our study case is the entire country of Italy, the study is performed in Italian.

GEOITALY DB

In this study, the RSS-to-GeoRSS conversion is performed with the help of some detailed Italian place names databases (DBs) available for Italy.

Merging all of the DBs listed in Table 2 results in a database called GEOITALY DB. Table 3 reports the categories of names that can be identified in GEOITALY DB.

Each toponym is geolocalized in the WGS84 geographic coordinate system. A geometric component is also stored as a linear, point or areal feature, depending on the characteristics of the toponym. We use the PostGIS open source DB for the DBMS, which was chosen for its ability to exploit the spatial relationships between toponyms and to integrate with the geographic information system (GIS).

To improve the efficiency of the geotagging process, we introduce hierarchical relationship among place names. Finding hierarchically dependent place names in news items strengthens the probability of correct matching. The most obvious hierarchy for the Italy is the classical administrative subdivision Regions-Provinces-Municipalities. Unfortunately, this chain alone is not sufficient for news geotagging. Not all local newspapers follow the same logic, and some preliminary tests using only this hierarchical chain resulted in low spatial accuracy. To overcome these drawbacks, GEOITALY has been expanded with additional geographic hierarchical chains, such as historical regions, valleys, mountain communities and tourist areas. Terms related to such subdivisions are often reported by journalists and can be used as additional clues to the location of the news item; therefore, a directory of place names, such as “Irpinia”, “Riviera Romagnola” and “Chianti”, was added. The relationship between the municipality and these geographical areas is not mutually exclusive, as it is in the administrative relationship.

Table 3
Geoltaly toponyms.

Features	Numbers
Regions	20
Provinces	110
Municipalities	8100
Geographical areas	203
Sub-municipal localities	≈770,000
Rivers	6239
Lakes	438
Mountains	1838
Road toponyms	≈570,000
Foreign place names	>1000

Local news bulletins tend to assume that the reader is aware of the general location of the news. Therefore, they often neglect to explicitly mention the name of the place where the event occurs. To avoid missing the geotagging of some local news, a table with the localization of every online bulletin that publishes a news item is added to DB GEOITALY. This table is updated every time a new online bulletin is detected by the system. At present, approximately 2000 bulletins are stored in GEOITALY. Each bulletin is associated with 1–3 areas of interest. The areas can be a province, a region, a municipality, a state, or a geographical area. These areas are arranged from the most specific to the most global.

Because the system gathers information from many different news bulletins, a *Focus* field is used to indicate the primary topic of the news bulletin. This facilitates the removal of undesired news. Currently, the focus of a news bulletin can be assigned to one of the categories in Table 1.

GEOITALY DB has a triple hierarchical structure (Fig. 1), in which the municipality is the smallest geometric entity and is stored as a polygonal entity. A municipality may be part of several higher-order polygonal entities, i.e., administrative areas (region and province), geographical areas (zones) or information areas (news bulletins). Recognizing any of these higher entities within the news increases the possibility that the identified municipality is the actual location.

The objective of the news geotagging process is the identification of the municipality where a geohazard has occurred. To assist the geotagging process, we include approximately 770,000 toponyms, derived from the official Italian Military Geographic Institute (IGM) maps at the 1:25,000 scale, in the GEOITALY DB. Each of these localities is characterized by WGS84 geographical coordinates and has a name that represents a neighborhood, a district, an area, a small village, or any other site that can be contained within a

Table 2
Source of databases for GeoItaly DB.

Source of database	Content
IGMI – Italian Geographic Military Institute	DB containing all the place names in 1:25.000 map, subdivided into categories (Urban, Hydrography, Roads and Orography)
ISTAT (2001) – Italian National Institute of Statistics	DB containing Italian administrative subdivision (regions, provinces and municipalities with hierarchical relationship between them)
ISTAT (1991) – Italian National Institute of Statistics	DB containing a list of smaller towns, with the municipality and the province to which they belong;
MATTM – Italian Ministry of the Environment	DB containing the features of regions, provinces and municipalities;
PT – Italian Post	DB of ZIP postal code, containing areas and names of the Italian zip codes, with the province
APAT – Agency for Environmental Protection	DB containing the vector digitization of Italian rivers, lakes and canals at 1:250.000 scale

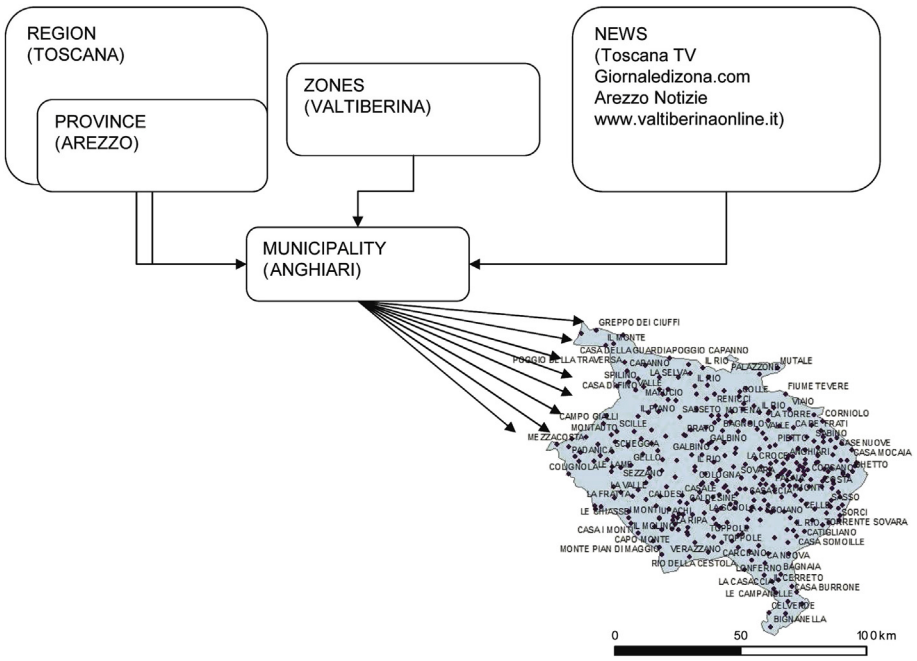

Fig. 1. Sample of the municipality of Anghiari stored in GeoItaly, its hierarchical chain (administrative, zonal and bulletins) and sub-municipality locations.

municipality. Locating one of these toponyms in the news can ensure a more accurate location of the news item because it confirms the item's relation to the municipality.

If the municipality cannot be determined because of a lack of information in the text, the news item can be referenced to an administrative area larger than the municipality (e.g., a province or a region), to a geographical area, or to a physiographic element (e.g., a river or a mountain). In this case, sub-municipality localities in the news are used to confirm the localization.

The database GEOITALY is a detailed database for the Italian territory, and it was used to create a geotagging custom service. The database GEOITALY contains other tables used by the geotagging process to change the scores (see below) given to every toponym found in the news.

Some local place names have other meanings in Italian; this is an inconvenience that may deceive the geotagging algorithm. For example, "Parete" is a town of Southern Italy but also means "wall". A table collecting approximately 500 potentially misleading place names was created to reduce the probability of this type of error.

In GEOITALY DB, all toponyms are stored with their official names. Journalistic and common names are often different from official names. Places are often referred to by the common version known by the readers (e.g., "Castel Santangelo" instead of "Castel Sant'Angelo", "Jesolo" instead of "Iesolo", or "Bozen" instead of "Bolzano"). An equivalent toponym solves this problem. Another table is used to store the adjective form used to describe the territory or the population of the primary population centers (e.g., Roman is equivalent to Rome, Partenopeo is equivalent to Naples).

System architecture

The news searching, geotagging and storing in the DB is performed with a 15-min update rate. The task of creating event maps or other information necessary for various applications is deferred to client computers interfaced with the computer containing the news DB. Fig. 2 illustrates this architecture.

The input to the geotagging process is an RSS news feed and the output is a GeoRSS news feed. The GeoRSS output contains geographical tags, which are necessary for locating the news items. A main program (*newsProcess.php*) connects to the Internet, retrieves the RSS news items, and geotags and catalogs the items in a database called NOTIZIE (Italian name for NEWS). The Internet is accessed every 15 min to ensure a continuous flow of information as the news populates the web. The program *newsProcess.php* is subdivided into 3 sub-modules, each providing the functions needed for a particular process:

- Acquisition: *newsSearch.php* script;
- Geotagging: *newsGeoLocate.php* script;
- Cataloging: *newsStore.php* script;

Acquiring news

The script *newsSearch.php* browses the web in search of event news. Basically, the program is a feed reader and collects RSS news from Google News, which serves as an aggregator of news from online networks. The script acquires news using classes provided by the SimplePie project. SimplePie is a very fast and easy-to-use feed parser (http://simplepie.org/). The entire block of feed obtained from Google News is managed as an array.

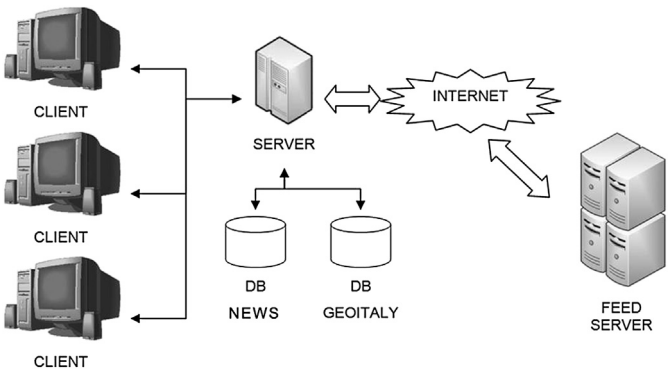

Fig. 2. System architecture.

Using the pre-processed feed from Google News provides several advantages:

- It can be queried via a web browser or used as a web service. Therefore, it can easily be integrated into a feed reader program;
- It provides feeds either in RSS or Atom format. RSS is chosen for this study because of the potential GeoTag information embedded in the news;
- There is no need for additional research for online news bulletins. Google claims to collect news from over 250 Italian sources, and we have found more than 400 different news bulletins in the query results.

The news is pre-classified by the Google News system before sending the response. Google News provides a grouping that accounts for several factors, i.e., title, text and time of publication of the item. It uses clustering algorithms to identify items that are related. With this technology, various news items concerning the same event can be stored in a single feed, significantly reducing the required DB storage space. For each of the multiple news items grouped into a single event, the system estimates the number of items that are grouped together. This number is then used as an index of relevance, indicating how much the news has been echoed in bulletins. Important news, which had been lost for some reason, is grouped with others as long as the news media echo it on the web.

The technology provided free of charge from Google News provides a powerful system for directly filtering news online. The Google News URL for feeds in Italian is http://news.google.it. A basic call to the service is implemented as follows:

http://news.google.it/news?q=*string*

where *q* parameter indicates the searched string in the text of the news. A combination of strings can also be searched using the logical operators AND and OR as needed.

A more specific call, used to obtain geohazard news, is implemented as follows:

http://news.google.it/news/search?&ned=it&output=rss&as_nloc=it&q=*string*

where the following optional arguments are used:

- *ned*: this parameter indicates which edition of news is searched, and *it* is used to specify news in Italian;
- *output*: this parameter indicates how the result should be provided, and *rss* indicates that the output should be in RSS format;
- *as_nloc*: this parameter is used to search news from a particular country, and *it* specifies that only pages of web sites registered in Italy are relevant;

The research presented in this article is primarily based on the argument “landslides”. The “floods” and “earthquakes” arguments were also researched for experimental purposes to provide a statistically more consistent total number of news items. The automatic search for news is performed by combining the primary words (synonyms in Italian) that may be related to each argument with the logical operator *OR*. Every word is searched in both singular and plural forms as follows:

- Italian landslide synonyms and related words: frana, frane, smottamento, smottamenti, franamento, franamenti, scivolamento and scivolamenti;
- Italian flood synonyms and related words: alluvione, alluvioni, straripamento, straripamenti, inondazione, inondazioni, allagamento and allagamenti; and
- Italian earthquake synonyms and related words: terremoto, terremoti, sisma, sismico, sismica, tellurica and telluriche.

For each news item, the following important information is derived directly from the RSS format:

- Id: unique identifier of the news item;
- Title: title of the news item;
- Content: content of the news item (HTML format);
- Description: summary of the news item (HTML format);
- Time: date and time of publication of the news item;
- Permalink: web link to the news item or news group in Google News;
- Lat: latitude of the position of the news event if it is in GeoRSS format; and
- Lon: longitude of the position of the news event if it is in GeoRSS format.

The *description* is only a summary containing the first few lines of the news item. The *content* is the entire HTML code that represents the news item in the bulletin. Unfortunately, several feed aggregators, including Google News, do not provide the *content*. In such cases, the *SimplePie* reading procedure duplicates the description. As a consequence, using Google News, *description* and *content* are the same. *Lat* and *Lon* are acquired because the SimplePie class is also designed to handle feeds distributed in the GeoRSS format. The process acquires the coordinates of the point from the GeoTag if these coordinates are present. The geotagging is performed only if *Lat* and *Lon* are not assigned (100% of the news registered during the test period).

Some useful information is not explicitly expressed in RSS but can be extracted from the standardized description provided by Google News. Therefore, some automatic search functions were developed to analyze the HTML code in the *description* of multiple news items, allowing the following information to be learned:

- Main title;
- Main news bulletin;
- Main article text;
- Titles of other news items; and
- Other bulletins with news on the same subject.

In the case of grouped news, Google News also counts how many news items are equivalent and aggregated. The process stores this number.

Geotagging news

After the acquisition process described in the previous section, the *newsGeoLocate.php* script starts geotagging the acquired news. The first action is to check whether the bulletin is a local bulletin, which can provide a clue to the approximate location of the event. The script begins data mining for the news *description*. The data-mining process searches for any of the Italian toponyms stored in GEOITALY DB.

After one or more place names are detected, the process geotags the news item based on the toponym that most likely indicates the location of the event. If more than one toponym is found, each one is assigned a *toponym score,* which is calculated according to several factors, such as nearby words in the same sentence, presence of capital letters, position in the sentence, position in the text (e.g., in the title), articles or prepositions introducing it and number of times the toponym appears in the sentence. A series of additional specific checks are then performed, i.e., a test about possible alternative meanings of the toponym and a test that the toponym is

not a person name or surname. If these tests are passed, the integrated GEOITALY DB is used to adjust the score, e.g., the presence of hierarchical geographical chains strengthens the score of the toponym at the highest level; similarly the score is increased if the toponym is located in the competence area of the local news bulletin that is reporting the news. Theoretically, the scores may vary without limits. In practice, more than 95% of the scores are between 0 and +300.

Toponym scores below 40 are interpreted as insufficiently reliable for geotagging. In this case, the score is set at 0 to avoid further processing, and the news item is assigned to an "unknown location".

Once all of the place names have been processed, they are ranked based on their *toponym scores*. The place name with the greatest value is selected as the reference place name of the news item. In the case where multiple place names have comparable high scores (a difference of 20 or less) and belong to the same geographic hierarchical chain, the one corresponding to the highest level is selected to provide a more accurate location.

Once the reference toponym has been identified, the script starts geotagging the news items. The GeoTagging is done using WGS84 geographic coordinates for point location. In the early stages of the research, the coordinates from the geocoding service offered by Google Maps were used, i.e., the script makes a call to the service, sets it for geocoding in Italy and passes the reference toponym. Unfortunately, this system was running slowly and was unstable when working with acquisition rates of 15 min. This occurred because of the double passage in the Internet channel, i.e., one pass to acquire the news feed and a second pass to acquire the toponym location. In addition, some differences with place names stored in GEOITALY DB thwarted the geotagging of some news. Therefore, the script was modified to obtain the geographical coordinates of the place name directly from the GEOITALY DB.

Currently, only the coordinates (latitude and longitude in the WGS84 system) are assigned to the GeoTag, regardless of the type of the place name or the geometry with which it is stored in the GeoItaly DB, i.e., for linear entities (rivers) or polygonal entities (municipalities, provinces, regions, geographical areas and lakes), the coordinates of the centroid are used (*ST_PointOnSurface* PostGIS function).

Cataloging news

After geotagging, the news items are in GeoRSS format. The script *newsStore.php* stores the news in another PostGIS DBMS database, NOTIZIE. The following information is stored:

- Latitude and longitude of the news item's geotag;
- Toponym: the primary toponym identified for the news item; and
- Location and place type.

The latter is a more immediately understandable string than the geographic coordinates. It includes the name and the typology of the place pertaining to the GEOITALY DB that was used to geotag the news (e.g., "Province of Florence", or "Municipality of Rome").

If the geotagging process does not find any toponym with a score higher than 40, placeType is set to UNKNOWN to filter out extremely uncertain bulletin/toponym associations. There are also 2 other possible values for the parameter placeType: NO PLACE and FOREIGN. The first is used when the news has no particular place of reference (e.g., a national law on flooding), and the latter is used when the event in the news item is located in a country other than Italy. Unknown, No Place and Foreign news can be manually or automatically filtered out.

A *placeScore* value is assigned to evaluate the reliability of the geotag. The *toponym score* identified during the geotagging process is used as a base score. If the news includes other toponyms belonging to other hierarchical chains with a comparable probability (lower by 20 or less), the score is lowered by 10. Otherwise, it is incremented by 10. In cases of manual geotagging, or if the news was georeferenced from the source, *placeScore* is automatically set to 300. If placeType is NO PLACE or FOREIGN, *placeScore* is set to 0.

Additional data-mining processes are performed on the text to assign other scores as follows:

- *eventScore*: index of the probability that the news item actually concerns the topic event;
- *timeScore*: estimated days between the time of occurrence of the event (as reported in the news) and the time of publication of the news item; and
- *newsNumber*: number of equivalent news items on the web, which can be used as an index of the importance of the news.

These data are useful for the classification of the information and for future applications (e.g., landslides with accurate timing and location can be used to validate forecasting models).

EventScore estimates the probability that a specific news item refers to the type of event that is the object of the search (e.g., landslides). With this score, any news that is not related to the events of interest, though it contains words of interest, can be filtered out (e.g., the Italian expression "frana in amore" contains the word "frana", which is the Italian for "landslide", but in gossip it is used as a metaphor meaning "to be hopelessly in love"). To calculate this score, an evaluation is performed after searching for certain categories of strings. Their presence or absence raises or lowers the score by a value depending on the type and length of the word found. For example, in a news item concerning a landslide, there is a high probability of finding words such as "collapse", "risk", "danger", "evacuation", "houses", "isolated", "street" and "viability", as well as the derived versions, such as "collapses", "evacuated" and "dangerous". Other strings, similarly, lower the score because they are normally used in other topics (e.g., sports, gossip, financial, political or administrative) and are rarely found in news concerning geohazards. Currently, the strings that change the *eventScore* index are grouped into the following categories: Flood, Earthquake, Politics, Weather, Landslide, Emotions, Meteorology, Entertainment, Economy, Damage, Eruption, Earth, Sport, Social events, Danger, Anthropic and Water. The words are stored in the NOTIZIE DB.

Another data-mining process determines whether the date of publication is near or far in time from the reported event (***timeScore***). The event could be recent, in the past (e.g., an anniversary) or in the future (e.g., scheduled or forecasted events). This is accomplished by searching for strings such as "two days ago", "in December", "in May of 2008", "last year Christmas" and "last week". The process uses *regular expressions*, a concise and flexible means to recognize strings of text. *timeScore* is an integer that estimates the lag, in days, between the date of the event and the date of publication. Negative and positive numbers refer to past and future events, respectively; 0 refers to an event occurring on the same day as its publication.

Another data-mining process estimates the importance of the event. The mediatic impact is an indirect reflection of the severity of the geohazard (a severe event is more widely and persistently reported in the news than an ordinary event). Without more reliable parameters, the number of equivalent news items estimated by Google News is found to be a good indicator. A ***newsNumber*** variable is therefore used to store this information.

Before the cataloging process, an additional check prevents double cataloging. The check is performed on 4 levels, i.e., the Id, the title, the Permalink and 75% of the *Content*. If these parameters are already stored in NOTIZIE DB, the news item is updated only if the parameters are changed (e.g., the accuracy, relevance or the equivalent number of news items is increased). This updating may happen when Google News changes the news grouping, and the process can thus obtain better results.

Classifying and displaying news

A WebGIS management system was developed to visualize the geotagged news (Fig. 3). The display is influenced by the definition of some classes for each stored score (*placeScore*, *eventScore*, *newsNumber* and *timeScore*). The classification thresholds used in this study are empirically established after a test period of 2 years of continuous acquisition. They are summarized in Table 4.

The position of the news items and their score classes can be viewed using WebGIS. It is possible to filter news based on time, type or precision. If necessary, the news items can be manually

Table 4
Score thresholds.

Score	Classes	Significance	Suitable for geohazard database
placeScore	placeScore < 60	Approximate location	no
	placeScore ≥ 60	Accurate geotagging	yes
eventScore	eventScore < 40	Not inherent	no
	$40 \leq$ eventScore < 60	Approximate inherence	no
	$60 \leq$ eventScore < 90	Discrete inherence	yes
	eventScore ≥ 90	Good inherence	yes
timeScore	timeScore < -30 days	Too past event	no
	-30days $\geq$ timeScore ≥ 0	Current news	yes
	timeScore > 0	Future event	no
newsNumber	NewsNumber $= 1$	Minor local event	yes
	NewsNumber $= 2$	Local event	yes
	$3 \leq$ newsNumber ≤ 20	Relevant event	yes
	newsNumber > 20	Major event with high national or international relevance	yes

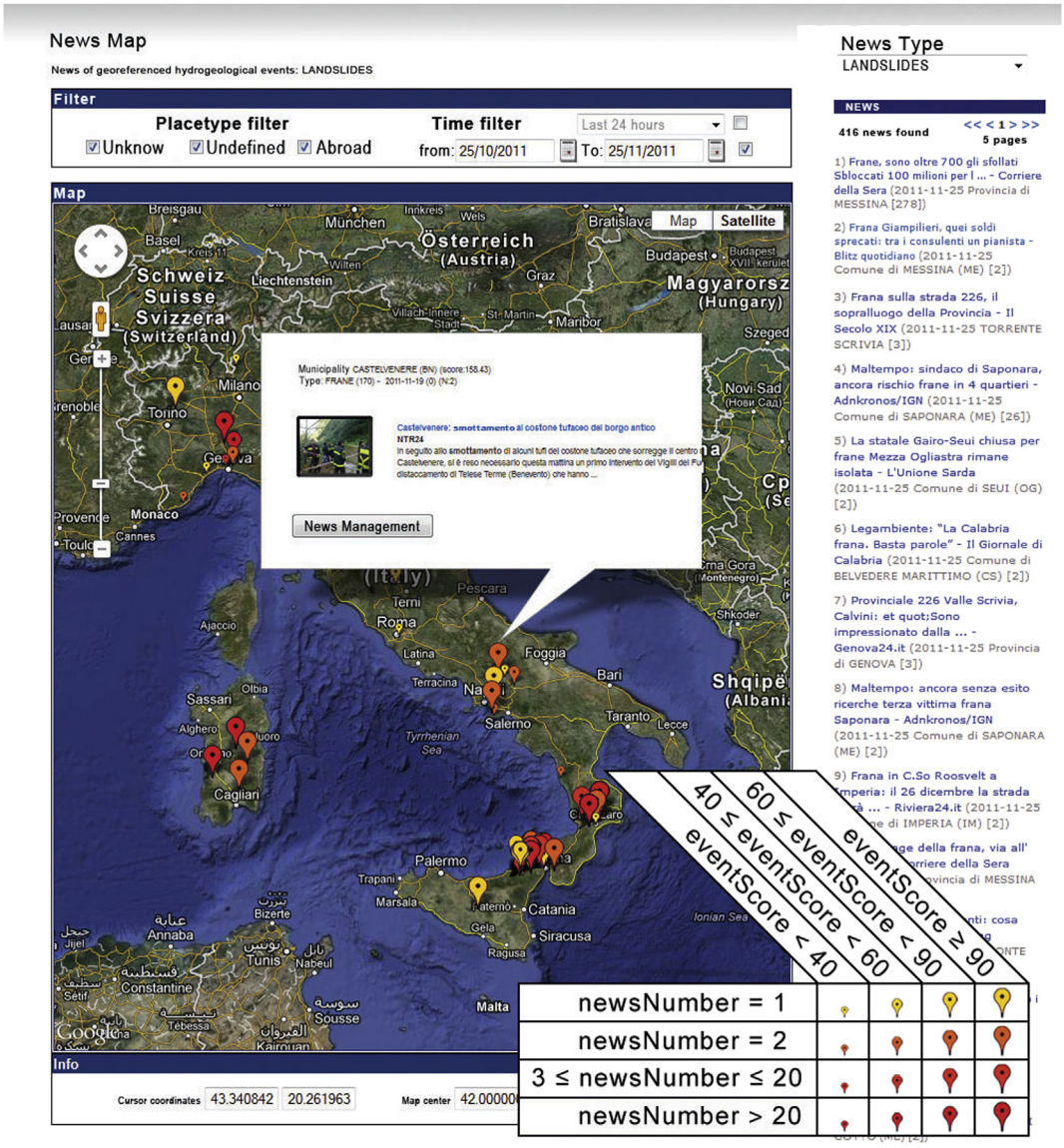

Fig. 3. News management and visualization via WebGIS.

placed, and their assigned parameters in the database can be changed.

Results and discussion

The system operated for 2 years, from November 30, 2009, to November 30, 2011, to test the feasibility of the process. During this test period, the system stored more than 20,000 news items, approximately 5300 items related to landslides, approximately 6500 items related to flooding and approximately 8700 items related to earthquakes. To build the final Italian geohazard database, some items were automatically filtered out according to the threshold values defined in Table 1. To evaluate the data-mining and geotagging processes, the trends of the values assigned to the news items for each stored score (*placeScore*, *eventScore*, *newsNumber* and *timeScore*) as a function of their cumulative frequency is shown in Fig. 4.

Approximately 10% of the collected news was excluded from the final geohazard database because of its approximate location (*placeScore* < 60) (Fig. 4a). The *placeScore* discontinuities at 40 and 300 are from values forced by the geotagging algorithm, i.e., values below 40 are forced to 0 to avoid subsequent calculations, and 300 is the default value assigned to previously geotagged news.

The *timeScore* values show a similar trend, as shown in Fig. 4b. Approximately 80% of the news items were published a few days after the occurrence of the described event and can thus be considered very reliable and are included in the geodatabase. Approximately 10% of the news refers to events that occurred over the previous year, with an almost linear trend. The discontinuities in the graphic are related to news reporting the recurrence of a past geohazard (3% anniversary and 1% double anniversary). Less than 1% of the news was discarded because it reports future dates (primarily related to scheduled events, suppositions or hypothetical sentences).

The *eventScore* trend is shown in Fig. 4c. Similar to the *placeScore* trend, a moderately increasing trend is observed. For a limited number of news items, the trend diverges; this phenomenon is from the data-mining process and reflects a significant decrease or increase of the score when a news item is clearly related or not related to a geohazard. Samples of news with very high or very low *eventScores* were analyzed to obtain the list of words used to calculate the *eventScore* value itself. Based on the *eventScore* threshold (Table 4), approximately 20% of the news was discarded because it was not related to a geohazard, approximately 20% was discarded because the probability that it referred to a geohazard was low, approximately 30% of the news was included in the geodatabase because the probability that it referred to a geohazard was good and approximately 30% of the news was included because of the high degree of certainty of the data.

The inventory maps related to landslides, floods and earthquakes are automatically defined and updated through the application of these filters. Fig. 5 shows the locations of the geohazard news collected in the test period.

An accurate validation of each event location is beyond the scope of this work. It is also not feasible for technical reasons. As stated in the introduction, a complete database of events that have actually occurred (ground truth) is not available. This information will be provided in part through this work.

The data in Fig. 5a show that in many cases, the concentration of news matches with the known landslide susceptibility (Catani, Casagli, Ermini, Righini, & Menduni, 2005; Manzo, Tofani, Segoni, Battistini, & Catani, 2012; Salvati et al., 2010) or with the concentration observed in landslide inventories (Guzzetti, 2000; Guzzetti, Stark, & Salvati, 2005; Martelloni et al., 2012; Trigila et al., 2010). In particular, during the investigated period, the Riviera Ligure was the most affected by landslides. Other areas are the Alps (especially the valleys in the northern Lombardia and Veneto), the Apennines between Tuscany and Emilia Romagna, the Campania region (both in the internal reliefs and in the coast of Naples and Salerno), southern and central Calabria, and the province of Messina in Sicily.

Similarly, the flooding news map (Fig. 5b) is generally in accordance with the existing inventories and hazards maps (Guzzetti et al., 2005; Salvati et al., 2010). The primary concentration of news is reported along the primary rivers (Po, Tiber, Arno and Adige) and in many areas where the hazard is associated with small but nonetheless dangerous streams or torrents, especially in the proximity of their mouths (e.g., Liguria, Adriatic coastline and Lombardy Alps).

Concerning earthquakes (Fig. 5c), the primary areas struck by major (L'Aquila) or ordinary (Gargano, Friuli Venezia Giulia, north eastern Sicily and north Calabria) magnitude events are adequately represented in the geodatabases. However, the system also geotags a relevant number of news items referring to secondary effects or to the local perception of the quakes.

Table 5 shows that not all of the collected information can be gathered from the national database. A large segment of the news retrieved by the data-mining process is classified abroad, but this does not represent an error because this news is filtered out by the

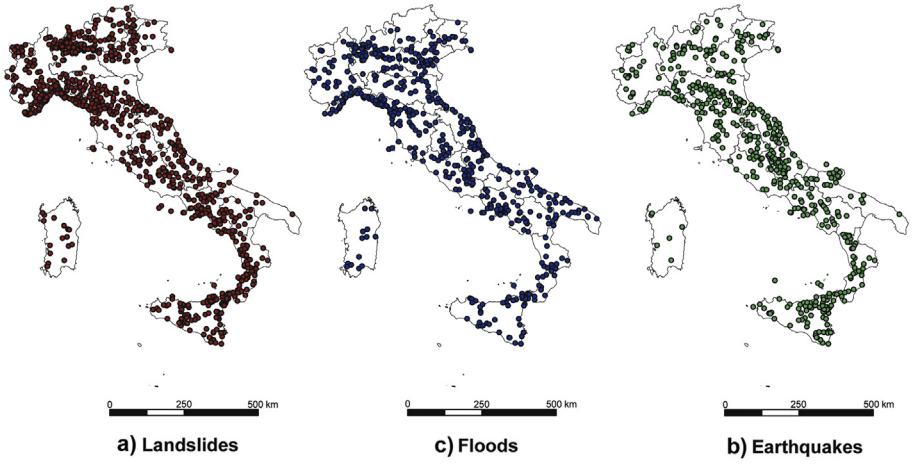

Fig. 4. Cumulated frequency distribution of the scores.

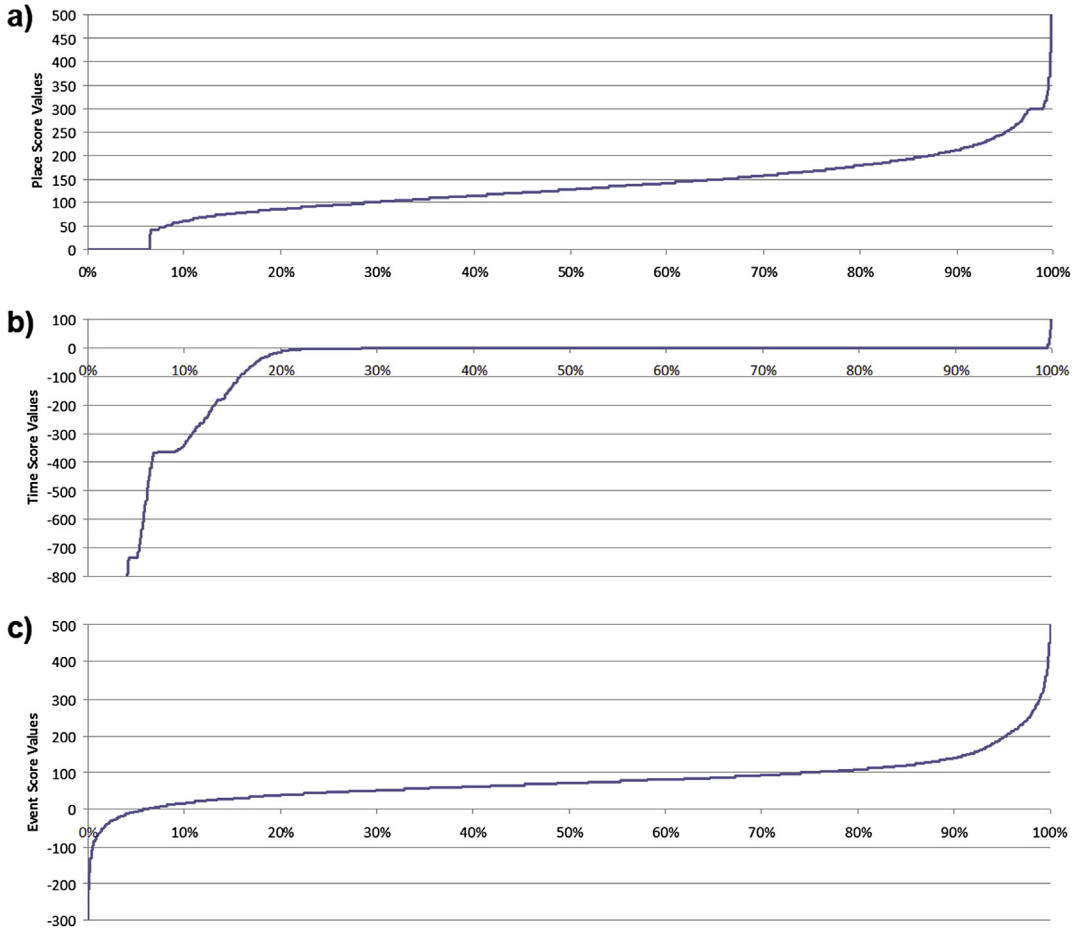

Fig. 5. Spatial distribution of geohazard news collected in the geodatabase during the test period (November 30, 2009–November 30, 2011).

subsequent steps of the data mining process. The primary error committed by the algorithm is related to the incapability of finding a place for approximately 5% of the news, which cannot therefore be automatically georeferenced and is classified as "unknown". A check on a random sample found that 33% of the events discarded with an "unknown" location could be manually placed with high spatial accuracy after checking the news. The rest of the items with "unknown" locations could be assigned to no position (NO PLACE, 62%) or outside Italy (FOREIGN, 5%).

In addition to the number of geohazards included in the geodatabase, it is important to estimate the accuracy of their spatial location (Table 6). The locations must be sufficiently accurate if the database is to be used in operational applications.

Landslides are geotagged with the highest accuracy, i.e., 71% of them are geotagged with the maximum accuracy possible at the scale of this study (municipality). Moreover, 92% of the landslides in the database have been automatically positioned with a sufficient accuracy to be fruitfully used for the validation/calibration of warning systems based on statistical rainfall thresholds, which in Italy and around the world are commonly applied at the regional or sub-regional scales (Baum & Godt, 2010; Cannon, Boldt, Laber, Kean, & Staley, 2011; Jakob, Owen, & Simpson, 2012; Lagomarsino et al., 2013; Rosi et al., 2012). The capability of the data mining process to retrieve and geotag landslide news is most likely related to the spatial extent of landslide phenomena. At the national scale, landslides can be considered point phenomena; therefore, data mining is able to accurately acquire the spatial location of these points.

Table 5
Number of discarded geohazard news items from the national database and their percentage of the total.

	Landslides news	%	Flood news	%	Earthquakes news	%
Geotagged in Italy	4725	87.79%	5071	77.50%	5876	67.20%
Foreign	366	6.80%	1209	18.48%	2428	27.76%
Unknown	291	5.41%	263	4.02%	441	5.04%

Table 6
Number of detected geohazard news items in Italy and the spatial accuracy of their location.

	Landslides news	%	Flood news	%	Earthquakes news	%
Municipality	3371	71.34%	3271	64.50%	3066	52.18%
Province	704	14.90%	800	15.78%	1810	30.80%
Region (<5000 km^2)	22	0.47%	22	0.43%	48	0.82%
Region (>5000 km^2)	355	7.51%	762	15.03%	710	12.08%
Geographic area (<5000 km^2)	129	2.73%	108	2.13%	143	2.43%
Geographic area (>5000 km^2)	19	0.40%	10	0.20%	25	0.43%
River	71	1.50%	88	1.74%	55	0.94%
Lake	11	0.23%	7	0.14%	1	0.02%
Mountain	43	0.91%	3	0.06%	18	0.31%

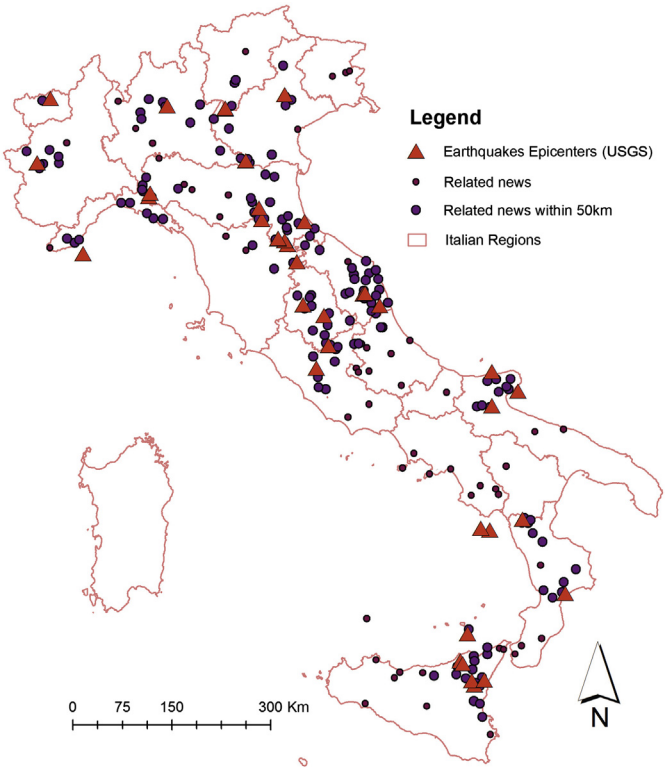

Fig. 6. Earthquake epicenters from USGS and related news.

With respect to floods, 64% of them are geotagged with the maximum accuracy level (municipality). A significant 85% of total occurrences are geotagged with an accuracy that is considered exploitable from an operational point of view (e.g., for the automatic updating of national scale archives for hydraulic risk assessment).

Earthquakes are characterized by the worst statistics. Only 52% of them are geotagged with the maximum accuracy level (municipality), and 87% of the total occurrences are geotagged with an accuracy suitable for applications at the regional and sub-regional scales.

The differences in the statistics displayed by the 3 geohazard typologies in Tables 5 and 6 are primarily related to the destructive impact and the spatial extent of the physical phenomena investigated.

These differences are reflected by the number of Italian geohazard news items that are related to events occurring abroad. The data mining process identified and filtered 2428 foreign earthquake news items (28% of the total earthquake news) and 1209 (18%) foreign flood news items. Only 5% of landslide news items (366 occurrences) retrieved by the data mining process was related to events occurring abroad. Obviously, the foreign geohazards are major destructive events; big disasters are considered more attractive to the general public as a news story and are thus reported in the news, even if they pertain to foreign countries. These occurrences are correctly filtered out by the process (which was aimed at setting up a national geodatabase) but suggest that geotagging is more complicated for earthquakes and (to a lesser extent) floods than it is for landslides. The higher destructive impact of the events may more often lead to the misplacement and inclusion of news items that are not directly related to the actual occurrence of the event and thus to reports of events occurring elsewhere.

In particular, the geotagging of earthquakes with a point is very challenging. From a scientific viewpoint, the epicenter is a point, but news items report the effects of the quake, which may widely impact the society at the regional or national scales. As a consequence, geotagging is problematic (e.g., many points, even located in distant areas, could refer to the same quake that was noticed and therefore reported by distinct local bulletins). At present, this constitutes a primary limitation of the proposed procedure, which is used only to map the perception of earthquakes by the population rather than mapping earthquakes *sensu strictu*.

However, as explained in the introduction, a methodology to geotag earthquake news is not urgently needed because, at present, several national or international network sensors effectively perform a similar task. Thus, earthquake news items were only used to test the procedure empirically, correlating them spatially with the seismic epicenters reported by conventional methods. Using the U.S. Geological Survey search page for global earthquakes (http://earthquake.usgs.gov/earthquakes/eqarchives/epic/ last visited on April 12, 2013), we have collected epicenters of earthquakes of magnitude greater than 3, located in the Italian area (maximum 25 km from the coast) and occurring during the test period. Using only the news suitable for the geohazard database (Table 4), we have verified that news items tend to cluster in the vicinity of the epicenter within 7 days of the event (Fig. 6). Each of the 38 earthquakes generated from 9 to 62 news items (average 30.3), with a distance from the epicenter ranging from 2 km to 282 km (average 48 km, standard deviation 53 km), and 2 earthquakes had no news within 50 km of the epicenter.

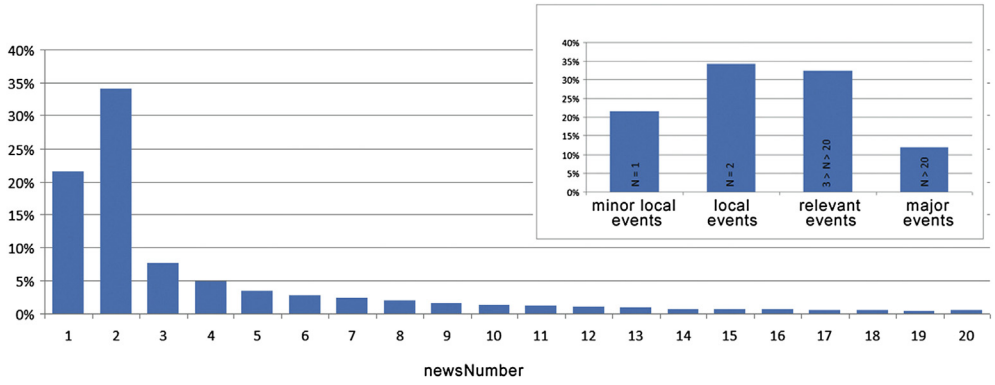

Fig. 7. Frequency distribution and classes based on newsNumber.

Concerning floods, their effects are more spatially confined (even if particularly disastrous events may have wide echoes in the web news) but very often are not assimilable to a point, so the geotagging process may result in inaccurate or uncertain locations.

As stated in the introduction, earthquakes and floods were primarily included to increase the number of data for the calibration of the geotagging and classification algorithms, while the primary focus of this work is on landslides because a methodology for a fast and automatic mapping of events is needed more for landslides than for the other geohazards.

The self-organization of the events included in the geodatabase, according to their magnitude, is analyzed in Fig. 7. If *newsNumber* is considered a proxy for the magnitude of an event or its associated risk level (the more relevant the societal impact of an event, the more widely it is reported by news), the trend observed in the frequency-magnitude relationship is partially in accordance with the power-laws typical for the geohazards considered (Bak & Tang, 1989; Dai & Lee, 2001; Guzzetti, Malamud, Turcotte, & Reichenbach, 2002; Malamud & Turcotte, 2006). A rollover point suggests an under-sampling of minor local events (which could be more frequent than what was reported because these events often do not produce enough damage to be considered newsworthy). After the rollover point, the frequency of events tends to exponentially diminish for increasing event magnitudes. The typical power-law trend cannot be fully established because the data-mining process is based on news and the low frequencies associated with catastrophic events are counterbalanced by the high relevance they receive on the Internet.

Conclusions

Internet news can be considered a constant source of information that is available in digital format; whenever a geohazard impacts society or economy with relevant consequences, news is reported on the Internet. We developed a data-mining process to scan this continuous stream of data and to identify news related to geohazards (landslides, floods and earthquakes) occurring in Italy. Pertinent news is then geotagged by means of a comparison between the possible toponyms found in the text of the news and the Italian geographical features, which are organized in a geodatabase. The service is independent of the content of the news because it can automatically geotag any type of news. Scoring the relationship between the event and the news is optimized only for the specific needs of this study (data mining and geo-localization of geohazards), but the methodology can be applied to other types of news in the future.

The proposed methodology is calibrated and applied to the country of Italy. After a test period of 2 years we demonstrated that the process is capable of performing a continuous analysis, building geohazard geodatabase on a national scale, storing a series of relevant information for each event (e.g., coordinates, timing of occurrence and description) and updating it in near-real time. The process is completely automated and scalable. It can also be applied in other countries or to other categories of news after a specific tuning of the keywords used by the data-mining algorithm.

The outcomes of this study represent an advance in the creation and management of geodatabases because the presented methodology creates and updates national scale databases at a low cost and with greatly reduced efforts. This procedure also allows continuous feedback from the real world. Information about geohazards can be fully exploited in extremely short times compared to traditional techniques that are based on remote sensing, field surveys or historical inventories. Therefore, the proposed methodology can be a valuable tool to assist with national scale hazards and risk assessments. The proposed application can be particularly useful in landslide studies, where complete national scale databases and inventories of dated and georeferenced landslides rarely meet the requirements needed to be effectively used for the calibration and validation of physical and statistical models.

References

Ahilan, S., O'Sullivan, J. J., & Bruen, M. (2012). Influences on flood frequency distributions in Irish river catchments. *Hydrology and Earth System Sciences, 16*, 1137–1150. http://dx.doi.org/10.5194/hess-16-1137-2012.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., et al. (2013). GloFAS – global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences, 9*, 12293–12332. http://dx.doi.org/10.5194/hessd-9-12293-2012.

Bak, P., & Tang, C. (1989). Earthquakes as a self-organized critical phenomenon. *Journal of Geophysical Research, 94*, 15635–15637.

Barriendos, M., Coeur, D., Lang, M., Llasat, M. C., Naulet, R., Lemaitre, F., et al. (2003). Stationarity analysis of historical flood series in France and Spain (14th–20th centuries). *Natural Hazards and Earth System Sciences, 3*, 583–592. http://dx.doi.org/10.5194/nhess-3-583-2003.

Barriendos, M., & Rodrigo, F. S. (2006). Study of historical flood events on Spanish rivers using documentary data. *Hydrological Sciences Journal, 51*(5), 765–783.

Baum, R. L., & Godt, J. W. (2010). Early warning of rainfall-induced shallow landslides and debris flows in the USA. *Landslides, 7*(3), 259–272.

Böhm, O., & Wetzel, K. F. (2006). Flood history of the Danube tributaries Lech and Isar in the Alpine foreland in Germany. *Hydrological Sciences Journal, 51*(5), 784–798.

Brázdil, R., Kundzewicz, Z. W., & Benito, G. (2006). Historical hydrology for studying flood risk in Europe. *Hydrological Sciences Journal, 51*(5), 739–764.

Brunetti, M. T., Peruccacci, S., Rossi, M., Luciani, S., Valigi, D., & Guzzetti, F. (2010). Rainfall thresholds for the possible occurrence of landslides in Italy. *Natural Hazards and Earth System Sciences, 10*, 447–458.

Brunsden, D. (1985). Landslide types, mechanisms, recognition, identification. In C. S. Morgan (Ed.), *Landslides in the South Wales coalfield. Proceedings Symposium, April 1–3, 1985* (pp. 19–28). The Polytechnic of Wales.

Cannon, S. H., Boldt, E. M., Laber, J. L., Kean, J. W., & Staley, D. M. (2011). Rainfall intensity-duration thresholds for postfire debris-flow emergency-response planning. *Natural Hazards, 59*(1), 209–236. http://dx.doi.org/10.1007/s11069-011-9747-2.

Catani, F., Casagli, N., Ermini, L., Righini, G., & Menduni, G. (2005). Landslide hazard and risk mapping at catchment scale in the Arno River basin. *Landslides, 2*, 329–342.

Cummings, C. A., Todhunter, P. E., & Rundquist, B. C. (2012). Using the Hazus-MH flood model to evaluate community relocation as a flood mitigation response to terminal lake flooding: the case of Minnewaukan, North Dakota, USA. *Applied Geography, 32*(2), 889–895.

Dai, F. C., & Lee, C. F. (2001). Frequency-volume relation and prediction of rainfall-induced landslides. *Engineering Geology, 59*(3–4), 253–266. http://dx.doi.org/10.1016/S0013-7952(00)00077-6.

Dikau, R., Cavallin, A., & Jager, S. (1996). Databases and GIS for landslide research in Europe. *Geomorphology, 15*, 227–239.

Galli, M., Ardizzone, F., Cardinali, M., Guzzetti, F., & Reichenbach, P. (2008). Comparing landslide inventory maps. *Geomorphology, 94*, 268–289.

Gräler, B., van den Berg, M. J., Vandenberghe, S., Petroselli, A., Grimaldi, S., De Baets, B., et al. (2013). Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation. *Hydrology and Earth System Sciences, 17*, 1281–1296. http://dx.doi.org/10.5194/hess-17-1281-2013.

GSC. (2011). *Landslides - recent events worldwide*. Geological Survey of Canada. http://gsc.nrcan.gc.ca/landslides/in_the_news_e.php. Last Accessed 10.10.12.

Guzzetti, F. (2000). Landslide fatalities and the evaluation of landslide risk in Italy. *Engineering Geology, 58*(2), 89–107. http://dx.doi.org/10.1016/S0013-7952(00)00047-8.

Guzzetti, F., Malamud, B. D., Turcotte, D. L., & Reichenbach, P. (2002). Power-law correlations of landslide areas in Central Italy. *Earth and Planetary Science Letters, 195*, 169–183.

Guzzetti, F., Stark, C. P., & Salvati, P. (2005). Evaluation of flood and landslide risk to the population of Italy. *Environmental Management, 36*(1), 15–36.

Hilker, N., Badoux, A., & Hegg, C. (2009). The Swiss flood and landslide damage database 1972–2007. *Natural Hazards and Earth System Sciences, 9*, 913–925. http://dx.doi.org/10.5194/nhess-9-913-2009.

Jakob, M., Owen, T., & Simpson, T. (2012). A regional real-time debris-flow warning system for the District of North Vancouver, Canada. *Landslides, 9*(2), 165–178.

Kirschbaum, D. B., Adler, R., Hong, Y., Hill, S., & Lerner-Lam, A. L. (2009). A global landslide catalog for hazard applications – method, results and limitations. *Natural Hazards, 52*, 561–575. http://dx.doi.org/10.1007/s11069-009-9401-4.

Lagomarsino, D., Segoni, S., Fanti, R., & Catani, F. (2013). Updating and tuning a regional-scale landslide early warning system. *Landslides, 10*, 91–97. http://dx.doi.org/10.1007/s10346-012-0376-y.

Lu, P., Casagli, N., Catani, F., & Tofani, V. (2012). Persistent scatterers interferometry hotspot and cluster analysis (PSI-HCA) for detection of extremely slow-moving landslides. *International Journal of Remote Sensing, 33*(2), 466–489.

Malamud, B. D., & Turcotte, D. L. (2006). The applicability of power-law frequency statistics to floods. *Journal of Hydrology, 322*, 168–180.
Manzo, G., Tofani, V., Segoni, S., Battistini, A., & Catani, F. (2012). GIS techniques for regional-scale landslide susceptibility assessment: the Sicily (Italy) case study. *International Journal of Geographical Information Science*, http://dx.doi.org/10.1080/13658816.2012.693614.
Martelloni, G., Segoni, S., Fanti, R., & Catani, F. (2012). Rainfall thresholds for the forecasting of landslide occurrence at regional scale. *Landslides, 9*(4), 485–495. http://dx.doi.org/10.1007/s10346-011-0308-2.
McKean, J., & Roering, J. (2003). Objective landslide detection and surface morphology mapping using high-resolution airborne laser altimetry. *Geomorphology, 57*(3–4), 331–351.
Montesarchio, V., Ridolfi, E., Russo, F., & Napolitano, F. (2011). Rainfall threshold definition using an entropy decision approach and radar data. *Natural Hazards and Earth System Science, 11*(7), 2061–2074.
Morelli, S., Segoni, S., Manzo, G., Ermini, L., & Catani, F. (2012). Urban planning, flood risk and public policy: the case of the Arno River, Firenze, Italy. *Applied Geography, 34*, 205–218. http://dx.doi.org/10.1016/j.apgeog.2011.10.020.
Nottingham, M., & Sayre, R. (2005). *The atom syndication format*. RFC 4287. Network Working Group http://tools.ietf.org/html/rfc4287. Last Accessed 21.12.12.
Papagiannaki, K., Lagouvardos, K., & Kotroni, V. (2013). A database of high-impact weather events in Greece: a descriptive impact analysis for the period 2001–2011. *Natural Hazards and Earth System Science, 13*, 727–736. http://dx.doi.org/10.5194/nhess-13-727-2013.
Petley, D. (2012). Global patterns of loss of life from landslides. *Geology, 40*(10), 927–930.
Reed, C. (2006). *An introduction to GeoRSS: A standards based approach for geo-enabling RSS feeds*. OpenGIS White Paper, OGC 06–050r3. Open Geospatial Consortium Inc. Date: 2006-07-19, Version 1.0.0.
Ricci, S., Piacentini, A., Thual, O., Le Pape, E., & Jonville, G. (2011). Correction of upstream flow and hydraulic state with data assimilation in the context of flood forecasting. *Hydrology and Earth System Sciences, 15*, 3555–3575. http://dx.doi.org/10.5194/hess-15-3555-2011.
Rosi, A., Segoni, S., Catani, F., & Casagli, N. (2012). Statistical and environmental analyses for the definition of a regional rainfall thresholds system for landslide triggering in Tuscany (Italy). *Journal of Geographical Sciences, 22*(4), 617–629.
Rossi, G., Catani, F., Leoni, L., Segoni, S., & Tofani, V. (2013). HIRESSS: a physically based slope stability simulator for HPC applications. *Natural Hazards and Earth System Sciences, 13*, 151–166. http://dx.doi.org/10.5194/nhess-13-151-2013.
Roth, R. A. (1983). Factors affecting landslide susceptibility in San Mateo County, California. *Association Engineering Geologists Bulletin, 20*(4), 353–372.
RSS Advisor Board. (2009). *RSS 2.0 Specification. March 30, version 2.0.11*. http://www.rssboard.org/rss-specification. Last Accessed 21.12.12.
Salvati, P., Bianchi, C., Rossi, M., & Guzzetti, F. (2010). Societal landslide and flood risk in Italy. *Natural Hazards and Earth System Science, 10*, 465–483.
Santangelo, M., Cardinali, M., Rossi, M., Mondini, A. C., & Guzzetti, F. (2010). Remote landslide mapping using a laser rangefinder binocular and GPS. *Natural Hazards and Earth System Science, 10*, 2539–2546.
Satriano, C., Wu, Y. M., Zollo, A., & Kanamori, H. (2011). Earthquake early warning: concepts, methods and physical grounds. *Soil Dynamics and Earthquake Engineering, 31*, 106–118. http://dx.doi.org/10.1016/j.soildyn.2010.07.007.
Schmocker-Fackel, P., & Naef, F. (2010). Changes in flood frequencies in Switzerland since 1500. *Hydrology and Earth System Sciences, 14*, 1581–1594.
Segoni, S., Leoni, L., Benedetti, A. I., Catani, F., Righini, G., Falorni, G., et al. (2009). Towards a definition of a real-time forecasting network for rainfall induced shallow landslides. *Natural Hazards and Earth System Science, 9*, 2119–2133.
Silva, A. T., Portela, M. M., & Naghettini, M. (2012). Nonstationarities in the occurrence rates of flood events in Portuguese watersheds. *Hydrology and Earth System Sciences, 16*, 241–254.
Soeters, R., & Van Westen, C. (1996). *Slope instability recognition, analysis and zonation, in Landslides, investigation and mitigation* (pp. 129–177). Washington, USA: Transportation Research Board, National Research Council, National Academy Press.
Thorndycraft, V. R., Benito, G., Rico, M., Sopeña, A., Sánchez-Moya, Y., & Casas, A. (2005). A long-term flood discharge record derived from slackwater flood deposits of the Llobregat River, NE Spain. *Journal of Hydrology, 313*(1–2), 16–31. http://dx.doi.org/10.1016/j.jhydrol.2005.02.003.
Tofani, V., Segoni, S., Agostini, A., Catani, F., & Casagli, N. (2013). Use of remote sensing for landslide study in Europe: an overview. *Natural Hazards and Earth System Sciences, 13*, 299–309. http://dx.doi.org/10.5194/nhess-13-299-2013.
Trigila, A., Iadanza, C., & Spizzichino, D. (2010). Quality assessment of the Italian landslide inventory using GIS processing. *Landslides, 7*(4), 455–470.
USGS. (2011). *Landslide events*. US Geological Survey. http://landslides.usgs.gov/recent/. Last Accessed 10.10.12.
Van Den Eeckhaut, M., & Hervás, J. (2012). State of the art of national landslide databases in Europe and their potential for assessing landslide susceptibility, hazard and risk. *Geomorphology, 139–140*, 545–558. http://dx.doi.org/10.1016/j.geomorph.2011.12.006.
Van Westen, C. J., Castellanos, E., & Kuriakose, S. L. (2009). Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. *Engineering Geology, 102*(3–4), 112–131.
Varis, O., Kummu, M., & Salmivaara, A. (2012). Ten major rivers in monsoon Asia-Pacific: an assessment of vulnerability. *Applied Geography, 32*(2), 441–454.
Wald, L. A., Wald, D. J., Schwarz, S., Presgrave, B., Earle, P. S., Martinez, E., et al. (2008). The USGS earthquake notification service (ENS): customizable notifications of earthquakes around the globe. *Seismological Research Letters, 79*(1), 103–110. http://dx.doi.org/10.1785/gssrl.79.1.103.
Wetmiller, R. J., Adams, J., & Woodgold, C. (2007). Canada's automated earthquake notification service. In *Proceedings of the Canadian Dam Association Annual Conference 2007, St John's, September 22.27* (pp. 13). (ESS Cont.# 20080640).
Wirtz, A., Kron, W., Löw, P., & Steuer, M. (2012). The need for data: natural disasters and the challenges of database management. *Natural Hazards*, 1–23. http://dx.doi.org/10.1007/s11069-012-0312-4.
Zhou, X., Persaud, N., & Wang, H. (2006). Scale invariance of daily runoff time series in agricultural watersheds. *Hydrology and Earth System Sciences, 10*, 79–91.