

Exploring a landslide inventory created by automated web data mining: the case of Italy



Abstract Nowadays, several systems to set up landslide inventories exist although they rarely rely on automated or real-time updates. Mass media can provide reliable info about natural hazard events with a relatively high temporal and spatial resolution. The news publication about a natural disaster inside newspaper or crowdsourcing platforms allows a faster observation, survey, and classification of these phenomena. Several techniques have been developed for data mining inside social media for many natural events, but they have been rarely applied to the automatic extraction of “landslide events”. This source of information allows continuous feedback from real world, and news concerning landslide events can be rapidly collected. In this work, the newspaper articles about landslides in Italy are automatically collected by an existing data mining algorithm, based on a semantic engine. The news has been analysed to assess their distribution over the territory and to verify the possibility of using them for hazard mapping purpose. In 10 years, from 2010 to 2019, the algorithm identified and geolocated 184322 articles referring to 32525 generical events (“news”). At first, the collected data underwent to a manual verification, followed by a classification based on news relevance, localization accuracy and time of publication. Then, these data have been used to identify the areas and the periods most affected by landslide phenomena. The analyses show that almost 42% of Italian municipalities have been affected by landslide. According to the results, the use of data mining is helpful for the creation of landslide databases where the day and the approximative location (municipality) of the possible landslide triggers are known. This database, in turn, can be used for scientific purposes, as the definition of the meteorological condition associated with landslide initiation, the validation of risk maps. It can also be used for a proper land use or risk mitigation planning, since the most landslide-prone municipalities can be defined.

Keywords Landslide · Social media · News

Introduction

Landslides are extremely widespread in the Italian territory, and they are, along with floods, the most frequent natural hazard, causing the greatest number of losses of human lives and damages to properties and infrastructures (Guzzetti 2000). Over the time, there has been an increase of the risk of landslide due to increasing anthropization of the territory (ISPRA 2020a) even near to unstable area. In Italy, the estimate annual losses caused by landslide are 3.9 billion Euros (Klose et al. 2016).

Landslide research chiefly relies on landslide inventories for a multitude of spatial, temporal or process analysis (Van Den Eeckhaut and Hervás 2012; Kirschbaum et al. 2015; Klose et al. 2015). These inventories can be created with several methods as

photo-interpretation, field surveys (Brunsden 1985) and remote sensing (Soeters and Van Westen 1996; McKean and Roering 2003; Lu et al. 2012; Bianchini et al. 2018; Solari et al. 2020) or retrieval of data from technical reports and/or newspapers (Kirschbaum et al. 2010; Görüm and Fidan 2021; Guzzetti et al. 2008; Klimeš et al. 2017; Vennari et al. 2014; Rosi et al. 2019) or a combination of them (Dikau et al. 1996; Rosi et al. 2012; Rosser et al. 2017).

All these traditional approaches are usually quite accurate, but they are also very demanding tasks, even at detailed scales (Brunsden 1985; Santangelo et al. 2010), hence time consuming.

Mass media is generally the first and primary source of information about hazards for the public (Fischer 1994). Literature studies indicate that the social sensors in terms of tweets and other social media websites report a natural disaster much faster than the observatories (Goswami et al. 2016). The data generated on social media provide a unique opportunity to capture disaster situations with a relatively high temporal and spatial resolution to map different events across various locations (Fan et al. 2018; Rachunok et al. 2019; Saltelli et al. 2020). There are several ways of using social media in disaster management, including data collection, analytic workflow, narrative construction, disaster relevant information extraction, geo-localization pattern/text/image analytics and the broadcasting of information through social media platforms (Carley et al. 2016). Because social media is widely used for various purpose, vast amounts of user-generated data exist and can be made available for data mining (Gundecha and Liu 2012). The data mining research has successfully produced numerous methods, tools and algorithms for handling large amounts of data to solve real word problems (Gundecha and Liu 2012). In recent years, artificial neural networks (ANNs) (Ashoghi et al. 2020), support-vector machines (SVM), and decision tree have been used extensively as the data mining models (Goswami et al. 2016). Systems using automated or real-time updates are still uncommon and only used for some types of natural hazards (Battistini et al. 2013, Battistini et al. 2017; Calvello and Pecoraro 2018), mainly earthquakes, floods and wild-fire, while creating a complete and updated database is more difficult for landslide (Galli et al. 2008; Santangelo et al. 2010). The methodology of Battistini et al. (2013, 2017) and Kreuzer and Damm (2020) allows to update in near real time the landslide database using the data mining technique inside online newspaper articles.

Newspaper articles can represent a relevant source of data for landslide scholars, and several authors used them to collect info about landslide events. Early works (Guzzetti et al. 1994; Cuesta et al. 1999; Devoli et al. 2007) were based on the manual search and collection of newspaper articles, while more recent works (Kirschbaum et al. 2010, 2015; Taylor et al. 2015; Klimeš et al. 2017;

Görüm and Fidan 2021) used automated procedure to identify landslide-related news.

For example, Kirschbaum et al. (2010, 2015) and Klimeš et al. (2017) made use of Google Alert associated with proper keywords, to identify the news items, and Taylor et al. (2015) used a set of Boolean search terms to query the Nexis UK newspaper archive.

Even if the news items have been gathered by automated procedure, the literature review revealed that the collected data usually have been manually analysed to identify the landslide location and the date of the event.

The newspaper articles used in this work have been harvested by a data mining algorithm named SECaGN (Semantic Engine to Classify and Geotagging News, Battistini et al. 2013). The data mining takes place within Google News as it considers with more completeness national and local newspapers. The identified data are automatically dated, located and arranged by the system and filed in a geodatabase in near real time. This source of information allows continuous feedback from real world, and the news related to landslides can be rapidly collected (the system is set to scan Google news every 15 min) and used in extremely shorter times compared to traditional techniques (Battistini et al. 2017). In addition, it allows to define a more complete landslide database, even considering events with less social impact and catastrophic effects.

The objective of this work is to get the spatial and temporal distribution of landslides phenomena for the Italian context through online news harvested by SECaGN. To achieve this goal, the newspaper articles retrieved from online newspaper have been firstly

validated and then classified into 3 classes: (i) article related to recent landslide events, where the landslide triggering date and its approximal location can be identified; (ii) article related to landslide but without information about the triggering date and a scarce location accuracy (province, region or geographical zone); and (iii) articles not related to landslides, which have been removed from the database. Landslide-related articles have been analysed to assess their spatial and temporal distribution and then compared with landslide hazard map and population living in landslide risk zones map. In this work, the possibility of using web data mining to create a landslide database over a large area has been explored, and it also resulted that this approach is not suitable for very detailed landslide inventories, since some technical data, as landslide type, volume or exact location, cannot be always retrieved from newspapers.

Study area

Italy is almost 300.000 km², and it is divided into 107 provinces and 7926 municipalities, most of them affected by landslide hazard (Figure 1A). Much of Italy consists of hilly and mountainous terrain subject to landslides of different types and sizes (Guzzetti 2000). Nowadays, in the IFFI database (which is the Italian Inventory of Landslide), 470.000 landslides are reported, involving an area of 20.000 km², representing the 6.6% of the national territory. The most common movement types are rotational and translational (slump and slide), debris flow and complex (as defined in Cruden and Varnes 1996).

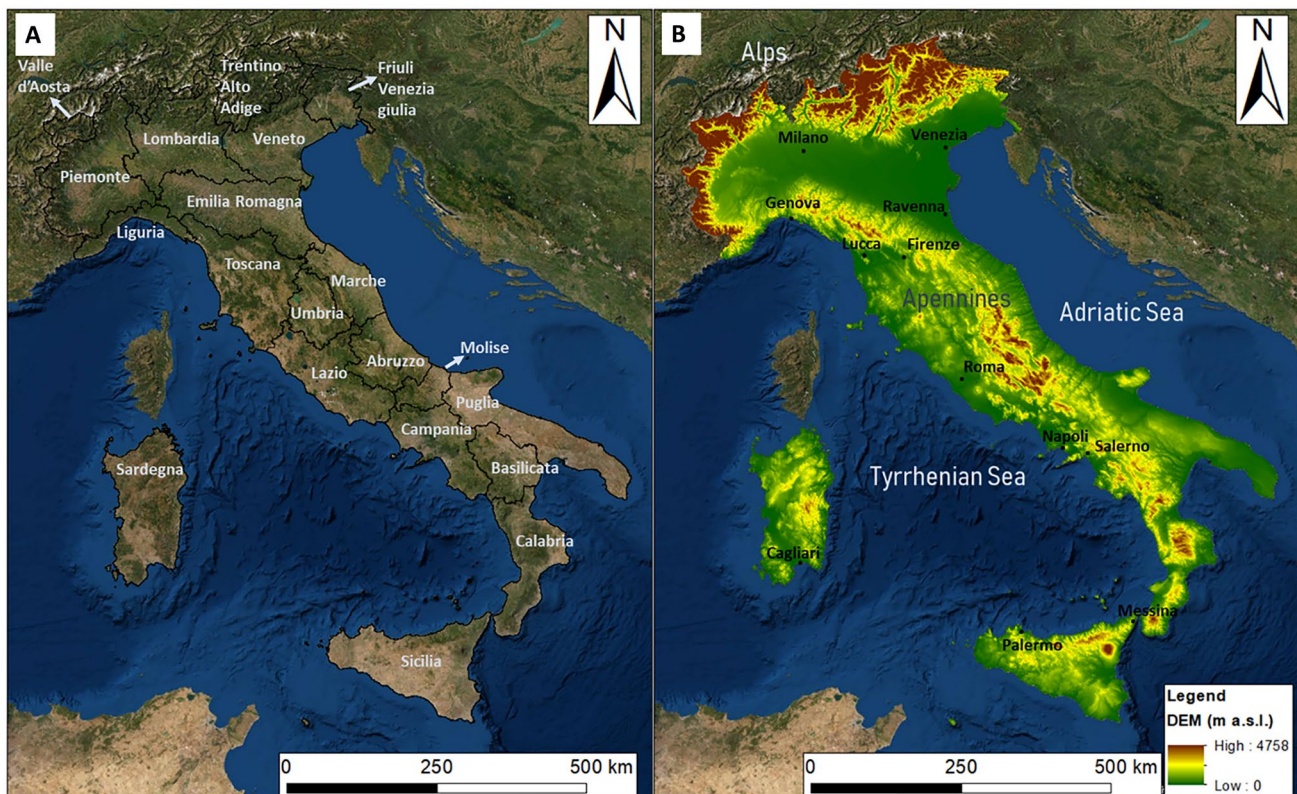


Figure 1 A Regions of Italy and B digital elevation model with the names of cited cities

The main relieves are the Alps, which span from East to West of the northern border of the country, and the Apennines, that cross the country from North to South (Figure 1B). In the alpine area, which is formed mainly of metamorphic rocks (Vai and Martini 2001; Salvatici et al. 2018), the most frequent phenomena are rock fall and debris avalanches (Agliardi and Crosta 2003; Panizza et al. 2011), while in the Apennines, which are formed mainly of arenaceous flysch (Vai and Martini 2001; Agostini et al. 2014; Rosi et al. 2018, 2021), the most common landslides are represented by rotational and translational landslides, both surficial and deep-seated.

The climate of Italy is mainly Mediterranean, with dry and warm summers and mild and wet winters; during winter, snowfall is frequent both on the Alps and on the Apennines, and the consequent snowmelt in the springtime often leads to the mobilization of landslides.

Material and method

Semantic Engine to Classify and Geotagging News (SECaGN)

SECaGN is an algorithm based on a mechanism of acquisition, management and publishing of online articles related to natural hazard (landslides, floods and earthquakes). It aims to get information about spatial and temporal distribution of the events. The automatic search for newspaper articles is performed combining primary words, synonyms, singular and plural forms (keywords) in Italian language related to the landslide argument. The data mining is applied inside Google News. After the acquisition process, a data filtering procedure is applied to separate non-relevant information from pertinent items. The data filtering takes place through the geotagging and the cataloguing of articles using three scores (Battistini et al. 2013):

- Place score: a score value is assigned to evaluate the reliability of the geotag.
- Event score: index of the probability that the news item actually concerns the topic event.
- Time score: estimated days between the time of occurrence of the event and the time of publication of the article.

All the newspaper articles that reach a minimum score are then filed in a geodatabase and their location can be viewed in a dedicated WebGIS (Figure 2A). The whole process is repeated every 15 min.

This data mining methodology was calibrated and tested in Italy during a test period of 2 years (November 2009–November 2011).

The process is completely automated and scalable. It can also be applied in other countries after a specific tuning of the keywords used by the data mining algorithm.

Manual supervision

The SECaGN algorithm identified 184322 newspaper articles about landslide events from 2010 to 2019. The retrieved articles refer to 32525 generical events or “news”.

It is to notice that each landslide event can be reported from 1 or more newspaper, based on its impact or on the relevance of the affected area; for example, small landslides involving a major road or an important city can have a vast media echo, while landslides involving minor roads or small villages are reported only by local newspapers.

In this way, the landslide event can be reported in several newspaper articles that are grouped in a single news, which hence refers to a specific landslide event.

Even if the SECaGN results were already tested in previous papers (Battistini et al. 2013, 2017), in this work, the news item underwent a manual verification and a classification based on their relevance, localization accuracy and time of publication. This classification (Figure 2B) allows to identify the most relevant news, in terms of temporal and spatial accuracy of landslide event identification.

For the classification, 3 classes have been defined (Figure 3, Table 1):

Class 1: “Near real time news”. In this category, all the news referred to ongoing or very recent landslide events (same day or a couple of days before) are classified. These news are also characterized by a high level of spatial accuracy (at least the municipality must be identified), with an approximation of few kilometres. Some news, with high temporal precision but low spatial accuracy, have been manually modified (if possible) based on article text, to reach the required level of approximation. The news in this class can be used for further analyses or modelling (Battistini et al. 2017).

Class 2: “News generically referred to landslides”. In this category, the news referred to past landslide with unknown triggering date (e.g. “the initiation (or finishing) of works aimed to risk reduction or to landslide remediation”) are stored. News with a low spatial accuracy (referred to provinces/cities or geographical areas) are classified in Class 2 as well. This kind of news is useful to identify those areas that have been affected by landslides in the past and for hazard/risk zoning.

Figure 2 Workflow of the work: **A** data mining and geotagging procedure; **B** news analysis

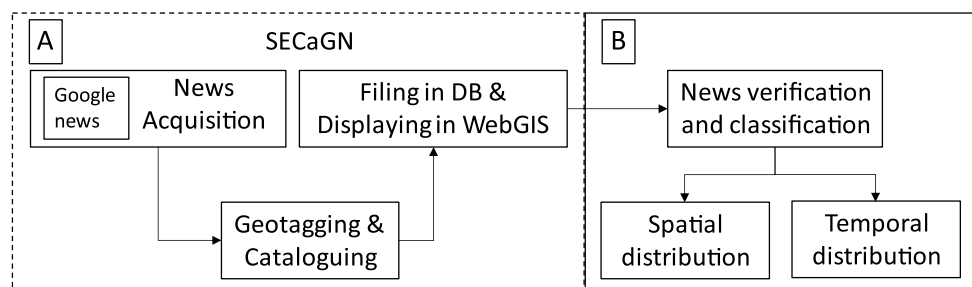


Figure 3 General distribution of the used news in Italy

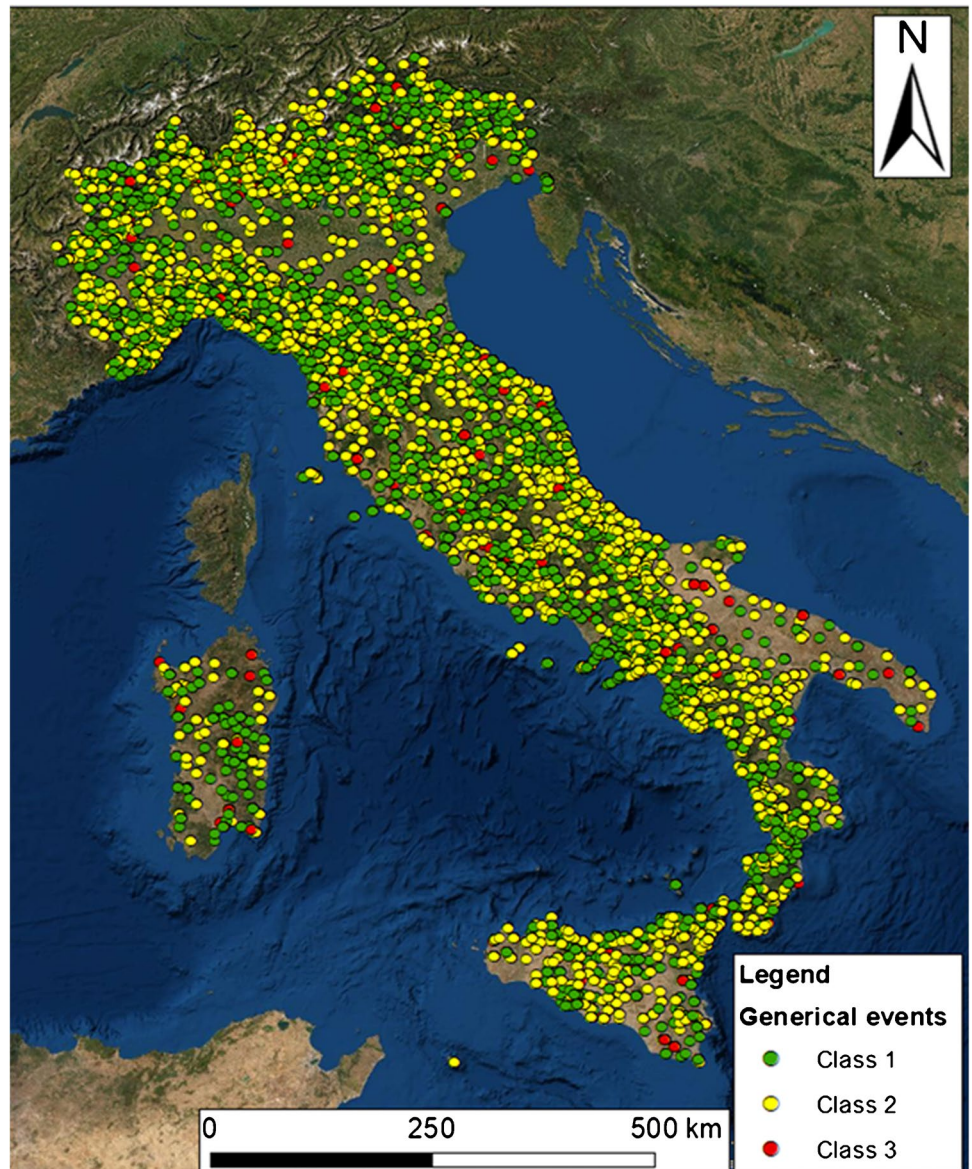


Table 1 Description of the 3 classes used to group the news.

Class	Time	Localization	Title example
1	Near real time	Municipality or village	Gallivaggio landslide, the video
2	The date of the event cannot be defined	City, region, lake, river	The funds for securing the landslides are now available, the works will start soon
3	-	-	Italy's economy does not grow up, the south slide down

Class 3: “News not related to landslides”. News not related to the landslide argument but whose semantic association lead to a misclassification. After this work, these news have been removed from the database.

The news classified in classes 1 and 2 have been then used to explore landslide distribution in Italy, both at region and province levels (in Italy, each region is divided into provinces, and each

province is divided into several municipalities) as well as to explore the temporal distribution of the news.

Headline text analysis

The headlines of each article have been analysed using Natural Language Processing (NLP) technique (Liddy 2001).

NLP is a computerized approach for textual analysis, and it provides several techniques to model the textual data. In this work, the word frequency technique has been used with the scope of identifying the most common associations of words both for “good” and “bad” news. The results of this analysis can help to improve the data mining algorithm.

Results

From 2010 to 2019, 32525 news have been gathered by the used data mining algorithm. Among them, 13275 news had useful information about the geo-localization and the date of landslide event; 1400 news have been corrected, attributing a more appropriate localization based on the text into the article.

According to the adopted classification criteria, the identified news has been classified as follows:

- Class 1: 13275 news (41%)
- Class 2: 18603 news (57%)
- Class 3: 647 news (2%)

This classification allowed to identify the “true news” (classes 1 and 2) and to reject the data not appropriate (class 3), reducing the data to be processed. About 41% of news reported information

relative to recent landslide, and only a minimum percentage of the database is made up by wrong news (2%) (Figure 4A). A textual analysis has been conducted to retrieve the frequency of words inside the headlines. In Figure 4B, C and D, the most frequent words of the headlines of the classes 1, 2 and 3 news are reported, respectively. The term “landslide” is present in all categories as first word widely used; indeed, in the class 1, the word “landslide” is present 8021 times, 10457 times in class 2 and 271 times in class 3.

After the word frequency analysis, the spatial distribution of the data was explored, as described below.

The used data mining algorithm cannot identify the exact location of a landslide, since it is not usually reported in newspapers; therefore, the data have been grouped on regional base (Figure 5A) and on provincial base (Figure 5B) to identify the areas with a higher number of landslide news. Class 2 news have been used only on regional scale aggregation since some of them do not provide an adequate localization accuracy for a more detailed analysis. According to the spatial distribution of the news, during last 10 years, 41.7% of the municipalities suffered at least one landslide.

The regions most involved by landslide are mainly in the northern area of the Country. Liguria and Lombardia are the regions with the highest number of news (classes 1 and 2) and therefore of articles publication (article referred to the same

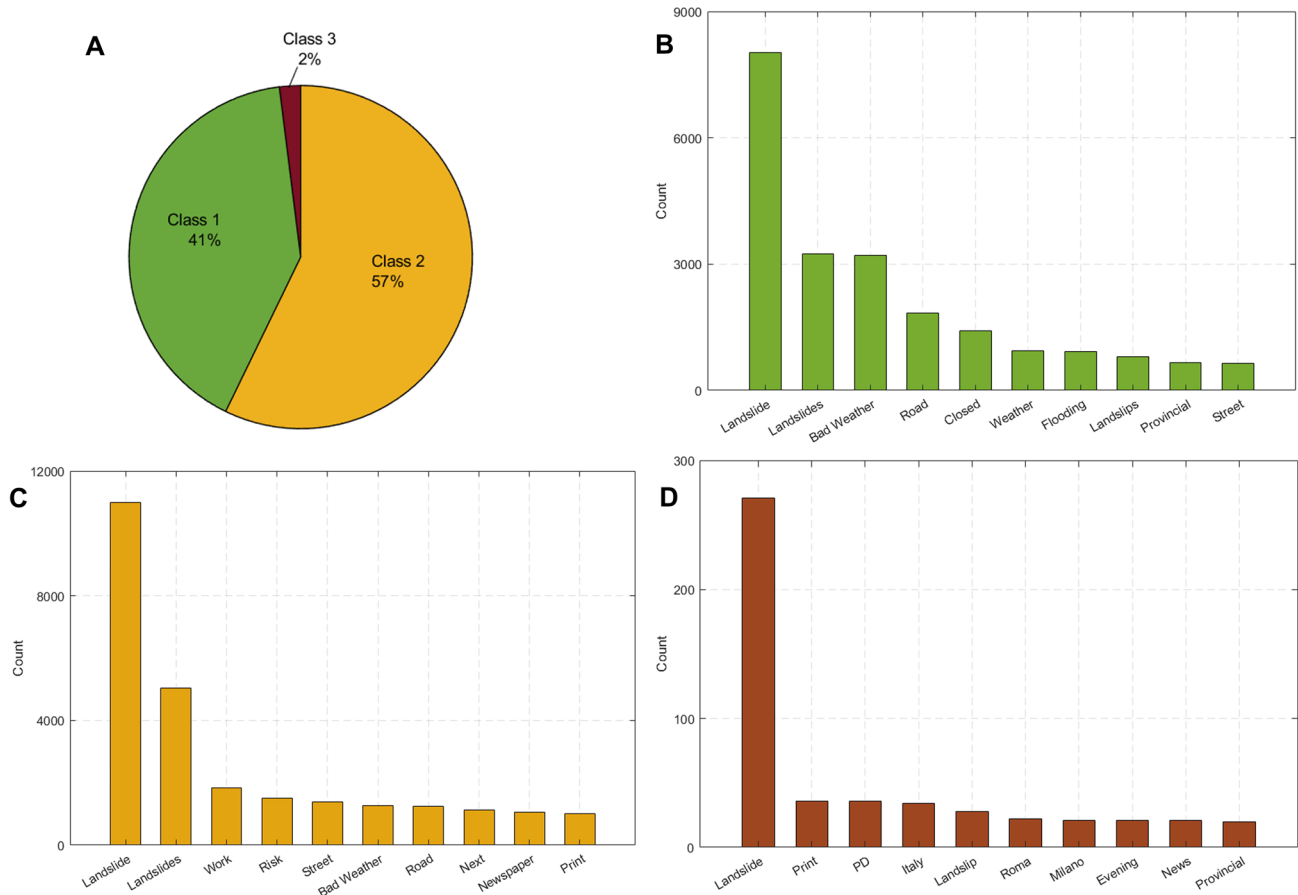


Figure 4 A Overall landslide news classification. B Words' frequency in the headlines inside class 1. C Words' frequency in the headlines inside class 2. D Words' frequency in the headlines inside class 3

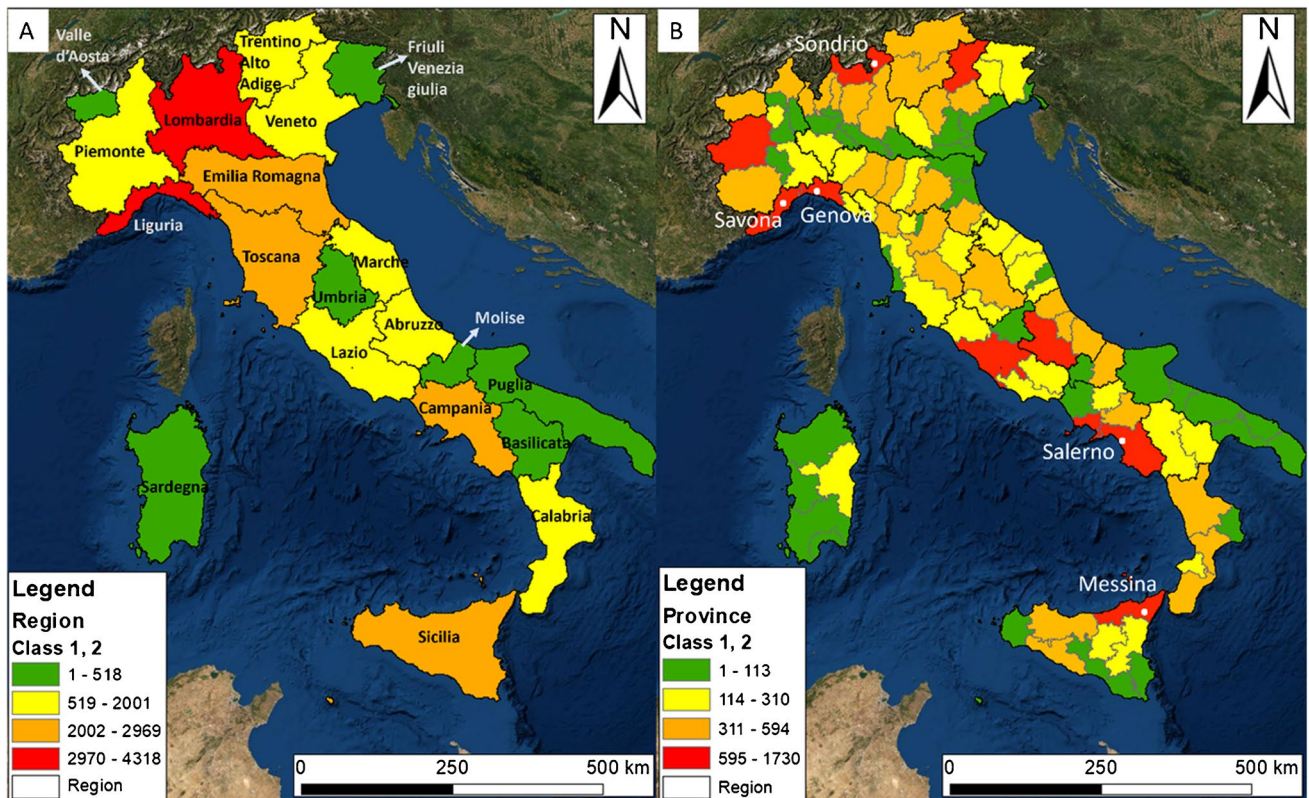


Figure 5 Spatial distribution of landslide news: **A** Regional aggregation with overall news (classes 1, 2); **B** province with only news about recent landslides (class 1). Genova is the province most affected by landslides, followed by Salerno, Messina, Savona and Sondrio. The Puglia region and the provinces along the North-East coast show a lower number of landslide events

landslide event are grouped into a single “landslide news”). For example, Liguria has 36451 articles referring to 4318 landslide news (classes 1 and 2, Figure 5A); among them, 19844 articles refer to 1174 recent “landslide events” (class 1, Figure 6A), and in particular Genova is the most affected province by landslides (Figure 5B).

Besides the alpine area, several other provinces over the country showed a relevant number of news (Salerno, Messina, Savona, Sondrio), and they are mainly located along the western coast (Tyrrhenian seacoast) and along the Apennines mountain belt (Figure 5B), which is historically affected by landslides, because of its geological origin and the high frequency of clayey slopes.

The Puglia region (Figure 5A) and the provinces along the North-East coast (Figure 5B) show a lower number of landslide news, because they are mainly flat areas and less landslides are obviously expected (Figure 6A), as well as the Southern part of Lombardia and Veneto, and the North-Eastern part of Emilia-Romagna region.

Figure 6A shows the distribution of only class 1 news (referred to recent landslide events) at regional scale; also in this case, Liguria is the region with the highest number both of articles and landslide events. Lombardia is the second region, regarding the number of landslide events, but with a lower number of articles, while Sicilia and Toscana are the second and the third region, respectively, in terms of published articles, even if with a lower number of landslide events.

From a temporal point of view (Figure 6B), the year with the highest number of landslide-related articles (blue bars) is the 2014, while the number of landslide events (orange line) showed a very sharp increase from 2017 (1243 events) to 2019 (2901 events).

Once a general overview of spatial and temporal distribution of news has been accomplished, a more detailed analysis about only class 1 news has been carried out.

Figure 7A displays a monthly distribution of the landslide events identified by the class 1 data; it shows that November, March and February are the months more involved by landslides.

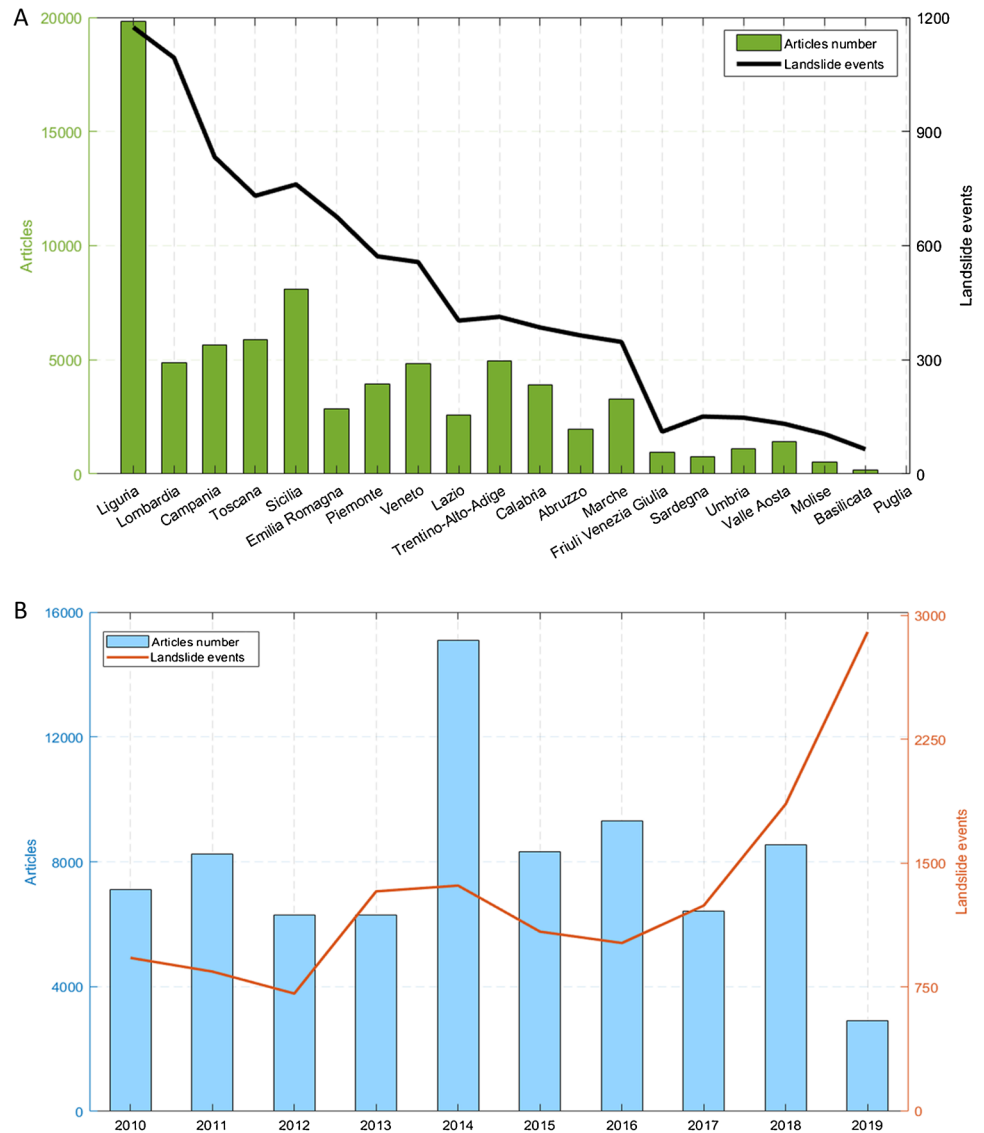
Indeed, November, in 10 years, reported 2093 landslide events with 20142 published articles (multiple articles can refer to the same landslide event, as described in the previous section), while July, June and September are the months with less events. For instance, in July 597 landslide events were reported by newspapers.

Class 1 news has been further analysed to identify the number of days with at least 1 landslide reported (Figure 7B).

The annual distribution (Figure 7B) follows a gradual increase of days with at least 1 landslide from 2015 to 2019; in this period, 8103 landslide events have been collected, distributed over 1378 days, with an average of almost 5 landslides each day (Figure 7B), while from 2010 to 2014, 5172 landslide events, distributed over 1236 days, were reported.

The number of days with at least 1 reported landslide event (landslide day) is higher the northern regions rather than in the southern ones, except for Sicily, the southernmost region, where

Figure 6 Comparison between the number of published articles and of landslide events. **A** Regional distribution, **B** temporal distribution. In both the panels, histogram represents the distribution of published articles about recent landslides (class 1), and the black and orange lines represent the number of landslide events; both of them are referred to class 1 data



a high number of landslide days is present (Figure 8A). Overall, 5 regions out of 20 had at least 450 days with landslide events, in the analysed period. Lombardia, Liguria, Campania, Sicilia and Toscana are the regions with the highest number of days characterized by landslides. In particular, 677 days with landslides have been identified in Lombardia, 572 in Liguria, 545 in Campania, 475 in Sicilia and 451 in Toscana (Figure 8A). The Puglia region has the lowest number of landslide days: in this region, 72 landslide events, distributed over 49 days, are present.

In a more detailed scale (Figure 8B), 4 provinces out of 107 have a high number of days with landslide events (180–301), while the average value is 23 days with landslides every year. For example, the Genova province is characterized by 915 landslide events, reported in 12942 articles, distributed over 301 days. The provinces that have less days with at least one landslide event are located along the North-East coast of Venezia, Rovigo, Ferrara and Ravenna.

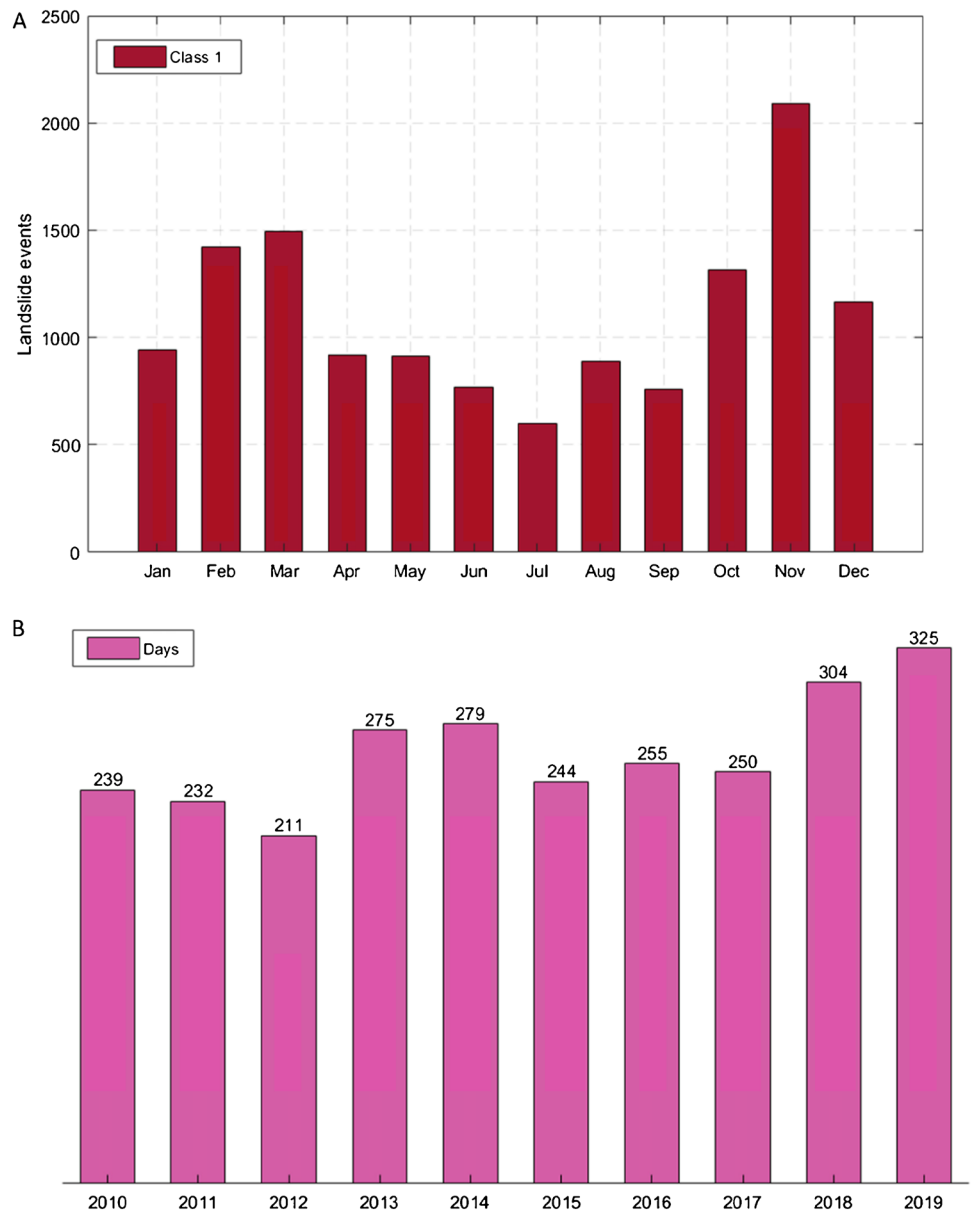
In general, results show that Liguria, Lombardia, Campania, Toscana and Sicilia are the regions with the highest number of both “landslide events” and “landslide days”.

Comparison with existing datasets

In order to validate the quality of the results, mainly of the spatial distribution of landslide events, a comparison with existing datasets about landslides has been made. The landslide hazard map of Italy (Trigila et al. 2018) and the map of population living in landslide-risk areas (Trigila et al. 2018) have been used.

These 2 maps have been processed to extract the percentage of area of each region affected by landslide hazard (Fig. 9B) and to calculate the percentage of population of each region living in zones affected by landslide risk (Fig. 9C). This operation was needed to account the differences in size and population of the different regions, which can vary greatly. Furthermore, some large regions (e.g., Lombardia, Veneto, Emilia-Romagna) are characterized by wide plain areas, and this will result in low percentages of territory affected by landslide hazard. The use of population at risk (as percentage of total regional population) was aimed to overcome this problem.

Figure 7 Temporal distribution of class 1 news. **A** Monthly distribution of “landslide events”; **B** the number of days with at least 1 landslide reported from 2010 to 2019



The comparison between the three maps in Figure 9 shows a good agreement between the distributions of landslide news, landslide hazard or people at risk for several regions, even if some anomalies can be identified. For instance, Valle d’Aosta shows a lower number of landslide news but a very high portion of the territory subject to landslide hazard (94%), or Lombardia has a high number of news and a low percentage of its territory subject to landslide hazard.

Then, the number of landslide events has been correlated with the aforementioned percentages to better verify the existence of a correspondence between these variables (landslide events, landslide hazard and population at risk). As shown in Figure 10A, there is a general correlation between the number of news (classes 1 + 2) and the areas affected by landslide hazard for each region as well as with the population living at risk. The distribution of the data shows some anomalies that are due to the morphology of the territory

and the size of the regions. As stated above, some large regions (Emilia-Romagna, Lombardia, Piemonte, Veneto and Sicilia) are characterized by large plain areas, so the percentage of hazardous zones (for landslides) is low, but not neglectable; this leads to the higher news/hazard ratio than in other regions as Liguria, Toscana or Trentino Alto Adige, where there are few plain areas.

Discussion

In this study, “landslide news” in Italy, automatically retrieved from web sources from 2010 to 2019, have been used to create a landslide database, which has been analysed to evaluate the spatial and temporal distribution of landslide events.

For the analysis, only newspaper articles reported inside the Google News aggregator have been considered, because it collects national and local newspapers, offering a better coverage of the

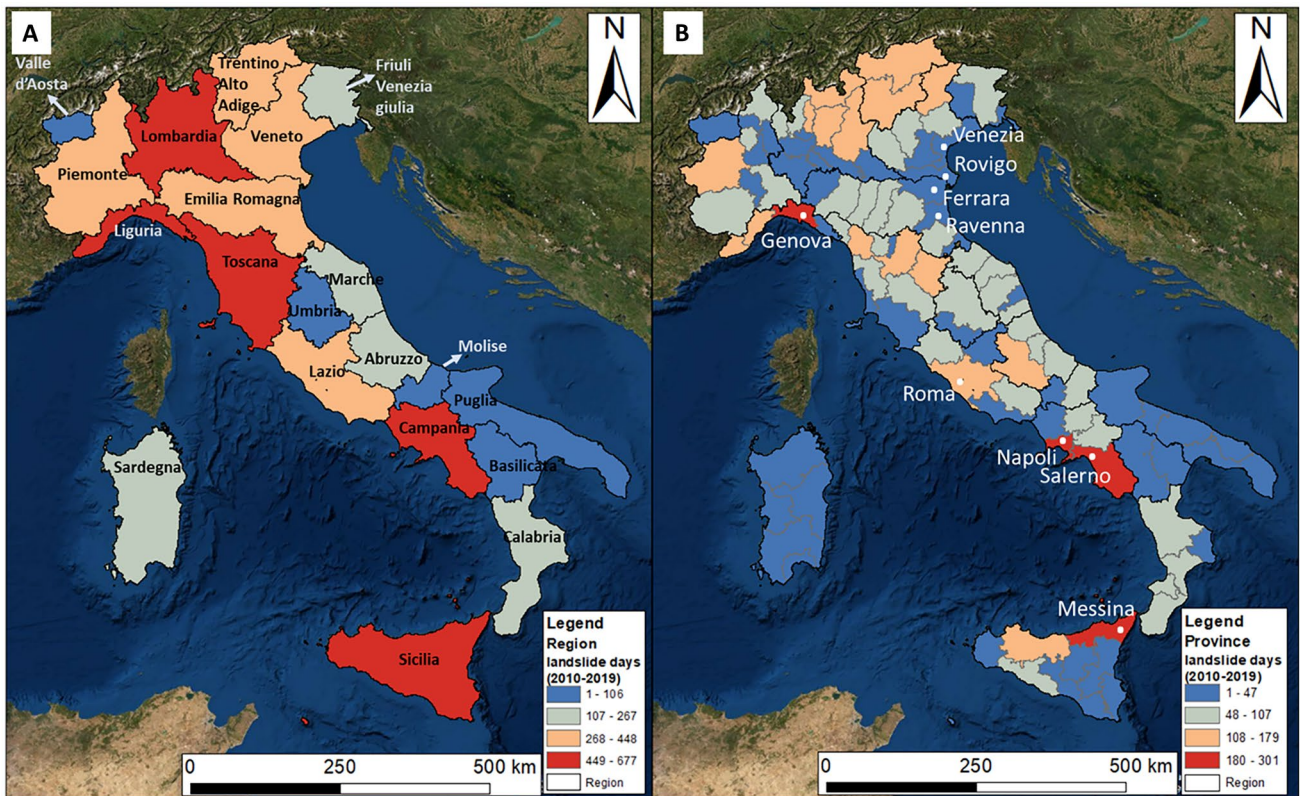


Figure 8 Spatial distribution of days with reported landslides. **A** Regional distribution. **B** Provincial distribution

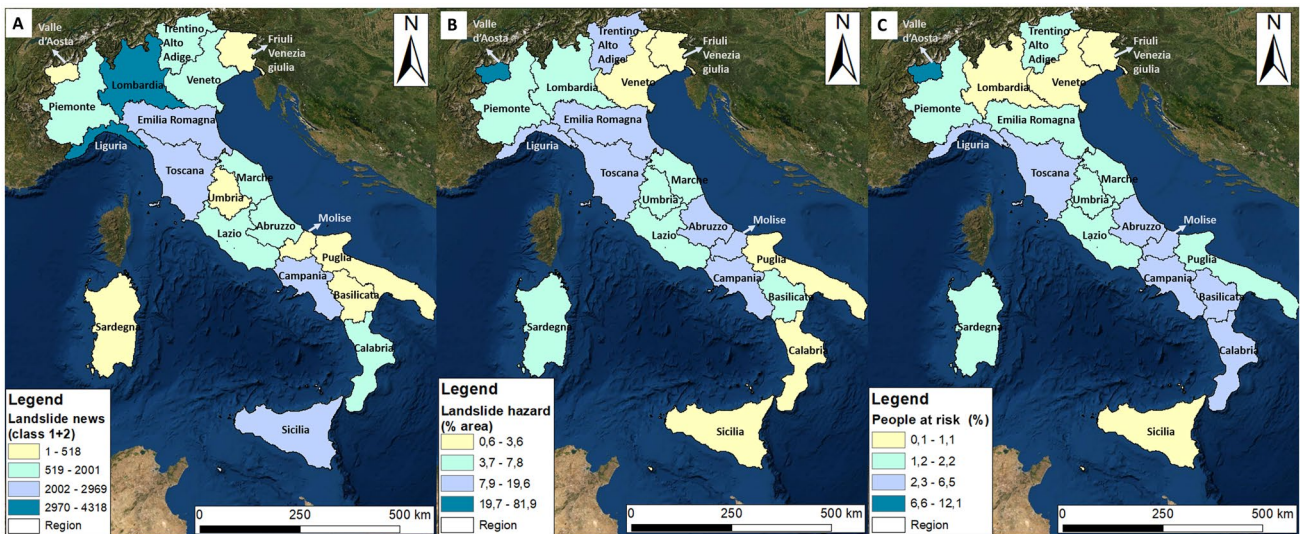


Figure 9 comparison between the distribution of landslide news (classes 1 and 2, **A**), landslide hazard (**B**) and people at risk (**C**)

data, and hence a better completeness in the creation of the landslides database.

Over 40% of news reported useful information (geo-localization, date) about recent landslides (class 1), while 57% of news can be used to identify an area involved by a landslide, but not the date of

triggering (class 2); both of them can be useful to analyse landslide events distribution and hence for landslide hazard estimation.

In 10 years, in Italy, 184,322 articles related to landslides have been released by online newspapers; among them, 78550 articles referred to 13275 recent landslide events (class 1).

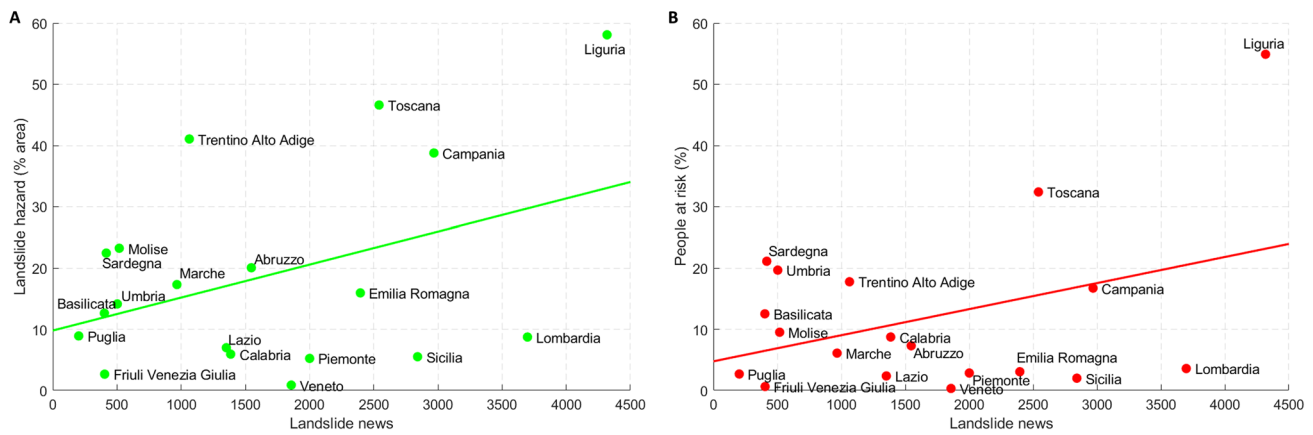


Figure 10 Correlation between the number of landslide news (classes 1 + 2 and the percentage of landslide hazard area (A) and the percentage of people at risk (B)

A textual analysis was conducted to get the frequency of words within headlines. Inside of the class 1 news, the majority of words refer to synonyms of the word “landslide”; within class 2, the most widely used words are referred to hazard, alert, weather forecast or anyway to past or future events without useful information about recent landslide events, while the words inside the class 3 are the result of a wrong association of word. However, in both categories (classes 1, 2, 3), the word that most widely used is “frana”, which is the Italian word for landslide, but it also has several figurative meanings, not referred to landslide events, that are often used in different contexts as sport or politics. The other commonly used word is “road”; this word is present in classes 1 and 2 with a high frequency, since landslides involving the road lead to a higher media coverage (high number of class 1 news) than landslides involving inhabited areas (e.g. forest), as well as the remediation work needed to restore the damaged roads (high number of class 2 news).

The landslide database analysis allowed to define a spatial and temporal distribution of landslide events. Considering only class 1 articles, the events are mainly present along the Alps and the Apennines.

The regions with more “landslide events” are mainly located in the northern part of Italy where the geological, geomorphological and climatic context of the Alps, along with permafrost melting and frost-thaw cycles, lead every year to several landslides (Giardino et al. 2004; Ratto et al. 2007; Cignetti et al. 2016).

Several areas along the Apennines are also highly involved with landslides.

Liguria is the region that shows the highest number of events and days associated with landslides; its territory, in fact, is characterized by steep slopes with few flat areas along the coast and in the valleys. These areas are very urbanized, that, in combination with the land use, involve a geomorphological evolution characterized by a high presence of landslide and flood events.

The areas less involved are located along North-East coast and in Puglia because they are mainly flat areas, most likely to be affected by other geo-hazards, as floods.

The temporal distribution of the articles and therefore of the landslide events increases from 2015 to 2019; the average number of days with landslides increased from 3 in the period 2010–2014 to 5 in the period 2015–2019. For this reason, it could be due to several aspects as the increasing number of high-intensity rainfall events or land use changes (Crozier 2010). Some authors have related the global climate change to a rise of the global temperature with a more frequent occurrence of extreme events in general (Rebetz et al. 1997; Easterling et al. 2000; Rosenzweig et al. 2008; Knight and Harrison 2009; Keiler et al. 2010) such as intense and localized precipitation. Theoretically, all these climatic parameters may influence the pre-conditions and triggering mechanisms of landslides and hence may lead to an enhanced frequency of landslides in general (Beniston and Douglas 1996). Furthermore, the inaccurate land use management can lead to the increase of mass movements in the whole Italian territory over time (ISPRA 2020a).

Landslide events have a certain seasonal distribution during the 10-year observation period. Indeed, during wet season (from October to April), the landslide events are more frequent since fall and winter are the rainiest period of the years.

Vice versa the frequency is lower during the dry season (from May to September), even if isolated landslide events can be found, usually related to severe storms that strike small areas (few tens of square km).

These results are in agreement with literature works where seasonal distribution of landslide was investigated in Campania (Cascini et al. 2014) and Toscana (Rosi et al. 2012) regions or at national level (Guzzetti et al. 2005; Calvello and Pecoraro 2018)

A relevant number of landslides have also been reported in February and March: These months coincide with the end of the winter and a rise in temperature, associated to snowmelts, which is a well know landslide triggering factor in Italy (Cardinali et al. 2000).

The year 2019 is the year with the highest number of “landslide events” and involved days; according to ISPRA (2020b), the mean cumulative rainfall of this year was 12% higher than the mean over the 1961–2019 period, and autumn and springtime were 47% and 19% more rainy than usual. The years 2013 and 2014 have a high

number of days with landslide news, because they present several articles, distributed over a long-time interval, about the Mont de La Saxe landslide in Valle d'Aosta Region. The Mont de La Saxe landslide is a rock fall type landslide threatening a valley with buildings, streets and a river (Giordan et al. 2015). It suffered several reactivations over the time that caused damages or lead to road closures, and each time new articles have been published, and more days with landslide news have been recorded.

The distribution of news (classes 1 + 2) shows a certain correlation with the percentages of landslide hazard areas and of people at risk. Liguria region has a high percentage of hazardous area (58%) and a very high percentage of people living in landslide risky zones. While Lombardia has a very high number of news, but low percentages of territory subject to landslide hazards and of people at risk; this is due to the fact that landslide are concentrated along the Alpine arc (see Fig. 5B), in the northern part of the region, where there is a lower population rather than in the southern, plain part.

Even if the used approach gave the opportunity of taking a picture of landslide distribution in Italy, it is worth to notice that mass media attention is not uniformly distributed across disaster-affected areas (Fan et al. 2020). The classification was necessary since each landslide event can be reported from one or more newspaper based on its impact of the relevance of the affected area; for example, small landslides involving a major road or an important city can have a vast media echo, while landslides involving minor roads or small villages are reported only by local newspapers. In some cases, the presence/absence of news could be affected by other factors such as disruption in communication services, socio-demographic factors (the events affecting socially vulnerable populations get less attention) and absence of points of attraction. Furthermore, a landslide can have more reactivates (see La Saxe landslide) during time and therefore more articles published. These factors can alter the real distribution of landslide hazard, leading from one hand to underestimate the presence of landslides in rural areas, forest or without a journalistic relevance and to the other hand to overestimate the hazards in most relevant areas from a journalistic point of view.

One last observation must be done on the spatial resolution of the used data. Since online newspapers are the used source of info, the accuracy of some parameters could be low, in fact inside newspaper articles some technical details, as the type of landslide, its dimension or volume are often missing; the exact location of the landslide is a parameter rarely available from this kind of source as well. For these reasons, the use of newspaper articles may be useful for analyses over large areas, but not to create detailed landslide inventory or for detailed analyses.

Conclusion

Newspaper articles inside Internet or crowdsourcing platforms can be regarded as a constant and continue source of information about a recent landslide with a high impact and consequences in terms of loss of infrastructure and human lives.

In this work, a spatial and temporal distribution of landslide events in the Italian territory has been presented. These analyses have been carried out using online landslide news harvested by SECaGN algorithm from 2010 to 2019. The news database was classified in three classes on the basis of news relevance, localization accuracy and time of publication so as distinguishing the “news

referred to recent landslide events” (class 1) from “news generically referred to landslides” (class 2) and “news not related to landslides” (class 3). This classification allowed to define, at national scale, the areas and periods mainly involved by landslide events. Around 41% of the news reported information relative to recent landslide events, and only 2% of the database is made up by wrong news. Through a semantic analysis, it has been possible to check the words with the highest frequency inside the headlines of newspapers. This allowed to define the principal words that describe the landslide events, which, in turn, can be used to properly tune up the data mining algorithm, limiting the news with wrong word association.

Based on the results of this work, it is possible to conclude that the events and news are increasing from 2015 to 2019 in the whole Italian context. November is the month with the highest amount of landslide events. Lombardia, Liguria and Campania are the regions that have the highest number of days characterized by landslide phenomena and consequently the highest number of news publication. Data also showed that ca. 42% of Italian municipalities have been affected by landslides in the observed.

More in general, this updated landslide inventory allowed to get the overall number of landslides events since 2010 to nowadays. Finally, this inventory can be used to get relationship between the more detailed news (class 1) and rainfall to create rainfall thresholds, but it can be also used for hazard and vulnerability assessment (class 1 and class 2 news).

This work showed that data mining is a reliable methodology to create a good landslide inventory in a relatively short time, even if with a coarser spatial accuracy than traditional inventories. Future developments could introduce the capability of being used for other hazards and in other languages, once appropriate dictionaries for the semantic engine will be available.

Author contribution

Text writing: R.F. and A.R. Data analysis: R.F. Conceptualization: A.R. and F.C. Manuscript reviewing: A.R. and R.F. Supervisor: N.C. and F.C.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Agliardi F, Crosta GB (2003) High resolution three-dimensional numerical modelling of rockfalls. *Int J Rock Mech Min Sci* 40(4):455–471. [https://doi.org/10.1016/S1365-1609\(03\)00021-2](https://doi.org/10.1016/S1365-1609(03)00021-2)
- Agostini A, Tofani V, Nolesini T, Gigli G, Tanteri L, Rosi A, Cardellini S, Casagli N (2014) A new appraisal of the Ancona landslide based on

- geotechnical investigations and stability modelling. *Q J Eng Geol Hydrogeol* 47:29–43
- Asheghi R, Hosseini SA, Saneie M, Abbaszadeh Shahri A (2020) Updating the neural network sediment load models using different sensitivity analysis methods: a regional application. *J Hydroinfr* 22(3):562–577. <https://doi.org/10.2166/hydro.2020.098>
- Battistini A, Segoni S, Manzo G, Catani F, Casagli N (2013) Web data mining for automatic inventory of geohazards at national scale. *Appl Geogr* 43:147–158
- Battistini A, Rosi A, Segoni S, Lagomarsino D, Catani F, Casagli N (2017) Validation of landslide hazard models using a semantic engine on online news. *Appl Geogr* 82:59–65
- Beniston M, Douglas GF (1996) Impacts of climate change on mountain regions. In: Watson RT, Zinyowera MC, Moss RH, Dokken DJ (eds) *Climate Change 1995 Impacts, Adaptations and Mitigation of Climate Change: Scientific–Technical Analysis*. Cambridge Univ. Press, Cambridge, pp 191–213
- Bianchini S, Raspini F, Solari L, Del Soldato M, Ciampalini A, Rosi A, Casagli N (2018) From picture to movie: twenty years of ground deformation recording over Tuscany region (Italy) with satellite InSAR. *Front Earth Sci* 6:177. <https://doi.org/10.3389/feart.2018.00177>
- Brunsdon D (1985) Landslide types, mechanisms, recognition, identification. In: CS Morgan (ed) *Landslides in the south Wales coalfield, proceedings symposium, the polytechnic of Wales*, pp 19–28
- Calvello M, Pecoraro G (2018) *Franelitalia: a catalog of recent Italian landslides*. *Geoenviron Disasters* 5(1):1–16
- Cardinali M, Ardizzone F, Galli M, Guzzetti F, Reichenbach P (2000) Landslides triggered by rapid snow melting: the December 1996–January 1997 event in Central Italy. *Mediterranean Storms (Proceedings of the EGS Plinius Conference held at Maratea, October 1999)*, Italy
- Carley KM, Malik M, Landwehr PM, Pfeffer J, Kowalchuck M (2016) Crowd sourcing disaster management: the complex nature of Twitter usage in Padang Indonesia. *Saf Sci* 90:48–61
- Cascini L, Sorbino G, Cuomo S, Ferlisi S (2014) Seasonal effects of rainfall on the shallow pyroclastic deposits of the Campania region (southern Italy). *Landslides* 11(5):779–792
- Cignetti M, Manconi A, Manunta M, Giordan D, DeLuca C, Allasia P, Ardizzone F (2016) Taking advantage of the ESA G-POD service to study ground deformation processes in high mountain areas: a Valle d'Aosta case study, Northern Italy. *Remote Sens* 8(10):852
- Crozier MJ (2010) Deciphering the effect of climate change on landslide activity: a review. *Geomorphology* 124:260–267
- Cruden DM, Varnes DJ (1996) Landslide types and processes. In: Turner AK, Schuster RL (eds) *Landslides — investigation and mitigation*. National Research Council, Transportation Research Board, National Academy Press, Washington DC, pp 36–75
- Cuesta MJD, Sánchez MJ, García, A. R. (1999) Press archives as temporal records of landslides in the North of Spain: relationships between rainfall and instability slope events. *Geomorphology* 30(1–2):125–132
- Devoli G, Morales A, Høeg K (2007) Historical landslides in Nicaragua—collection and analysis of data. *Landslides* 4(1):5–18
- Dikau R, Cavallin A, Jäger S (1996) Databases and GIS for landslide research in Europe. *Geomorphology* 15:227e239
- Easterling DR, Meehl GA, Parmesan C, Changnon SA, Karl TR, Mearns LO (2000) Climate extremes: observations, modeling, and impacts. *Science* 289:2068–2074. <https://doi.org/10.1126/science.289.5487.2068>
- Fan C, Mostafavi A, Gupta A, Zhang C (2018) A system analytics framework for detecting infrastructure-related topics in disasters using social sensing in *Advanced Computing Strategies for Engineering*. Springer, Cham, pp 74–91
- Fan C, Wu F, Mostafavi A (2020) A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access* 8:10478–10490. <https://doi.org/10.1109/ACCESS.2020.2965550>
- Fischer HW (1994) *Response to disaster: fact versus fiction and its perpetuation*. University Press of America, Washington, p 160
- Galli M, Ardizzone F, Cardinali M, Guzzetti F, Reichenbach P (2008) Comparing landslide inventory maps. *Geomorphology* 94:268–289
- Giardino M, Giordan D, Ambrogio S (2004) G.I.S technologies for data collection, management and visualization of large slope instabilities: two applications in the Western Italian Alps. *Nat Hazards Earth Syst Sci* 4:197–199 205
- Giordan D, Manconi A, Allasia P, Bertolo D (2015) Brief communication: On the rapid and efficient monitoring results dissemination in landslide emergency scenarios: the Mont de La Saxe case study. *Nat Hazards Earth Syst Sci* 15(9):2009–2017
- Görüm T, Fidan S (2021) Spatiotemporal variations of fatal landslides in Turkey. *Landslides* 18(5):1691–1705
- Goswami S, Chakraborty S, Ghosh S, Chakrabarti A, Chakraborty B (2016) A review on application of data mining techniques to combat natural disasters. *Ain Shams Eng J* 9:365–378. <https://doi.org/10.1016/j.asej.2016.01.012>
- Gundecha P, Liu H (2012) Mining social media: a brief introduction. *New Directions in Informatics, Optimization, Logistics, and Production*, pp 1–17. <https://doi.org/10.1287/edu.1120.0105>
- Guzzetti F (2000) Landslide fatalities and evaluation of landslide risk in Italy. *Eng Geol* 58:89–107
- Guzzetti F, Cardinali M, Reichenbach P (1994) The AVI project: a bibliographical and archive inventory of landslides and floods in Italy. *Environ Manag* 18(4):623–633
- Guzzetti F, Stark CP, Salvati P (2005) Evaluation of flood and landslide risk to the population of Italy. *Environ Manag* 36(1):15–36
- Guzzetti F, Peruccacci S, Rossi M, Stark CP (2008) The rainfall intensity duration control of shallow landslides and debris flows: an update. *Landslides* 5(1):3e17
- ISPRA (2020a) Land use, spatial dynamics and ecosystem services. Report 2020. (In Italian)
- ISPRA (2020b). Climatic indicators of Italy in 2018 (in Italian)
- Keiler M, Knight J, Harrison S (2010) Climate change and geomorphological hazards in the eastern European Alps. *Philos Trans R Soc* 368:2461–2479. <https://doi.org/10.1098/rsta.2010.0047>
- Kirschbaum DB, Adler R, Hong Y, Hill S, Lerner-Lam A (2010) A global landslide catalog for hazard applications: method, results, and limitations. *Nat Hazards* 52(3):561–575
- Kirschbaum D, Stanley T, Zhou Y (2015) Spatial and temporal analysis of a global landslide catalog. *Geomorphology* 249:4–15
- Klimeš J, Stemberk J, Blahut J, Krejčí V, Krejčí O, Hartvich F, Kycł P (2017) Challenges for landslide hazard and risk management in 'low-risk' regions, Czech Republic—landslide occurrences and related costs (IPL project no. 197). *Landslides* 14(2)
- Klose M, Damm B, Highland L (eds) (2015) *Geohazard databases: concepts, development, Applications locations [Special Issue]*. *Geomorphology* 249:1–136
- Klose M, Maurischat P, Damm B (2016) Landslide impacts in Germany: a historical and socioeconomic perspective. *Landslides* 13(1):183–199
- Knight J, Harrison S (2009) Sediments and future climate. *Nat Geosci* 2:230. <https://doi.org/10.1038/ngeo491>
- Kreuzer TM, Damm B (2020) Automated digital data acquisition for landslide inventories. *Landslides* 17:2205–2215
- Liddy ED (2001) Natural language processing. In: *Encyclopedia of Library and Information Science*, 2nd edn. Marcel Decker, Inc, New York
- 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. *Int J Remote Sens* 33(2):466e489
- McKean J, Roering J (2003) Objective landslide detection and surface morphology mapping using high-resolution airborne laser altimetry. *Geomorphology* 57(3e4):331e351
- Panizza M, Corsini A, Ghinoi A, Marchetti M, Pasuto A, Soldati M (2011) Explanatory notes of the geomorphological map of the alta Badia Valley (Dolomites, Italy). *Geogr Fis Din Quat* 34(2011):105–126
- Rachunok BA, Bennett JB, Nateghi R (2019) Twitter and disasters: A social resilience fingerprint. *IEEE Access* 7:58495–58506
- Ratto S, Giardino M, Giordan D, Alberto W, Armand M (2007) *Carta dei fenomeni franosi della Valle d'Aosta, scala 1: 100.000-Regione Autonoma Valle d'Aosta, Assessorato Territorio, Ambiente e Opere Pubbliche*
- Rebetz M, Lugon R, Baeriswyl PA (1997) Climatic change and debris flows in high mountain regions: the case study of the Ritigraben torrent (Swiss Alps). *Clim Chang* 36:371–389. <https://doi.org/10.1023/A:1005356130392>

- Rosenzweig C, Karoly D, Vicarelli M, Neofotis P, Wu Q, Casassa G, Menzel A, Root TL, Estrella N, Seguin B, Tyrjanowski P, Liu C, Rawlins S, Imenson A (2008) Attributing physical and biological impacts to anthropogenic climate change. *Nature* 453:353–357. <https://doi.org/10.1038/nature06937>
- Rosi A, Segoni S, Catani F, Casagli N (2012) Statistical and environmental analyses for the definition of a regional rainfall threshold system for landslide triggering in Tuscany (Italy). *J Geogr Sci* 22:617e629
- Rosi A, Tofani V, Tanteri L, Tacconi Stefanelli C, Agostini A, Catani F, Casagli N (2018) The new landslide inventory of Tuscany (Italy) updated with PS-InSAR: geomorphological features and landslide distribution. *Landslides* 15:5–19. <https://doi.org/10.1007/s10346-017-0861-4>
- Rosi A, Canavesi V, Segoni S, Dias Nery T, Catani F, Casagli N (2019) Landslides in the mountain region of Rio de Janeiro: a proposal for the semi-automated definition of multiple rainfall thresholds. *Geosciences* 9:203
- Rosi A, Segoni S, Canavesi V, Monni A, Gallucci A, Casagli N (2021) Definition of 3D rainfall thresholds to increase operative landslide early warning system performances. *Landslides* 18(3):1045–1057
- Rosser B, Dellow S, Haubrock S, Glassey P (2017) New Zealand's national landslide database. *Landslides* 14(6):1949–1959
- Saltelli A, Bammer G, Bruno I, Charters E, Di Fiore M, Didier E, Espeland WN, Kay J, Lo Piano S, Mayo D, Pielke R Jr, Portaluri T, Porter TM, Puy A, Rafols I, Ravets JR, Reinert E, Sarewitz D, Stark PB et al (2020) Five ways to ensure that models serve society: a manifesto. *Nature* 582:482–483. <https://doi.org/10.1038/d41586-020-01812-9>
- Salvatici T, Tofani V, Rossi G, D'Ambrosio M, Tacconi Stefanelli C, Masi EB, Rosi A, Pazzi V, Vannocci P, Petrolo M, Catani F, Ratto S, Stevenin H, Casagli N (2018) Application of a physically based model to forecast shallow landslides at a regional scale. *Nat Hazards Earth Syst Sci* 18:1919–1935. <https://doi.org/10.5194/nhess-18-1919-2018>
- Santangelo M, Cardinali M, Rossi M, Mondini AC, Guzzetti F (2010) Remote landslide mapping using a laser rangefinder binocular and GPS. *Nat Hazards Earth Syst Sci* 10:2539–2546
- Soeters R, Van Westen C (1996) Slope instability recognition, analysis and zonation, in *Landslides, investigation and mitigation*. Transportation Research Board, National Research Council, National Academy Press, Washington, p 129e177
- Solari L, Bianchini S, Franceschini R, Barra A, Monserrat O, Thuegaz P, Bertolo D, Crosetto M, Catani F (2020) Satellite interferometric data for landslide intensity evaluation in mountainous regions. *Int J Appl Earth Obs Geoinf* 87:102028. <https://doi.org/10.1016/j.jag.2019.102028>
- Taylor FE, Malamud BD, Freeborough K, Demeritt D (2015) Enriching Great Britain's National Landslide Database by searching newspaper archives. *Geomorphology* 249:52–68. <https://doi.org/10.1016/j.geomorph.2015.05.019>
- Trigila A, Iadanza C, Bussetini M, Lastoria B (2018) Landslides and floods in Italy: hazard and risk indicators. *ISPRRA, Rapporti* 287(2018):172 (In Italian)
- Vai F, Martini IP (2001) *Anatomy of an Orogen: the Apennines and adjacent Mediterranean basins*, XVIII edn. Springer Netherlands, Netherlands, p 633. <https://doi.org/10.1007/978-94-015-9829-3>
- 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
- Vennari C, Gariano SL, Antronico L, Brunetti MT, Iovine G, Peruccacci S, Terranova O, Guzzetti F (2014) Rainfall thresholds for shallow landslide occurrence in Calabria, southern Italy. *Nat Hazards Earth Syst Sci* 14:317e330. <https://doi.org/10.5194/nhess-14-317>

Rachele Franceschini · Ascanio Rosi (✉) · **Nicola Casagli**

Department of Earth Sciences, University of Florence, Via Giorgio La Pira, 4, 50121 Florence, Italy
Email: ascanio.rosi@unifi.it

Filippo Catani

Department of Geosciences, University of Padova, Via G. Gradenigo, 6, Padova, Italy