Universität für Bodenkultur Wien
University of Natural Resources
and Life Sciences, Vienna

Master Thesis

Classification of torrential behaviour of torrents in the Province of Trento

Submitted by

Valentino GOTTARDI, BSc

in the framework of the Master programme

Alpine Naturgefahren/Wildbach- und Lawinenverbauung

in partial fulfilment of the requirements for the academic degree

Diplom-Ingenieur

Vienna, February 2022

Supervisor:

Univ.Prof. Dipl.-Ing. Dr.nat.techn. Johannes Hübl
Institute of Mountain Risk Engineering
Department of Civil Engineering and Natural Hazards

Statutory declaration

I hereby declare that I am the sole author of this work. No assistance other than that which is permitted has been used. Ideas and quotes taken directly or indirectly from other sources are identified as such. This written work has not yet been submitted in any part.

Trento, 27/02/2022

Valentino Gottardi

Acknowledgements

I have finally achieved this important goal. Study and live abroad is not always easy but surely it is a great experience. I have to share the credit of this success with lots of people.
First of all I'd like to thank my entire family for supporting and encouraging me during this period of my live.
Thanks to Prof. Johannes Hübl for his openness to this collaborative work between the Institute of Mountain Risk Engineering and the Servizio Bacini Montani. Thanks to Dipl.-Ing. Micha Heiser for the helpful and expert suggestions. Thanks to the Servio Bacini Montani of the Autonomous Province of Trento and particularly to dott. Ruggero Valentinotti and dott. Gabriele Bertoldi for the continuous helpfulness and the valuable inputs.
I want also to thank old and new friends for accompaning me on this study period. In particular Kilian for the kind help and hospitality.

Abstract

The aim of this thesis is the development of a classification model of small alpine catchments located in the Autonomous Province of Trento. The classification is based on the probability that intense torrential phenomena would be generated (i.e., debris flows or hyper concentrated currents).
To do this, I used both the Provincial databases relating to the watershed-fan system and dejection fans and the cadaster of torrential and fluvial events. On the one hand, the first represents graphically, and through descriptive parameters of geomorphological type, 1697 binary basin-fan systems distributed over the entire Provincial territory. On the other hand, the second one reports 3522 events - collected and described according to various sources and with varying reliability - which took place over a period of time ranging from 1570 to 2016.

In this dissertation, all the fluvial and torrential events characterized by good reliability were extracted using the *R* software. Furthermore, this subset has been manually inspected in the *QGIS* environment. Spatial selection of signals located on the fans area followed this manual inspection.
For the construction of the model, this thesis proceeded with the selection of the basin-fan binary systems characterized by a very precise behavior. In this study, 103 of the extracted basins generated at least one debris flow - or debris flood in the fan area, and no alluvial phenomenon. While 36 basins, on the other hand, generated only bedload transport - or clear water discharge events, but no intense phenomena on their own fan.

The selection of the basins' most descriptive variables was carried out in *R* environment, applying the Principal Component Analysis. This approach was used to maintain all the starting variables and to avoid the loss of a considerable amount of information, which is inevitable in the manual selection process of highly correlated variables.
Given the small number of available basins, the bootstrapping resampling technique was used to create theoretical model's train and test samples.

Once the train and test samples were built, the model could be trained. A Logistic Regression was applied using only the first 6 main components in order to simplify the input parameters without losing excessive data variability. A $0-1$ or *unlikely* - *likely* binar variable is the output of the proposed model. To refine the predictive capacity of the model, the prediction threshold was chosen, in order to maximizes the parameter called Youden's J.

Some prediction models available in the literature have been applied to the same database in order to better evaluate the predictive capabilities of the proposed numerical model. Then, the results of both the comparison and of the proposed model were displayed using tables and graphs.

Zusammenfassung

Ziel dieser Masterarbeit ist es ein Verfahren zur Einstufung kleiner alpiner Wildbacheinzugsgebiete in der Provinz Trient zu entwickeln.
Die Klassifizierung basiert auf der Eintrittswahrscheinlichkeit von starken Wildbacherereignissen (z.B. Murgänge). Dies wurde durch die Anwendung der Datenbank der Provinz Trient für die Kartierung der Wildbacheinzugsgebiete und der Kegel sowie des Katasters der geologischen und hydrologischen Ereignisse ermöglicht. Die Kartierung stellt landesweit 1697 duale Systeme Einzugsgebiet-Kegel graphisch als auch nach deskriptiven Parametern dar. Das Kataster zeichnet 3522 Ereignisse auf, welche zwischen dem Jahr 1570 und 2016 vorgekommen sind.

Durch das Softwareprogramm *R* wurden von Wildbachprozessen verursachte Ereignisse mit statistischer Zuverlässigkeit ausgewählt. Das Programm *QGIS* wurde danach angewendet, um die ausgewählten Ereignisse geographisch einzuordnen. Nur jene, welche sich auf einem Schwemm- oder Murkegel ereigneten, wurden für die weitere Datenbearbeitung in Betracht gezogen.

Das Modell wurde mittels 139 Einzugsgebiete konstruiert und getestet. 103 Einzugsgebiete zeigen mindestens einen Murgang auf Kegelbereich und kein Hochwasser. Nur reines Hochwasser oder Geschiebetransport treten im 36 Einzugsgebieten auf.

Die Auswahl der deskriptiven Variablen der Einzugsgebiete erfolgte durch den Einsatz der Principal component analysis im Softwareprogramm *R*, sodass, gegenteilig zum manuellen Verfahren, die gesamten Informationen der Variablen mit hoher Kollinearität beibehalten wurden.
Angesicht der kleinen Anzahl an verfügbaren Einzugsgebieten wurden die zwei Trainings- und Testmuster des numerischen Modells mit Einsatz des Verfahrens Bootstapping with replacement kreiert.
Die erste 6 Principal components wurden für die logistische Regression angewendet. Output des Modells ist eine 0-1 oder *unwahrscheinlich-wahrscheinlich* Variable. Alternative Cutoff-Wert und Tuningparameter wurden ausgewählt um die Youden‘s J zu maximieren.

Das Risikoeinschätzungspotential des numerischen Modells wurde anhand von Daten anderer Modelle aus der Literatur getestet. Die daraus erfolgten Ergebnissen wurden sowohl in tabellarischer als auch in graphischer Form dargestellt.

Sommario

L'obiettivo di questa tesi è quello di sviluppare un metodo di classificazione dei piccoli bacini torrentizi alpini ubicati in provincia di Trento. La classificazione si basa sulla probabilità che vengano generati fenomeni torrentizi intensi (colate detritiche o correnti iperconcentrate).
Per fare ciò sono stati utilizzati i database provinciali relativi alla mappatura di bacini torrentizi e conoidi di deiezione ed il catasto degli eventi geologici e fluviali. I primi rappresentano graficamente e tramite parametri descrittivi di tipo geomorfologico 1697 sistemi binari bacino-conoide distribuiti sull'intero territorio provinciale. Il secondo riporta 3522 eventi raccolti e descritti secondo varie fonti e con varia attendibiltà a partire dall'anno 1570 fino al 2016.

Utilizzando il software *R* sono stati estratti gli eventi riferibili a processi torrentizi caratterizzati da una buona attendibilità. In ambiente *QGIS* è stato eseguito un controllo manuale delle singole segnalazioni e una successiva selezione spaziale dei soli eventi posizionati in area di conoide. Per la costruzione del modello si è proceduto alla selezione dei sistemi binari bacino-conoide caratterizzati da un comportamento ben preciso. In questo modo sono stati estratti 103 bacini che hanno generato almeno un fenomeno di colata detritica o di corrente iperconcentrata in area di conoide e nessun fenomeno alluvionale. 36 bacini hanno invece generato solamente trasporto di fondo o piene liquide ma nessun fenomeno intenso sul proprio conoide.

La selezione delle variabili descrittive dei bacini è stata effettuata in ambiente *R* utilizzando la Principal component analysis in modo da non perdere una notevole quantità di informazione, inevitabile nel processo di selezione manuale di variabili caratterizzate da elevata collinearità.
Dato il ridotto numero di bacini disponibili, per creare i due campioni di training e test del modello numerico si è impiegata la tecnica di resampling del bootstrapping.
Una volta costruiti i campioni di training e test, è stata applicata una regressione logistica utilizzando solamente le prime 6 componenti principali in modo da semplificare i parametri di ingresso senza perdere una eccessiva variabilità del dato. Il risultato dell'analisi è una variabile binaria del tipo $0-1$ o improbabile-probabile. Per perfezionare la capacità predittiva del modello è stata scelta la soglia di predizione che massimizza il parametro denominato Youden's J.

Per meglio valutare le capacità predittive del modello numerico proposto sono stati applicati ai dati disponibili alcuni modelli previsionali reperibili in letteratura e spesso utilizzati in ambito alpino. I risultati del confronto e del modello proposto sono stati poi rappresentati in forma grafica e tabellare.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Watershed management

The watershed management history in Tirol dates back to the 13^{th} century. Already in 1277, damages on the protection wall of the Talfer torrent by Bozen were mentioned. While in 1400 the construction of the shore walls along the Fersina torrent began, financed by one third of the duties on forest products.
One of the milestones of torrent control in the Alps is represented by the construction of the *Pont'Alto* check dam in the 1537, commissioned by the Prince bishop Cardinal Bernado Clesio to protect the city of Trento. This is an important step in the management of torrential processes: it was the first time that a mitigation structure was design to operate in the upper part of a torrent reach. Till that point, all the efforts posed served only to passively protect houses, infrastructures and agricultural areas.
Since its construction the Pont'Alto check dam was destroyed many times by the Fersina torrent, and the last reconstruction was performed in 1850, with hight of about 40 m. The stability of the check dam was ensured with the construction of the *Controserra Madruzza*, a dam at the foot ot the old one, concluded in 1882.

In the territory of the Habsburg Monarchy, after the destructive flooding events of both $16^{th} - 20^{th}$ September and 27^{th} October 1882 began, the establishment of a *forest-technical service* (*k.k. forsttechnische Abteilung für Wildbachverbauungen*), completed in 1884. As has already been done in France, this service adopted an integral watershed management by the combination of forestry and technical measures. In this way the protection concept were extended to the whole catchment area, by applying an intergal management (BM für Land- und Forstwirtschaft, 1984).

Not only floods but also mass movements like landslides, rock falls and rock avalanches occured in the past and are still likely to occure nowadays too. In many provincial areas, avalanches threaten both humans and infrastructures. For example during the White Friday on 13 Dec. 1916, various avalanches caused a lot of casualties between the soldiers on the dolomitic front. Figure 1.1 shows an original photo of an avalanche occured in Vermiglio (Neue Zürcher Zeitung, 12 December 2016).

Figure 1.1: 13 Dec. 1916, Vermiglio (source: K. u. k. Kriegspressequartier, Lichtbildstelle–Wien, Österreichische Nationalbibliothek

Another event to be mentioned is the flood that took place in 1966. It happened on 4 November 1966 and produced significant damages expecially to Trento and the province's eastern part. Moreover, damages were reported in many other Italian cities. The heavy consequences on the provincial economy and consequently the diffuse instability of the territory lead to a strong mitigation measure's construction campaign, which was performed until the 80's.

Nowadays, the main objective of the provincial services is to apply new scientific knowledge in order to optimize watershed management efficiency.
Since 2006, the maintenance and control of the entire water network - rivers, torrents and lakes included in the public water list - glaciers and all mitigation measures were tasks of the Service for Torrent Control (*Servizio Bacini Montani - SBM*), which is part of the Department for Forest and Mountain Resources (*Dipartimento Risorse Forestali e Montane*) of the Autonomous Province of Trento (*Provincia Autonoma di Trento - PAT*).
Counting administratives and construction workers, this service employs more than 300 people. The Service is divided into four District Offices, responsible for the corresponding territorial zones. Panel 1.2a of Figure 1.2 shows the organogram of the Service, while panel 1.2b depicts the spatial extension of the four District Offices (modified from SBM presentations).

The present work is made thanks to the collaboration of dott.for. Ruggero Valentinotti and dott.for. Gabriele Bertoldi of the Planning Office, Technical Support and Public Water Property (*Ufficio Pianificazione, Supporto Tecnico e Demanio Idrico*) of the SBM.

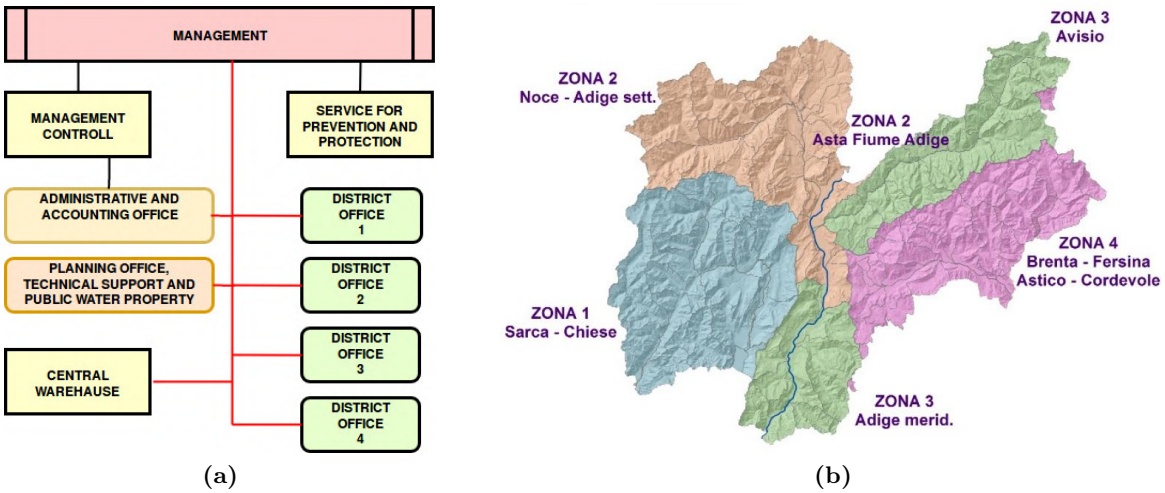


Figure 1.2: SBM's organogram and District Offices

1.2 What is an alpine torrent?

Many features have to be described in order to define an alpine torrent. A general but comprehensive definition is given in Bergmeister et al. (2009): an alpine torrent is a natural stream, with durable or temporary water flow. Its course is at least locally steep and the discharge is strongly variable. Erosion of sediment is performed by short and intense floods with high transport capacity. Erosion take place on the the entire watershed and along the thalweg. The sediment deposition can be done both outside or inside the torrent banks and to the main river. A general torrent basin is formed by a drainage basin, the area which collect the precipitation, and a depositional fan, where the water mass and the transported sediments are deposited.
Everything that happens in a torrent is due to the quick hydrological response of the basin and thus to the sudden increase of the flood wave. Related to this are the processes of sediment mobilization, transport and resulting deposition.
An example of typical alpine catchment is shown in Figure 1.3.

1.2.1 Classification of alpine torrents

There are several ways to classify torrent basins according to various parameters. Amongs the others, in the Middle-European literature, the following parameters are considered: torrent activity, location, sediment availability and origin and main process type.

Torrent classification depending on temporal origin of the sediment sources

The following geological classification, proposed by Stiny (1910, 1931) is still commonly used. The characteristic considered by it is the formation period of potential sediment sources. Following this approach, is possible to distinguish between:

- *Old sediment torrent*: the transported sediment was originated during past valley and torrent bed formation. The deposition of the erodible sediments sources is completely finished and an increase from external sources will not take place. The sediment sources are generally

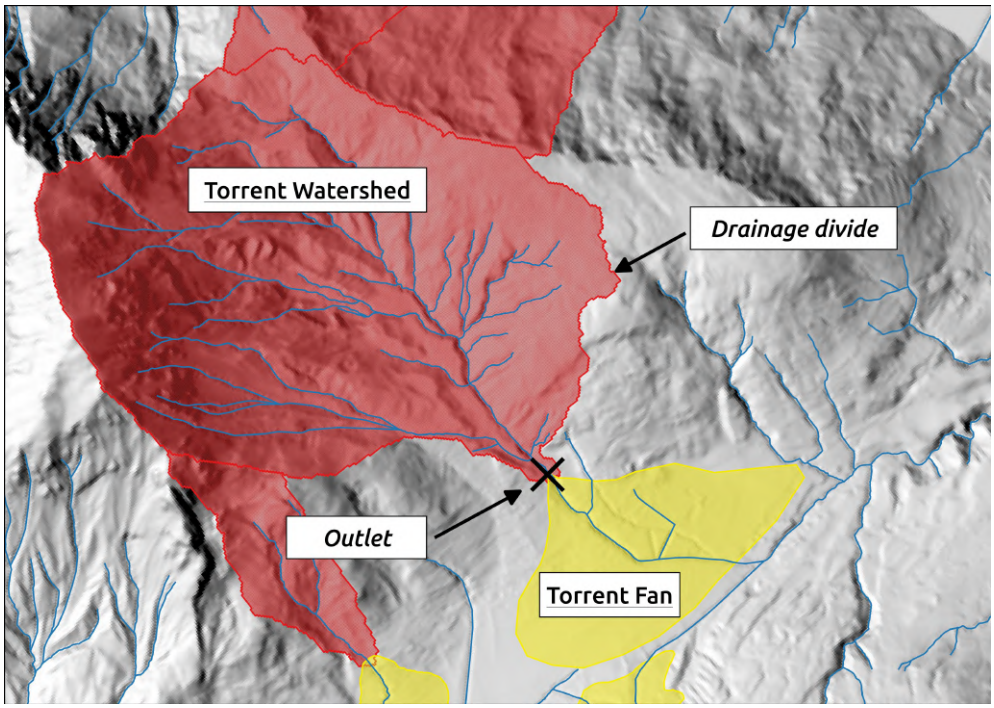


Figure 1.3: Schematic representation of a torrent basin

moraine, old thalweg sediments, fluvioglacial deposits, old rock avalanche deposits, stabilized slope deposits, fossile fan, etc..

- *Young sediment torrent*: the transported sediment is continuously or sporadically formed by rock weathering. The sediment formation is performed after the cutting of the valley. This torrent type is common in the region where limestone is the main lithological formation.

- Stiny proposed also the *mixed sediment torrent* group, where both sediment sources are represented. Torrent with different features are classified as *special torrent*.

Torrent classification depending on main transport process

Aulitzky (1984) proposed a 2-steps torrent classification for the evaluation of the general torrent behaviour at the fan or to a defined point into the lower watershed. In order to performe the classification, the available sediment sources have to be described. Fundamental requirement is the identification of a sufficient number of geomorphological features or silent witnesses (Stummen Zeugen). In the first step, it divides the torrents into the following four types, with a related common process behaviour at the fan:

- torrents generating debris flow with very high peak discharge (*Murstoßfähige Wildbäche*);

- torrents dominated by debris flow (*Murfähige Wildbäche*);

- torrents dominated by bedload transport (*Geschiebeführende Wildbäche*);

- torrents generating only clear water flood and suspendend load (*Nur hochwasserführende Wildbäche*).

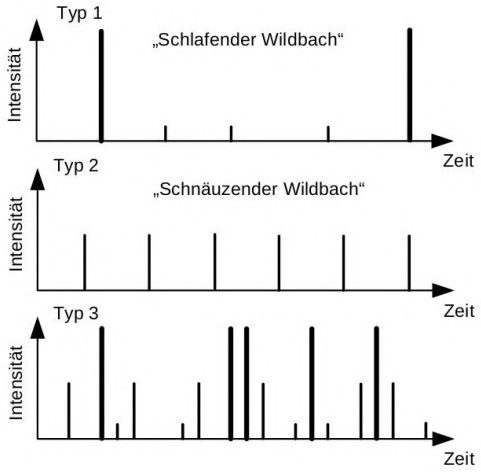


Figure 1.4: Schematic representation of event frequency for torrent basins (from Hübl, 2006)

While the second step aims to classify erosion type and tendency in order to evaluate the torrent dynamics.

Torrent classification depending on event frequency

Each torrent is characterized by different base and variable disposition and this leads to very different torrent behaviour. The diagrams depicted in Figure 1.4 (from Hübl, 2006) show, for a certain torrent basin, the reported events on a temporal scale. The bars height represent event intensity.
The upper diagram is related to a torrent with low sediment production and soil with high water storage capacity. Given these features, the torrent reacts with very high intensity, but only when hit by extreme precipitations (Type 1, *schlafender Wildbach* or sleeper torrent).
When the sediment production is low to medium and the soil's water storage capacity is quite low, sediments can be mobilized and transported already by precipitation with medium intensity. This results in more or less equal time-spaced events with low-to-medium intensity (Type 2, *schnäuzender Wildbach*).
In the case where sediments have variable production time and where the variable diposition changes strongly during the years, the events appear with uneaven time intervals and fluctuating magnitude (Type 3).

Torrent classification depending on basin morphology

Salzer (1886) proposed the torrent classification depending on basin morphology. As depicted in Figure 1.5 (from Bergmeister et al., 2009), **alpine regions'torrents** are commonly characterized by contiguous watershed and fan areas. Where the valley bottom is positioned under a morphological step (such as vertical or subvertical rock walls of valleys with glacial origin), the depositional fan is connected to the upper part through narrow and deep canyons or waterfalls. In any case, it is easy to define the area of sediment erosion from the deposition zone (Schneuwly-Bollschweiler et al., 2012). The torrent path is generally short and steep and, in many cases, thalweg and

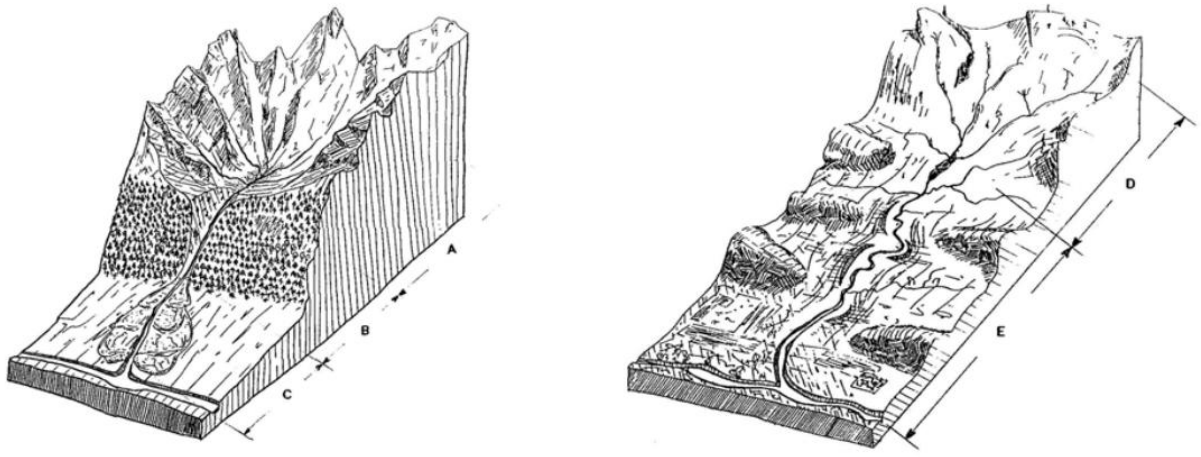


Figure 1.5: Torrent classification according to Salzer (1886) (modified from Bergmeister et al., 2009). Torrent of the alpine regions (left): A = drainage area, B = canyon, C = deposition area. Torrent of the mountaneous and hilly regions (right): D = drainage area, E = transport zone.

slopes are covered by erodible sediments. The discharge changes strongly and, after periods of precipitation absence completely dry thalwegs are common. The discharge rise quickly during storms or snow melting leading to very intense processes like debris flows.
Different arrangements are to find in **torrents of mountaneous and hilly regions**, where the torrent path is generally long and with low slope. Here, a well-defined depositional fan is substitute by an elongated transport zone. In this last part, the water flow repeatedly erodes and deposits sediments, while the thalweg divide itself into various channels or builds meanders. Thus, a sharp distinction between sediment production area and sediment deposition zone is difficult to made.
The water discharge is normaly steady and abundant. After durable and strong precipitation, the flow rate became imposing and can bring to flooding of the valley bottom with a short return intervall. This torrent type is typical of pre-alpine areas.

1.2.2 Torrent watershed

Depending either on natural or human made features of the upper part of a torrent, the drainage basin can be classified into hydrographical or hydrological drainage basin. The first type is delimited by geomorphological forms such us ridges and crests - called either drainage divides or watershed boundaries - which direct the surficial flows toward the pour point or outlet. If the geology highly affects the flow of surficial and underground water, the extension of the actual drainage basin can strongly diverge from the hydrographical basin. In this case it is called hydrological drainage basin. Moreover, human made water abstraction or import can produce dimensional variations.

The torrent's response to a certain precipitation is graphically represented by Figure 1.6 with an **hydrograph**, which shows precipitation and discharge to a certain gauging station on the same graph. Many influencing factors modify the form of the hydrograph, both in height and

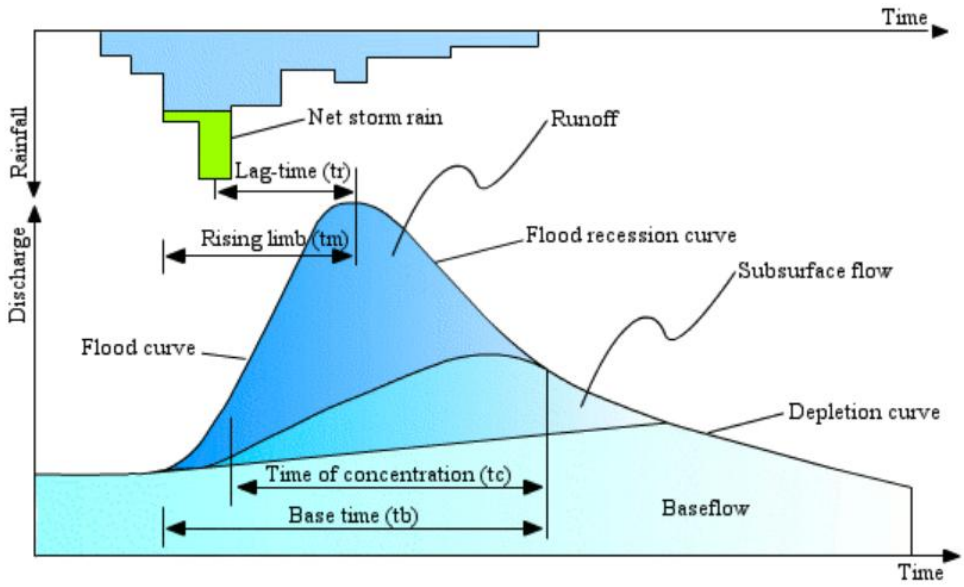


Figure 1.6: Example of a storm hydrograph (source: http://echo2.epfl.ch)

persistence.
The shape of the drainage basin together with the drainage density, drived by watershed geology and morphology, are important factors for the definition of the hydrograph, as well as land use and vegetation cover. This is due to the different flow velocity of surface runoff between slope and channel reach. For the description of watershed form several indexes have been proposed (e.g Gravelius compactness coefficient, Form factor, etc.).
For a circular drainage basin, the torrent's hydrograph can often be described as *flashy* because it will have a fairly steep rising limb and a high peak discharge. This is because all points in the drainage basin are roughly equidistant from the channel, thus all the precipitation reaches the torrent at the same time.

When considering the size, in respect of small basins, large watersheds will have low peak discharges, because they catch more precipitation, but at the same time they will have longer lag times because the water takes longer to reach the main channel.
Basins with steep slopes will have a high peak discharge and a short lag time since the water can travel downhill faster.
Finally, the drainage density of a basin will affect the lag time and the steepness of the falling limb. Basins with lots of streams and rivers (a high drainage density) will have a short lag time and a fairly steep falling limb because water will drain out of them quickly.

Both the prevalent soil and the rock type influence heavily the discharge formation. Non-porous and impermeable rocks develop high peak discharge and a short lag time, since impermeable rocks won't let water percolate through them, and thus forcing the water to travel via overland flow. Permeable rocks shape more gentle hydrograph because water can travel also via groundflow. Unconsolidated soils allow water to infiltrate and to be stored in the water-bearing layer or aquifers.

In order to understand the potential discharge formation, soil variability during the year, extremely dry or moist, as well as frozen soils have to be considered.

Precipitation intensity and persistence will strongly affect hydrograph's shape. Orographic storms are commonly considered as the most important driving factor in the intense torrential events. They are generally characterized by high intensity and short duration, and also the spatial distribution is uneven or localized. Hail and snow precipitations further alterate torrent response, with delayed and, in some cases, enlarged discharge.
The vegetation cover is responsible of interception, which is by low intensity precipitation a very important process. Vegetation can also disperse water into the air through transpiration and evaporation, and in order to assess the effectiveness of rain interception, water evaporation and transpiration, the vegetation type is crucial.
Sealing wide areas with the use of concrete and tarmac increases surface runoff and reduces the amount of stored water, increasing peak discharge and reducing lag time. Furthermore, the drainage density is strongly affected by artificial channels and pipes. Expecially for the alpine regions, alteration of the original slope morphology (for example for the creation of ski slopes) can modify the original torrent's hydrograph.

1.2.3 Torrent fan

As already mentioned, after sediment and water have been collected from the watershed and transported through the channel network, they arrive to the fan. A torrent fan can develop where the channel slope get smaller than a certain threshold. In many cases this happen at the torrent stream's confluence to the main valley, or at the foot of a geomorphological step. The slope reduction, together with the decrese of flow confinement, leads to the temporary or permanent deposition of transported sediment, which can be selective or not, in relation to the transport process.
Focusing the attention only to torrents, debris flow and bed load transport events edify the depositional fans. As explained in the following paragraphs these two types of fan are fairly different. From this classification debris fans resulting from deposition of avalanches are excluded, although many torrential valleys are also avalanche paths, and depositional cones of rock falls and rock avalanches.

As shown in Figure 1.7a a torrent fan has generally an half-circle form and can be divided into several parts. The transversal section is convex, whereas the longitudinal section can be concave, with constant inclination or convex. The transition zone, between watershed and fan, is called *fan apex* (Eisbacher and Clague, 1984). In the case that the sediment deposition take place upstream of the geomorphological step and the fan expand along the channel, is possible to define a *fan neck*. A *valley course* is present if the torrent flows over the valley bottom before reaching the river junction. The discrimination between *proximal and distal fan area* or *upper and lower fan segment* is sometimes important because of the different transport processes behaviour and thus the present depositional features.

Of major interest when dealing with torrential processes is the knowledge of the fan's development stage. It is important because the morphology of a fan can drastically be modified by extreme events of the torrent's basin itself or from the valley river. In many cases, fans experienced influences from neighbour torrents and rivers, expecially in narrow valleys. Moreover, the construction of major infrastructures - like highways, railways or big structures like reservoirs - strongly alterate the natural fan development. Considering the lack of land available for construction and the high population density, this is particularly true for the Alps.
Hübl (2006) describes some typical evolution paths of torrent fans. It is possible to recognize the natural evolution forms (see Figure 1.7b) only where human activity did not modify too much the natural fan morphology.

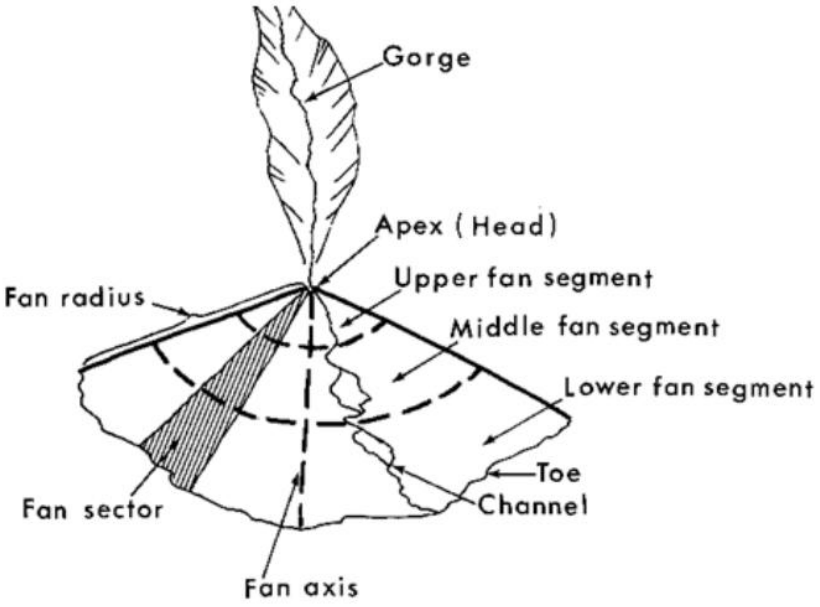
When only the situation within the torrent basin is considered, is possible to state that periods with high sediment mobilization lead to a general rise of the fan channel bed or fan surface. Periods with low sediment supply from the watershed generates localized erosion processes at the fan.
The main driving factors in the evolution of a fan are sediment supply, from the drainage area, and sediment erosion, from the valley main channel. A fan in dynamic equilibrium presents no significant fan channel deepening or sediment deposition.

Depending on basin geology, the sediment supply can be either steady during long periods or vary within relative short time windows. A classical example of increased sediment availability is the period immediately after a major flood event, wildfire or wind storms, which can destabilize not only channel bed and banks but also entire valley slopes. While another example is the current permafrost thawing and glacier melting, which deliver loose material to the initiation zones.
Mitigation measures likes consolidation dams retain a big amount of debris, with the result of an enlarged erosion capacity of the downhill torrent reach. The same statement is valid for big impoundments of hydropower plants. The erosion related to sediment supply reduction is located in the proximal fan area and proceed toward the fan foot.

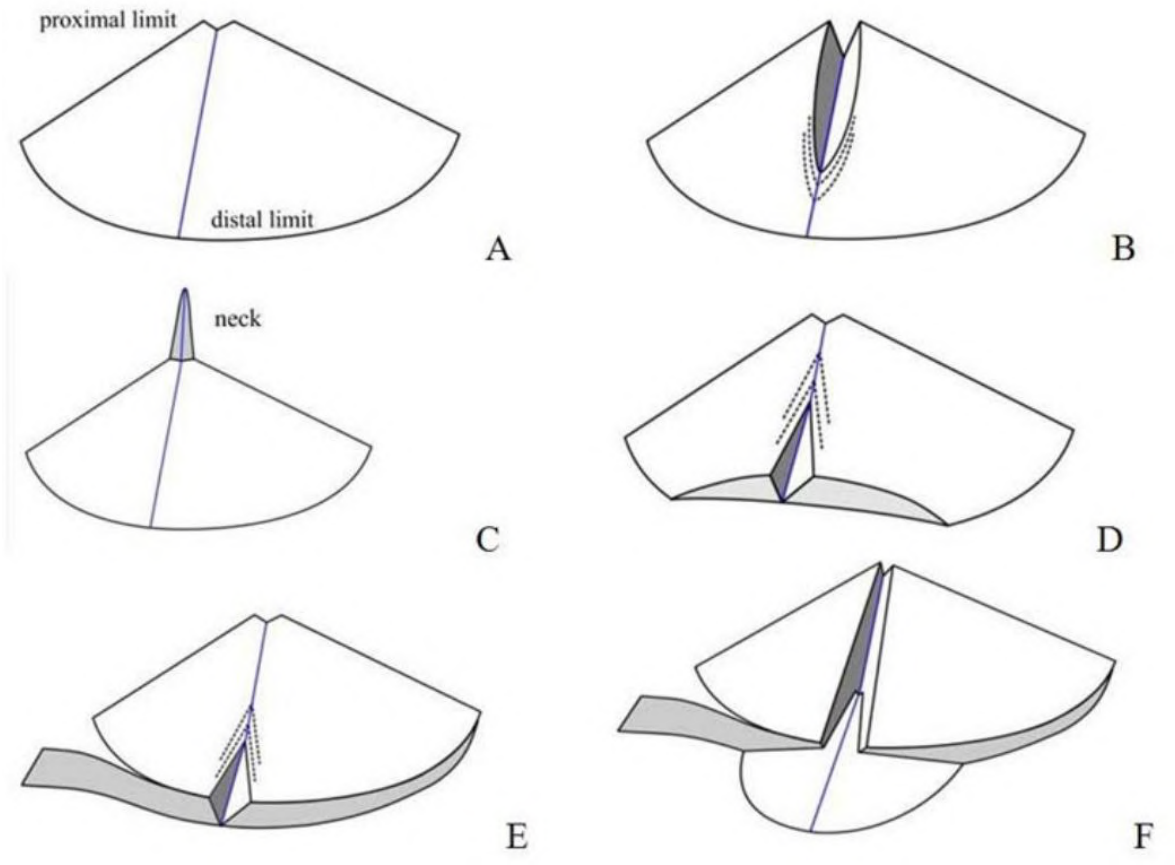
In the case of lateral movement or vertical changes of the valley channel's level, the erosion of the fan channel begin in the distal fan area and heads to the fan apex. A successive new formation of a secondary fan is then possible.
Considering the presence of deep fan channels or raised parts, is possible to classify fan portion into *active or recent deposition area* or *inactive or fossilized deposition area*.

1.3 Torrential processes

Within a torrent basin and thanks to the high energy of the flowing water, a wide variety of processes can happen. In this thesis, I am interested in phenomena related to fluvial events and thus I will not consider snow avalanches, rock falls, landslides and other, gravitational or not. More in detail, the focus will be posed to different paths of sediment erosion, transport and deposition.



(a)



(b)

Figure 1.7: Descriptive elements of a debris fan (from Eisbacher and Clague, 1984) and possible development stages: A. Fan in dynamic equilibrium; B. Downward erosion due to system changes in the watershed; C. Deposition above the proximal limits of the fan due to increased sediment transfer to the fan; D. Retrogressive erosion of the distal limits of the fan due to lateral change of the base level; E. Retrogressive erosion of the distal limits of the fan due to vertical change of the basal level; F. Situation E with additional formation of a secondary fan (modified from Hübl, 2006).

In montaneous and alpine basins, the effects of the water discharge have to be summed to the effects of solid material moving with the water. In many cases the solid component produce the main damages. Therefore, the study of erosion, transport and deposition processes to be expected for a certain basin is of essential importance in order to develop credible event scenarios, to create appropriate hazard and risk zoning plans and to design suitable mitigation measures.

1.3.1 Sediment erosion

Sediment erosion consists of disgregation of solid material by an agent, in this case by water. Other erosion agents are glaciers, snow, wind and waves. When speaking about the torrent, erosion takes place all over the basin. Of mayor interest for the subsequent transport and deposition is the erosion in the drainage basin and along the channel network.
Steep terrain and high difference in torrent basins'altitude, are responsible of the great water's gravitational potential energy. During the flow, this potential energy is converted into kinetic energy, expecially where the flow is less disturbed by turbulences or roughness elements. Several different ways of erosion are possible, depending on the water flow energy and solid material characteristics.
The first way is the *hydraulic action*, where the traction force of the water removes rock particles from torrent's bed and banks. Expecially for torrents, this type of erosion is responsible for the main amount of sediment mobilization. Its effect is magnified by high velocity sections like rapids and waterfalls. *Hydro-abrasion* is due to the abrasive action of the mobilized debris on channel bed and banks and it is as more effective as bigger and numerous are the dragged sediments (Suda, 2011). Where the main lithology is represented by limestone or dolomite (anything containing calcium carbonate), also *corrosion* plays an important role. The slighty water's acidity dissolves this kind of rocks into the water. With minor effects, *cavitation* can contribute to weakening rocks through small shockwaves generated by implosion of small air bubbles originating by flow turbulence (Suda, 2011).

The solid material eroded and loaded into the water flow can be mobilized from the channel network or from the valley slopes. In the first case, sediments are removed by the direct action of the water flow and the classical evolution processes of the **widening** (*Seitenerosion* in german) and **degradation** (*Tiefenerosion*) of the torrent thalweg will be visible. Effects of these phenomena is the **progressive backward incision** or **retrogressive erosion** of the torrential valley. In the second case, the solid material comes to the drainage system through surficial or deep landslides or flushed from slope surficial runoff. The last option is named *surficial erosion* (*Oberflächenerosion*). Avalanches, rock falls and other processes contribute to the delivery of erodible debris to the stream.
Important sediment sources are indicated in Weber (1964) and depicted in Figure 1.8. Torrent bed degradation is normally related to progressive backward torrent bed incision, it is charachterised by lowered thalweg due to erosion on loose material and can be named *file erosion* (*Feilenanbruch*). A special variation is the *scour erosion* (*Kolkerosion*), which happen at the bottom of

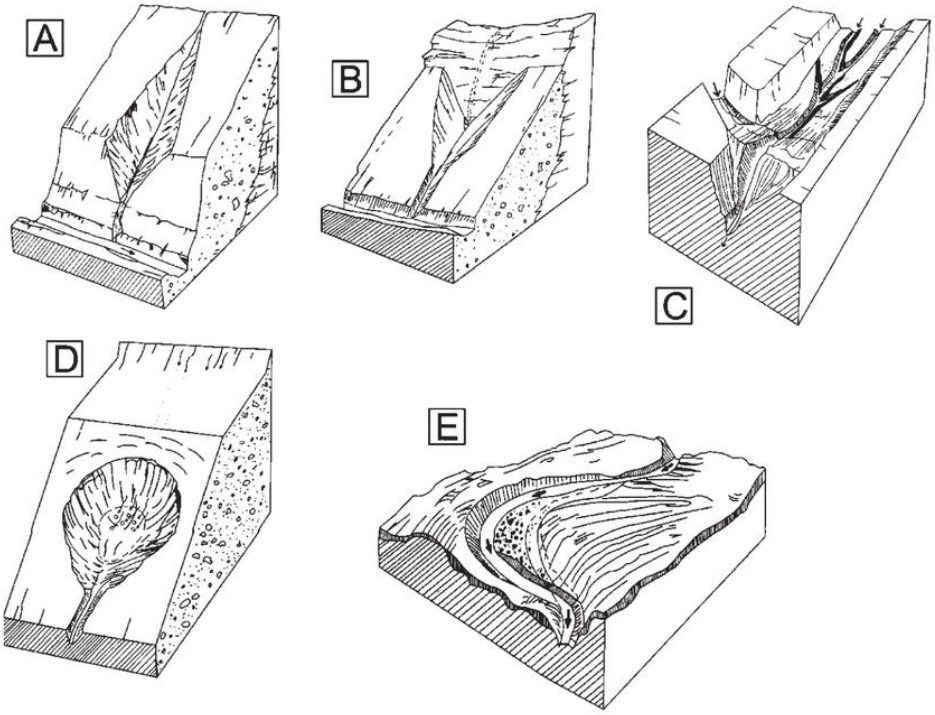


Figure 1.8: Typical sediment sources in torrent catchments: A. file erosion; B. wedge erosion; C. dam breake; D. slope failure; E. lateral erosion (from Weber, 1964)

waterfalls or any morphological steps in the channel. This erosion type is also called *wedge erosion* (*Keilanbruch*).

Widening is performed on river and torrent banks at bends and meanders (*Uferanbruch*). The erosion is due to the direct action of the water flow on the bank's material or through channel bed excavation and following bank's destabilization.
Local sediment sources are present on the valley slopes in form of rotational or translational landslides (e.g. *Muschelanbruch*). The formation of debris dams is common, often supported by the presence of blocks or logs with different origins, at the confluence of lateral channel reaches. These temporary deposits can suddenly collapse or been eroded causing a *dam break* (*Dammanbruch*). Similar results are given by the breach of moraine, by sudden release of water from pockets in glaciarized headwaters, by obstruction of under-dimensioned passages or bridges, etc. (Schneuwly-Bollschweiler and Stoffel, 2012).

Different sediment sources are distinguished by various mobilization velocity and mobilizable volumes. With these two parameters it is possible to estimate a *solid material mobilization intensity* (Bergmeister et al., 2009). Punctual sediment sources as landslides and dam breakes produce a sudden availability of a huge amount of mobilizable sediment. The other torrential erosion forms, although producing big quantities of solid material, are generally slow. For this reason the resulting mobilization intensities differ strongly. The graphic representation of this concept is shown in Figure 1.9.

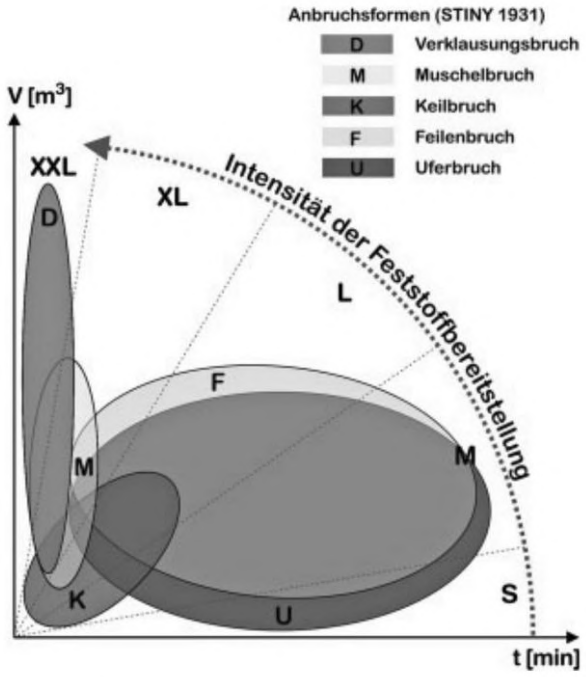


Figure 1.9: Mobilization intensity of the sediment sources (from Hübl, 2007)

Hübl (2007) states the crucial role of mobilization intensities on the sediment transport processes development along the torrent's channel. Low and medium mobilization intensities (sectors S and M on the graph of Figure 1.9) are responsible of debload transport, instead higher mobilization intensities (sectors L, XL and XXL) lead to intense transport phenomena likes debris flood and debris flow.

1.3.2 Sediment transport

The physical phenomenon that moves mineral and vegetal material of different sizes, generally from the catchment heads or channel reaches to the deposition areas, is called sediment transport. Really intense torrent processes, likes debris flows and debris floods, can deliver to the valley floor huge amounts of material within few minutes. Bedload transport is less intense but long lasting events are responsible of mobilization of big quantity of sediments. Expecially for risk management and mitigation structures design, the deep knowledge of the transport processes is of essential importance.
Granulometry of moved material strongly differs between transport processes. For a first classification it is enough to consider, besides water share, fine and coarse material portions plotted into the ternary phase diagram (see Figure 1.10).

Many authors proposed a classification of the different transport processes into suspended load, bedload, debris flood and debris flow. In addition to these, a torrent can transport material as dissolved load, that is made of soluble rocks like gypsum, halite and limestone, or by flotation of wood particles, waste, tanks, etc..
Suspended load, bedload and debris flood are selective transport types, since only certain sediment

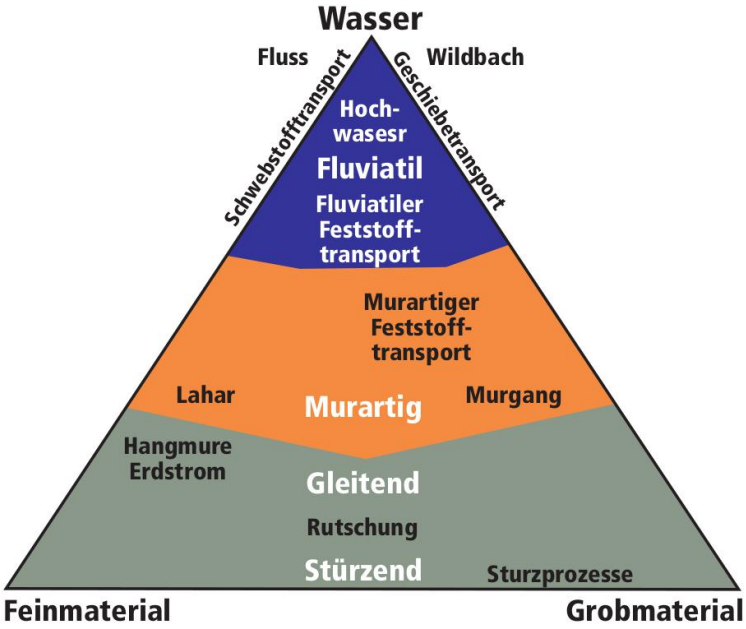


Figure 1.10: Transport processes depending on percentages of component granulometry (Rimböck et al., 2013; Phillips and Davies, 1991)

granulometry are transported by the flow. Conversely, debris flows are a sort of mass transport because any diametrical class is indiscriminately moved by the flow.
Depending on their available energy, torrents are able to carry a certain material mass, which is called *stream capacity*. Flow's discharge, velocity and depth, are responsible for the different sediment's maximum size that can be moved by the flow, which is named *stream competence*.
When the sediment is freely available for the transport, the transport capacity is the only limiting factors and the torrent is classified as *transport limited*. Again, if the erodible material is lacking, the torrent can be defined as *sediment limited*.
Is important to mention that the solid discharge is more variable in comparison with the liquid discharge. The first one is highly performed only during medium discharge and flood periods, whereas during low discharge times only dissolved load is completely represented (Ferro, 2002).

Clear water discharge and suspended load

When the water discharge is low, only few sediments are suspended or in solution. In this case, we speak about clear water discharge.
The suspended load, or wash load, is an important transport type during low-to-medium discharge periods, expecially in the case of big rivers. It is a nearly continuative process which does not cause big aggradation problems. In the case of torrents and small alpine streams, suspended load still important but its bearing on total sediment transport is diminished and often ignored. Turbulent motions are responsible of this phenomenon. The typical dimensional classes of sediment that are moved in suspension are clay and silt. Only strong turbulence within the water flow are able to suspend sand particles.
Suspendend load can be quantify through the concentration, either using material's weight over

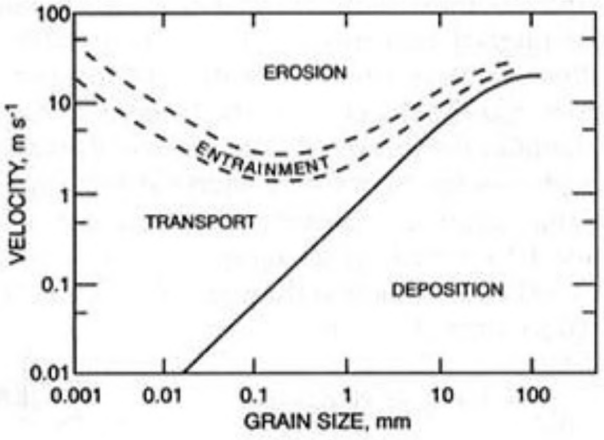


Figure 1.11: Flow velocity - Grain size relationships (from Knighton, 1998)

volume unit (N/m^3) or more commonly using mass over volume unit (g/l or mg/l). The density of the mixture water-sediment is not significantly increased and the value of 1000 kg/m^3 is normaly adopted. The distinction between suspenden load and bedload is based mainly on sediment's granulometry and considering the stream turbulence.

Bedload transport

When the material is transported mainly close to the bed, and the sediment granulometry is composed by sand ($2\ mm - 0,063\ mm$), gravel ($63\ mm - 2\ mm$) and cobbles ($2000\ mm - 63\ mm$), the transport process is called bedload transport. It generates the classical streambed translation and the movement phenomena are a) *rolling*, b) *sliding* and c) *saltation*. The largest particles are transported by traction. These particles are rolled along the bed, since the flow doesn't have enough energy to move these large particles in any other way. Slightly smaller particles are transported by saltation, since the flow has enough energy to lift the particles off the bed, even though the particles are too heavy to travel by suspension.
Bedload transport is the main transport process for many alpine torrents and can move huge amount of sediment and generate morphological changes of torrent reaches and banks. The debris comes from stream bed, expecially in the case of valley rivers, and from bank erosion and slope failure.
The density of the mixture water-sediment, also named as *bulk density*, is increased but is less than 1300 kg/m^3. The volumetric concentration of solid material can be up to 20%. Important feature is that the flow velocity of the transported debris is always slower than the water's flow velocity. Streams where bedload take place can be described by newtonian flow behaviour.
Bedload began when the *shear stress* τ, produced by the flow, exceeds the *critical shear stress* τ_{cr}, which is a function of sediment dimension, density and depositional pattern (armoring). For this reason, the sediment transport is *selective*, since every discharge stage can mobilize only determined debris granulometries. Graph of Figure 1.11 shows the relationships between grain size and flow velocity for the sediment erosion, transport and deposition (from Knighton, 1998).

Debris flood or hyperconcentrated flow

This kind of flow is a transitional phase between heavy fluvial bedload transport and debris flow. In other words, it is a transitional phase between selective transport and mass transport processes. Debris floods are also considered as not fully developed debris flows.
The distribution of solid material within the vertical flow's section is independent from particle's mass and geometry. This means that fine and coarse elements are equally arranged. Moreover, the flowing velocity of big debris is approximately the same of the flowing water. Debris flood's peak discharge remains in the same order as that of a flood, even if magnified by a *bulking rate* of up to 2–3, approximately (Costa, 1984; Hungr et al., 2014). Differently to debris flows, debris floods can occur also in large and flat watersheds. According to Hungr et al. (2014), large and damaging debris floods are often related to outburst and sudden drainage of moraine-dammed pro-glacial lakes, called *glacial lake outburst floods* (GLOFs).
Debris represent from 20% to 40% of the volumetric solid material concentration, causing the density of the mixture water-sediment to rise and being between 1300 kg/m^3 und 1700 kg/m^3. The rheology properties of this type of flow can still be contemplated as newtonian.
The distinction between bedload and debris flood and between debris flood and debris flow is not sharp. For this reason, from the examination of silent witnesses, erosional scours and depositional patterns, ambiguous results are possible.

Debris flow

Debris flows are unselective transport processes characterized by non-newtonian flow regime. Debris and water flow with the same velocity as a mixture. This process is characterized by high mobility, erosion potential and impact force. Debris flow process will be further described in the section 1.4.

1.3.3 Sediment deposition

When the torrent loses its energy, it is not able to transport its load anymore and deposition of both sediment and wood take place. There are several reasons why a torrent could lose energy.
In natural systems, the reduction of channel gradient leads to a slower flow velocity and thus a minor kinetic energy. This happen normally in the fan area or in any other part of a channel network characterized by low bed slope. Deposition take place also when the water flow reachs a lake or an artificial deposition basin. Another common phenomenon is the formation of deposits associated with lack of confinement, due to a lower flow depth in sections with wide channel's bed and resulting minor flow velocity.
Natural roughness elements - like boulders, tree trunks and stumps - or artificial ones - like debris flow brakes - cause localized turbulences and interfere with the flow velocity distribution. The result is not only sediment deposition but, because of contraction of the flow lines, also restricted zones of erosion. Water abstractions, can aventually reduce water discharge up to the point where sediment transport in not possible. Many mitigation measures aim to selective or unselective sediment deposition, adopting one or a mix of the above phenomena.

The sediment deposition depends on terrain, sediment and water volumes and material charachteristics (Bergmeister et al., 2009). Sediment supply to fan and other deposition area are episodic, and generally related to major precipitations, hailstorms or snowmelts (Schneuwly-Bollschweiler and Stoffel, 2012).
During a storm event a torrent can react in different way, making the identification of processes and the reconstruction of event scenario difficult.

Deposits from fluvial processes (suspended load and bedload) are normally built by fine and well sorted sediments and take, for example, the form of gravel bars, sand fields, etc.. Conversely, debris flow events build unsorted deposits, often with inverse granulometries sorting. Depositional forms and patterns of debris flow will be further described in section 1.4.

Silent witnesses

During the deposition, specific geomorphological features are formed. Intuitively, after an event the identification of torrential process or combination of them, is possible by looking at specific depositional patterns, geomorphological forms or silent witnesses (*Stumme Zeugen* in german) (from Hübl, 2003).
Aulitzki (1992), defines silent witnesses as geomorphic and biologic evidences of past mass-movement events founded in the field. Two types of silent witnesses exist: large-scale and small-scale geomorphic evidences.
Large-scale geomorphic evidences are represented by morphological features and indexees of drainage basin and fan, which are often able to suggest the most probable dominant process and event magnitude. The topography of the fan, as explained in 1.7, reflects the current state, development and erosion processes within the watershed.
If we look closer at fan and channel network it is possible to find *small-scale geomorphic evidences*. They are represented by depositional and/or erosional forms such as gravel bars, sorted and unsorted sediment banks, levees, mudlines and silting, big blocks, etc.. Moreover, also biotic elements, like tree (dendrogeomorphology), lichens (lichenometry) and vegetation analysis are important information sources. From the observation of this marks is possible to estimate, with a certain approximation, process parameters like maximum flow depth, cross section area, flow velocity, event magnitude, transport capacity and process type (Schneuwly-Bollschweiler and Stoffel, 2012; Kaitna and Hübl, 2013).

1.4 Focus on debris flow

Debris flows are a wide-spread hazardous phenomenon in mountainous terrains all over the world. As other torrential processes, they occur periodically on established paths, usually gullies and first- or second-order drainage channels. This means that a debris flow is specific to a given path and deposition area (debris fan). Debris flows are characterized by typical movement form and deposition features.
D'Agostino (2006) defines this process as flows constituted by water and solid material with an

high volumetric solid concentration. Their high viscosity, of several orders of magnitude greater respectively to water's viscosity, did not allow to compare their movement to the clear water flow. The definition from Hungr et al. (2014), states that a debris flow is a very rapid to extremely rapid surging flow of saturated debris in a steep channel, characterized by strong entrainment of material and water from the flow path.

Debris flows are highly unsteady phenomena, which show commonly rapid and sudden surges. These surges are followed by water-rich turbulent flows, with high suspendend load and low content of big debris (Pierson, 1980). A debris flow event may consist either of a single surge or many. The non-uniform flow behavior is well documented by Hübl et al. (2009) in the study on the austrian torrent *Lattenbach*, where 11 surges were measured during a 8 minutes period.
The peak discharge of the largest surge of an event can be estimated at several hundred of cubic meters, and may be more than one order of magnitude greater than the most extreme hydrological flood calculated for the same watershed (VanDine, 1985). The extreme discharge is responsible for great flow depth and velocity, which reflect the high impact loads and the ability to move large boulders.

The bulk density of a debris flow lies between 1700 kg/m^3 and 2400 kg/m^3. Its typical volumetric solid concentration ranges from 40% to 70% and the transported solid material is regardless of the granulometry even distributed all over the discharge section.
The volumetric solid concentration varies not only between different events, but also within the longitudinal profile of a singular surge. Its value is calculated using the equation (1.1):

$$V_{SC} = \frac{V_s}{V_t} = \frac{V_s}{(V_s + V_l)} \tag{1.1}$$

with:
V_{SC} = volumetric solid concentration of the flow;
V_s = volume solid fraction;
V_l = volume liquid fraction;
V_t = total volume of the flow.

Considering its rheology, this phenomenon is classified as a *non-newtonian* flow process.

1.4.1 Debris flow dynamic

The energy dissipation within a debris flow is caused mainly by two factors: a) external friction (e.g. between the flowing mass and the channelbed); b) internal friction (due to viscosity, rounding of bedload material, impulsive contact between the particles, damping of fluid motion due to suspended material, friction between the grains, turbulences in the fluid, mixing of the material, etc.) (from Kaitna, lessons). Considering the previous factors, it is possible to state that:

- increasing the concentration of fine particles (silt and clay), also the viscosity increases;
- a laminar or quasi-laminar flow substitute the turbulent flow from a certain mass viscosity;

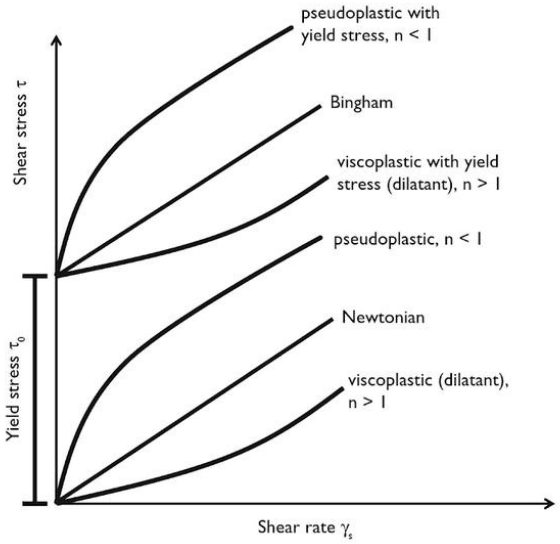


Figure 1.12: Rheological characterization of different fluids, modified from Kaitna (2006)

- a shear or slip surface develops between channel bed and debris flow. This means that shear stress occures only in the portion directly in contact with the channel bed, whereas the central part flows as a plug.

A debris flow is a multiphase fluid involving mixture of sediment, water and air. The behaviour of such a mixture is defined by its rheology and, in the case of debris flow, the fluid is classified as a non-newtonian flow (Iverson, 1997; Iverson, 2014). Selby (1993) states that the response to shear stress of a debris flow depends on: a) relative proportions of the components; b) sediment's grain-size distribution; c) physical and chemical properties of the sediments; d) temperature (lower influence).

For the description of the flow regimes of debris flows, several numerical models have been developed. Roughly, the used models can be grouped into rheological models or two-phase models. Rheological models treat the debris flow approximately as a homogene fluid, whereas in two-phase models, water and sediments are considered separately.

Rheological models are commonly used for many simulation purposes and are usefull for a first decription of debris flow behaviour (see Figure 1.12). They consider the relationship between *shear rate* and *shear stress*. The shear rate τ represent the flow velocity, whereas the shear stress γ the resistance to the flow.
For a newtonian fluid, as water, the flow behaviour of laminar flow, can be described by the equation (1.2):

$$\tau = \mu * \gamma \tag{1.2}$$

where μ represent the *dynamic viscosity* and is responsible of the slope of the line (for water is about 1 $mPa * s$)

The easier model for the description of a non-newtonian fluid, as a debris flow, is the Bingham-

Model (see equation (1.3)):

$$\tau = \tau_0 + \mu * \gamma \tag{1.3}$$

where the variable τ_0 is the so called *yield stress* parameter, which have to be exceeded by the driving force, or shear stress, in order to set in motion the fluid mass.
Variations of the two base models, as for example pseudoplastic, viscoplastic or dilatant, were proposed to better describe the different flow regimes.

Several authors proposed classification of debris flows based on different parameters. Due et al. (1986) use sediment granulometry as classification criterion dividing phenomenon into *mud flow*, *mud-rock flow* and *water-rock flow*. Wan & Wang (1994) use the density of the mixture sediment-water in order to obtain the classes of *viscous debris flow*, *sub-viscous debris flow* and *non viscous debris flow*.
Of major interest is the classification from Takahashi (1991), which is based on the prevailing flow regime. Macroviscous and inertial debris flows are described, the last ones are then further subdivided into four classes.

Macroviscous debris flow: is characterized by a laminar flow, which can be described and modelled by pseudo-newtonian flow models introducing an appropriate viscosity parameter. It occures when the sediment concentration is high or when the interstitial fluid is highly viscous (due to the presence of clay).

Inertial debris flow: are phenomena mainly driven by frictional and collisional forces. In the **stony debris flow**, sediments are distributed on the entire vertical section and big particles are drived and suspended by dispersive pressures generated by collisions and friction. Its turbulence is low and the flow is quasi-laminar. **Immature debris flows** are characterized by the concentration of sediments in the lower part of the flow section. Also in this type, the shear stress is due to the collisional forces. A **turbulent mud flow** occures when the turbulence of the interstitial fluid is able to suspend only small grained sediments. The turbulent flow regime is highlited by large vortex. The **hybrid of stony and muddy debris flow** is characterized by a mature or immature debris flow in the bottom part of the vertical flow section an by a turbolent flow in the upper part.

1.4.2 Debris flow basin

A debris flow process generally occures in basins characterized by small extension and large relief, calculated as difference in altitude between the higher point of the catchment and the outlet.
Considering an imaginary debris flow basin, it is possible to identify three main areas: *initiation zone*, *transport and erosion zone* and *deposition zone* (see Figure 1.13). VanDine, (1996), states that a slope gradient bigger than 25° is needed for the initiation of a granular debris flow. Transport and erosion need slope greater than 15°. A first partial deposition, in form of debris levees, takes place where the slope of the channel is lower than 15° or the flow confinement becomes to small. The deposition on a fan usually occures when the channelbed declivity decrease unter the 10°

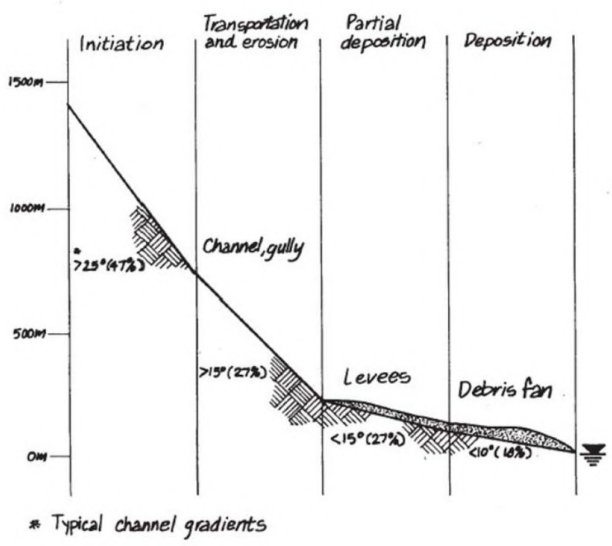


Figure 1.13: Approximate slope thresholds for debris flow's initiation, transport, erosion and deposition (from VanDine 1996)

threshold. As already mentioned, the slope values can vary depending on flow confinement, sediment features (matrix, lithology, geometry, granulometry, etc.), water content and flow type.

Initiation zone: it is the area, generally with steep morphology, where the debris flow is formed. In small alpine basins, sediment availability and stepness are not, at least locally, limiting factors. The water supply is made by precipitation and snow melting, but other sources are sometimes significant.

Transport and erosion zone: also called transit zone, it is the area which connects the initiation zone with the lower torrent reach. The debris flow assumes during the flow its typical structure with a steep front, a slurry body and a dilute tail. The erosive potential of a debris flow surge is mainly due to the presence of the bouldery front. The erosion and accumulation of soil, rocks and woody material into the flowing mass is called *entrainment*. Different authors state that the mass growth of a surge can be up to an order of magnitude. The eroded material originates from channel's bed and banks, slopes instability and any mass movement connected to the channel network. Bagnold (1954) suggests that block and other big particles are stored in the front because they flow in the upper part of the vertical section, where the flow velocity is bigger.

Deposition zone: also called runout area, often corresponds with the debris fan. Channel slope and confinement are leading factors for the deposition of debris flow, as well as the presence of obstacles to the flow, both natural (trees, big blocks, abrupt channel curve, etc.) and artificial (check dams, bridges, buildings, etc.).

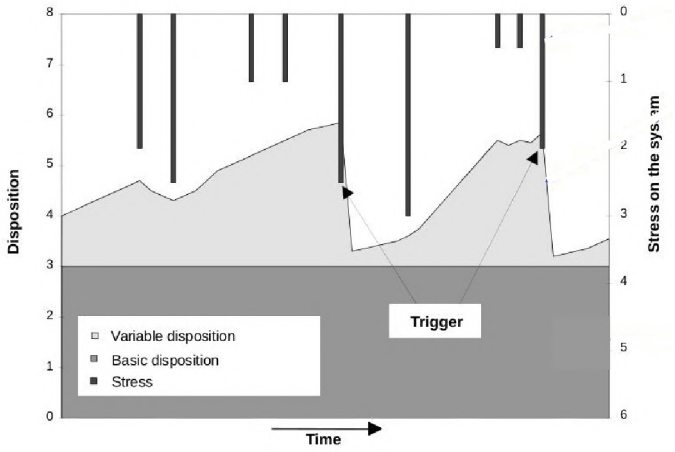


Figure 1.14: Basin dispositions and triggering events (from Zimmermann et al. 1997)

1.4.3 Debris flow disposition

Three main factors influence the susceptibility or disposition to the occurence of debris flows in a torrent basin (Iverson 1997):

- steepness of slopes and streambed (potential initiation zones);
- availability of loose or unconsolidated sediments (weathered rocks, quaternary morane, alluvium soils, etc.);
- hydrologic characteristics of potential initiation zones (sufficient water supply).

A clear explanation of the debris flow's occurrence mechanism on a certain torrent basin is given by Zimmermann et al. (1997). The author proposed to divide the basin's disposition into basic and variable disposition as shown in Figure 1.14. The sum of the two terms represents the actual debris flow disposition. Rainstorm events or snow melting periods represent the stress on the system. A debris flow event can be triggered only when the stress and the basin disposition exceed certain values, which are not constant but they vary over time.

The basic disposition is invariable or varies only on the long-term (decades to centuries), for example for the reactivation of fossile landslides. It expresses the general susceptibility of the torrent system for debris flow events. Location and magnitude of a possible event are strongly related to the basin's basic disposition. Watershed's relief, main geology, presence of freely erodible sediments likes morene, colluvium, rock fall's deposits, etc. strongly affect the basin disposition. More unstable factors influence the basin's variable dispostion like seasonal precipitation, presence of ice and snow, hydrologic conditions, vegetation periods, etc. The variability is measured on the mid-term (days to decades). The variable disposition is important expecially for the definition of debris flow frequency, but has a certain control also on event's magnitude.

1.4.4 Debris flow initiation

Many causes are likely to trigger a debris flow. For the alpine region the most important are surely short and intense rainfalls in the form of convectional and orographic storms. Expecially during late spring time, the snowmelt can trigger debris flow phenomena. In glacial areas lake outburst or sudden drainages are to consider. Volcanic eruptions, causing melting of snow and glaciers, and earthquakes are also possible triggering factors for debris flows.
Typical initiation scenarios for debris flow are listed in the following paragraphs.

- **Slope instability**: transition from landslide to a channelized debris flow after absorption of water (Selby 1993). Slope failures on the gully slopes deliver large quantities of material, which are deposited near the torrent bed and are easily mobilized by the water flow. A failure of the gully-head lead directly to the formation of a debris flow through absorption of water and formation of flow waves.

- **Channel bed instability**: liquefaction of bed material by *impulsive loading*. Similarly at the liquefaction process of a saturated landslide deposit, caused by the additional and sudden load of a new lanslide, Selby (1993) explains the liquefaction of bed material with the undrained load (generally from a slope failure) causing the increase and propagation of the pore water pressure and the subsequent motion of the bed material. Moreover the rapid and large water and solid discharge can destabilize and liquefy the torrent bed material. The flow erodes and entrains up to the point that it develops into a debris flow.

- **Blockage and release**: water pressure and flow erosion can breakdown temporary sediment dams (formed by lateral valleys, landslides, avalanches, etc.) leading to the sudden mobilization of big amount of solid material. This phenomenon is called *dam break*. The same happens when a step-pool-system or a glacial lake collaps. If undermined or overturned, human made transversal structures (e.g. consolidation check dams), are likely to be damaged, causing a *dam failure* and releasing the stopped sediments.

- **Firehose effect**: in the case that the channel begins at the bedrock/colluvial contact, entrainment occurs during a *firehose effect* where water-rich runoff from gulleys in bedrock cliffs pours on downslope colluvium. The firehose effect has been described in many studies (Coe et al. 2008).

1.4.5 Debris flow transport

The flow velocity of a debris flow is strongly variable and depends on many factors. The most important are surely solid material concentration and debris granulometry, together with channel features like form, steepness, sinuosity, macro-roughness and flow confinement. The most commonly measured flow velocity are in the range between 1 m/s to 10 m/s, but in many cases velocity of up to 20 m/s are observed locally (Major 1997).

During the motion, a debris flow surge has a typical form characterized by three main part as depicted in Figure 1.15 from Bardou (2003):

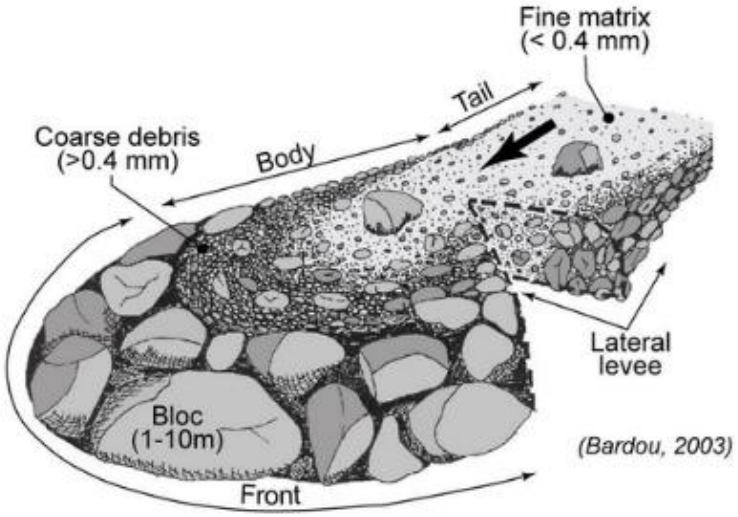


Figure 1.15: Debris flow surge (modified from Bardou 2003)

- **Front** or **snout** is the frontal section of a debris flow surge. As a result of channelization, a debris flow surge grows and becomes fronted either by a boulder concentration (Pierson, 1980) or a turbulent "head" (Davies, 1986). In many cases the snout contains big dimensioned boulders, blocks and woody material, it is only partially saturated and has an high erosive potential and a vertical shape. When the sediment granulometry is rich in fine material, the snout becomes a flat shape. In any case, this is the portion which shows the greater flow depth along the surge. A short but evident precursory surge of muddy material, frequently foreruns the real debris flow surge.

- **Body** is the median section of the surge. It is normally saturated with water obtaining the properties of a viscous mass. Generally it is more fluid than the front and is also made of finer sediments. Its upper surface is approximately parallel to the channel bed and the flow depth nearly constant. The high mobility of this part induces the collection of big particles to the front and therefore increases the erosive potential. Due to sediment interaction, matrix pressure and viscosity, boulders and large blocks are entrained, rolled and moved via suspension. Depending on flow's rheology, the body can move both in laminar or, more frequently, turbulent flow.

- **Tail** is the terminal section of the surge. It is normally more dilute and its flow depth is lower than in other parts. The viscosity descreases, following the significant reduction of the solid concentration to the point where turbulences lead to the transition toward a debris flood process.

1.4.6 Debris flow deposition

The partial deposition of a debris flow can occur up to steepness lower than 15° (corresponding to 27%) where the flow confinement in the channel began to be too small. The flow confinement can

be calculated using Equation (1.4). For example in the transport zone , the confinement grade is often lower than 5. The real deposition process of granular debris flows generally needs bedslopes between 18% and 5%.
Moreover, debris flows are largely affected by the sudden loss of water from the body.

$$Confinement\ grade = \frac{Flow\ width\ [m]}{Flow\ depth\ [m]} \tag{1.4}$$

From the vertical section of a debris flow's deposit it is possible to notice the general lack in homogeneity and dimensional sorting (Bergmeister et al., 2009). Expecially for granular debris flows, the deposit is formed by a big amount of large debris, weakly stratified. It is possible to observe an *inverse sorting*, where large blocks and boulders lie on top of a fine sediments. The typical flow section is semicircular or U-shaped, because of shear stress into the flow mass (VanDine 1996).

VanDine (1996) states that the deposition forms of a granular debris flow on a fan are dependant on the features of sediment, fan and channel section and on the presence of flow obstacles. As shown in Figure 1.16a three groups were proposed. It is possible to observe the presence of fluvial deposition forms, like alluvial terraces, formed by the deposit's erosion from the water-rich debris flow tail or flood recession curve.

- **Debris sheets** or **lobes** are lobate deposits, composed of heterogeneous granulometry. Generally they are located on the debris fan but, in many cases, they are also present on the side of the transport channel. Often debris sheets exhibit branching and a variable thickness. Observing european debris flows, with a magnitude between 10000 m^3 and 50000 m^3, Innis (1983) noticed debris lobe's thickness approximatively ranging from $1,1\ m$ to $1,7\ m$. The flow rheology has an important role in the deposition of debris lobes. On the one hand, mud flows or fine sediment-rich debris flows are charachterized by well-defined form, with sharp margins and abundant fine grained matrix. Whereas granular debris flows shown thicker and irregular lobes, formed by big boulders and mixed with a sandy matrix. Often pressure ridges are visible on the surface (Kaitna and Hübl 2013).

- **Debris plugs** are deposition forms which take place in the channel. The section blockage can be partial or total. In the case of a partial occlusion, sediments are deposited on the stream bed or close to the banks. When the entire section is occluded, than the plug can present a swelling and bigger flow depth. The formation of debris plugs causes damming of flow paths and consequent branching flow and avulsion to new areas. This erratic behavior extends the potential hazard across wide areas (Santi, 2017). Small debris plugs are often visible in the transport zone, where the channel suffers a sudden slope reduction.

- **Debris levees** are typical deposition forms, caused by the lateral loss of mass from the flowing debris flow front if the flow cross-section is too wide. After the passage of the entire surge, only the lateral debris embankments remain. They are normally present in the transport zone and in the upper fan segment. Debris levees indicate the happening of a debris flow for many decades. These deposition forms can be many meters high and

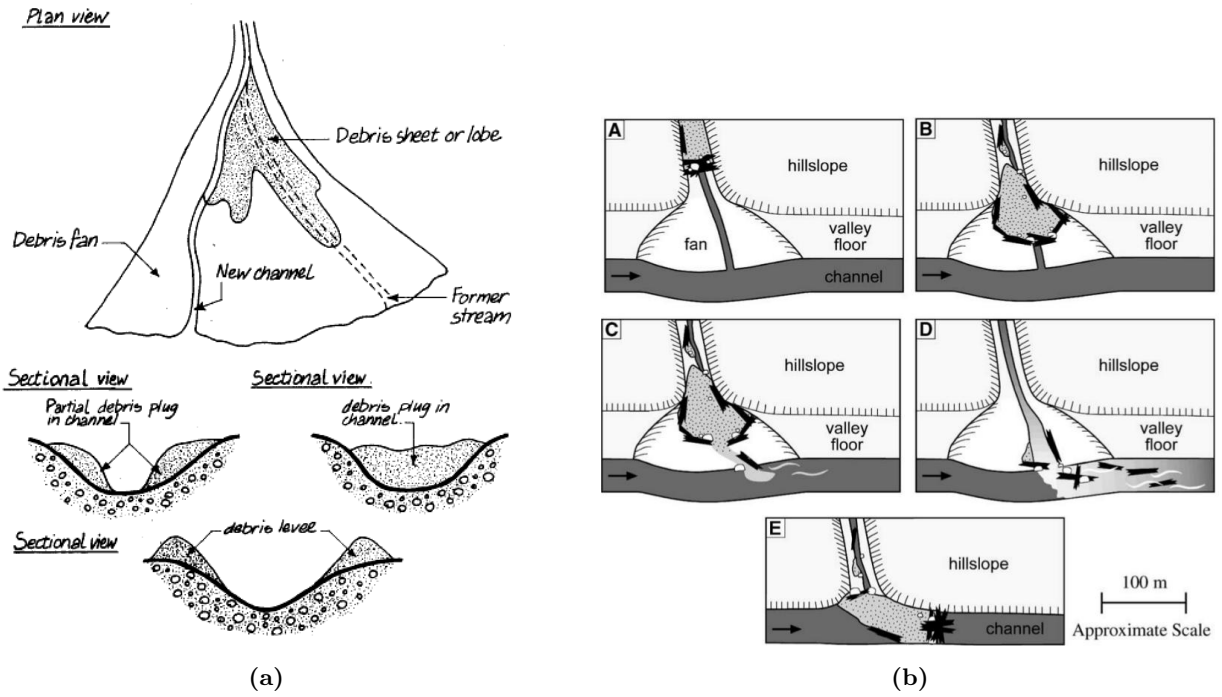


Figure 1.16: (a) Deposition forms of a channelized debris flow (from VanDine 1996); (b) Debris flow deposit types: A) debris dam and sediment wedge formed in tributary valley, B) deposit perched on fan surface, C) lobe of perched deposit overrans fan, D) deposit overrans fan surface, E) debris dam formed in mainstream river in narrow valley that prevented fan formation (from May and Gresswell 2004)

are characterized by steep flanks. They are several meters long and are present both along straight and curved channel reaches. In the second case, because of the centrifugal force, they show an higher crest elevation on the outside of the curve called *superelevation*, permitting also the back calculation of the flow velocity (Mizuyama and Uehara 1981; Costa 1984; Hungr et al. 1984; Bulmer et al. 2002; Scheidl et al. 2014). Bigger sediments are often located in the upper part of the levee. Muddy debris flows form well-defined, regular and round levees, very cohesive once dry. Granular debris fows, instead, form levees without clear limits, very loose levees once dry and with an approximately triangular vertical shape (Bardou 2002, 2003). Matrix-supported debris flows form levees rich in fine sediments and charachterized by flattened crests.

In addition to debris flow's magnitude, flow velocity and depth, the location of the deposits relative to the fan and mainstem river channel has particular inportance. This is particularly true if we are intereted in hazard and risk analysis or in event detection. In this sense, May and Gresswell (2004) propose a classification of debris flow's deposits based on location of the deposition process. Figure 1.16b shows the five suggested classes.

A The debris flow's deposit has no direct contact with the fan or mainstem river channel. Deposition takes place before the flow reaches the fan or valley bottom.

B Deposition takes place when the debris flow arrives at an existing fan or in the area immediately upstream of the fan ($< 100\ m$) that had been back-filled with sediment. The depositional material is out of reach of the valley stream.

C One or several debris lobes cut a path through the fan, partially reaching the mainstem channel.

D The debris flow overruns the fan, scouring the surface of it, and delivers the majority of its mass to the mainstem channel.

E Deposit type which does not interact with a debris flow fan, and formes a large debris dam in the mainstem river channel. It happens where the river prevented the formation of a fan in narrow valleys.

1.5 Mitigation measures

The design and implementation of mitigation measures aims to reduce the existing risk to an accepted level of residual risk (Zollinger 1986). This risk reduction can be achived through **active** and **passive** measures, which affect respectively hazard and potential damage (Zollinger 1985). Every process stage can be influenced by active mitigation measures. Therefore, magnitude and frequency of the process by initiation, transport and deposition are altered. If the probability of occurence is modified, it is possible to speak of *disposition management*, while if the process itself is influenced, we speak of *event management*. Passive mitigation measures are used to reduce the potential lost, decreasing goods and people presence into the endangered area or decreasing the vulnerability of the exposed objects. (Bergmeister et al., 2009)

The mitigation measures can be either **structural** or **non-structural**. The structural ones are normally structures, made of construction material and plants and the silvicultural maintenance operations (Hübl et al. 2005). Non-structural measures are constitute by zoning plans, warning and evacuation plans, insurance, etc. The structural mitigation measures can have a singular functionality or work together with other structures of the same type or with other features (which is called *functional chain*).

Another important feature of mitigation measures is the period of functionality. **Permament** measures can be both of structural and non-structural type. **Non-permanent** measures are active only by given system's conditions (e.g. threshold exceeding, event early-warning, etc.).

Many different mitigation measures and protection concepts can be found in alpine catchments. The systematic classification of all the mitigation measures can be done considering many parameters or features. Among others, construction material, structure geometry and design, process type, relative position to the channel axis and function type represent the most useful classification criterion. Just to focus the principal characteristics of the used structures, classifications based on position and function will be broadly described.

1.5.1 Sorting by position

The classification of stucture/measure considering its position relatively to the main movement direction is useful and permits a good and quick description of the used protection concept. It is possible to distinguish between linear and transversal mitigation measures. A third group is composed by measures with a surface effect.

Linear structure: is arranged along the flow main direction (process direction). These structures are normally used to deflect flow-processes and prevent the flooding and deposition of material into endangered areas. They are walls, banks and deflection dams.

Transversal structure: is arranged across the flow main direction. Various functions can be achieved by these structures and, considering their geometry, it is possible to identify solid body barriers, open barriers, ramp, etc.

Measure with surface effect: is located in the mountain slopes and prevents erosion or landslides. They are both non-structural (silvicultural operation) and structural (drainage, earth movement and shaping, slope stabilization).

1.5.2 Sorting by function

In the field of torrent management, a general overview of the functions conducted by mitigation measures is given by the austrian guideline (ONR 24800). It is mostly difficult to identify an unique process form or mitigation objective. For this reason, in the praxis is considered worthwhile to work and suggest a mitigation concept, which is based on many related mitigation funtions.

Diversion: (german *Ableitung*) covers all the measures which aim to carry the flow process out of the endangered area in the shorter way. There is a similarity with the correction of river (meander's cut off, realigment).

Stabilization: (german *Stabilisierung*) covers all the measures which aim to ensure the conservation of the actual conditions of torrent bed, banks and valley's sides. The main objective is the protection against lateral and vertical streambed erosion.

Consolidation: (german *Konsolidierung*) covers all the measures which aim to stabilize the slopes upward the structure through the up-lifting of the streambed. Secondary effects of these structures are the retention of considerable volume of debris, reduction of bed inclination and flow velocity, with a general decrease of the transport capacity.

Bypass: (german *Umgehung*) covers all the measures which aim to collect and transport water and debris discharge away from fragile torrent reaches or fan areas. The bypass can be done through open channels or pipes.

Retention: (german *Retention*) covers all the measures which aim to hold back water or sediment. Water retention takes place naturally along the channel with an attenuation of the flood wave (routing). It can be performed using reservoirs (steady retention) or activation of flood areas (flowing retention). The retention of sediment occures naturally in channel reaches with low declivity and confinement. It can be also achieve through check dams or deposition areas.

Dosage: (german *Dosierung*) covers all the measures which aim to the temporary retention of water into a basin in order to mildern the flood's peak discharge to an unproblematic limit. The retained water volume is released with an appropriate delay. Also the sediment transport is affected by this operation.

Filtering: (german *Filterung*) covers all the measures which aim to the selective retention of coarser blocks, timber and tree stumps, which can potentially lead to jam formation under bridges or narrow flow sections. Fine particles can unhindered pass through.

Energy conversion: (german *Energieumwandlung*) covers all the measures which aim to reduce the energy of flow processes. This is made taking advantage of structure's hydraulic features or freefall from dam's edge. The flow velocity is decreased and therefore also the flow process turns out to be transformed.

Deflection: (german *Ablenkung*) covers all the measures which aim to the change of process orientation away from the endangered area.

1.6 Literature review

The definition of the most likely natural hazard process for a given watershed is not a new topic in the technical literature. Various approaches have been used to identify basins prone to debris flows or other natural processes on a regional scale.
The strong qualitative and quantitative variability of available information sources in different countries leads to different methods being used. The proposed procedures are only appropriate for the study area and not directly applicable to other regions. Similarly, factors founded to guide the occurrence of a certain process in one study area are often characterised by low predictive power in other circumstances.

Many authors, starting for example with Melton (1965), have proposed approaches and indices for the identification of basins prone to debris flow or for the prediction of the dominant sediment transport process of a certain basin. Among the various studies of the last two decades on this topic, some interesting approaches are reviewed and presented.

In 2000, Ceriani et al. used morphological parameters to characterise alluvial fans. Statistical techniques were used to discriminate the dominant torrential activity among a sample of 97 alluvial fans in the central Italian Alps. Many geomorphological parameters were taken into account,

both for the drainage basin and the alluvial fan. The analysis aims to highlight any statistical relationship between the parameters and to analyse the relationship between the parameters considered and the type of transport process.

Coe et al. (2003) used historical and stratigraphic records of debris flows, which occurred on fans along a highway corridor in central Colorado, U.S.A., and coupled them with geomorphological parameters of the drainage basins (Melton's number). The result of their work is a regression model suitable for estimating the mean recurrence interval of debris flows on fans without stratigraphic or historical records.

An easy and interesting example of a model for predicting the occurrence of a given process is given in Ayalew and Yamagishi (2004). They worked on an inventory of 87 landslides in the Kakuda-Yahiko mountains, central Japan. A logistic regression was used to study the relationship between a dependent variable (presence/absence of the process) and several independent variables (such as lithology, slope gradient, aspect, etc.). This mapping procedure reveals that, for the study area, landslides, and in many cases also the resulting channelled debris flows, are strongly related to proximity to the road network, aspect and slope gradient.
Similar results were obtained by Ravellino et al. (2008) for the Campania Apennines in southern Italy. Landslides and the resulting debris flows are often linked to the proximity of escarpments and the forest road network, especially to sagging road segments.

A GIS-based identification procedure of debris flow-dominated channels was proposed by Cavalli and Grisotto (2006) for the upper Avisio basin in the Autonomous Province of Trento. The semi-automatic approach identifies debris flow channels by analysing DEM-derived maps. Potential trigger sites are identified by a relation between the local slope and the contribution area. Deceleration and deposition zones are evaluated only by means of local slope thresholds.
Carrara et al. (2008) compare the available models of debris flow susceptibility in the Alpine environment. Both statistical and physics-based models were considered, as well as slope unit and raster cell reference unit. In each case, a strong discriminating power between debris flow and flood dominated basins was found for the parameters of slope gradient, land use as pasture or without vegetation cover, availability of debris material and active erosive processes.
Horton et al. (2008) worked on a regional scale debris flow susceptibility map. Their GIS-based approach automatically identifies the likely area of origin taking into account slope angle, plane curvature, upstream contributing area, lithology and land use. Using a DEM with 10 m resolution, a simple assessment of debris flow spread is then performed.

An alternative sediment-budget model was proposed by Dong et al. (2009) for a region of Taiwan. It is used to predict the temporal variation of debris volume stored in a debris flow prone watershed. Basin topography and the debris volume stored in the sediment source area of the watershed help evaluating its debris flow susceptibility. The susceptibility index is calculated using channel gradient, catchment area, form factor and available debris volume as causative parameters.

Welsh and Davies (2011) focus the analysis to identify alluvial fans likely to be affected by debris flows. The used DEM derived parameters are the Melton number and the watershed length. Calibration and testing are performed using data from well known debris flow catchments in New Zealand. Linear regressions were carried out to test the relationships between observed evidence for debris flows, debris floods and fluvial processes and the indipendent parameters.

For an Italian area in the southern Apennines, Santangelo et al. (2012) use topographic indices to predict the flood susceptibility of alluvial fans. First, alluvial fans are recognised using GIS analysis. These fans are then classified in terms of their dominant transport process based on stratigraphic and statistical analyses. The authors state that for this specific geological context, the best discrimination between debris flow and flood processes is achieved by means of two correlated variables, the slope of the feeder channel and the length of the fan. By applying a logistic function all fans without any information are classified into one of the two groups formed.
Based on this research, Scorpio et al. (2016) performed the analysis of multiscale maps of geomorphological features related to flood-prone alluvial fans. Different features were used by adapting the resolution of the analysis to the scales of the maps, from small scales to detailed studies at larger scales (1:50,000-1:1,000).

Heiser et al. (2015) examined 84 well-documented catchments in the Austrian Alps in order to predict potential types of dominant flow processes in steep catchments. Several statistical models were trained against 11 morphometric parameters. The model with the best performance can be used to predict the most likely torrential process in a watershed.

Blais Stevens and Behnia (2016), for the Yukon Alaska Highway Corridor, Canada, proposed a qualitative heuristic model of debris flow susceptibility. A susceptibility index was defined using geomorphological parameters such as slope angle, slope aspect, surface geology, plane curvature and proximity to drainage system. Model validation was carried out by comparing the model results with the inventory of debris flow deposits compiled from aerial photos, high-resolution satellite images and field verifications.

A model for regional debris flow susceptibility analysis was proposed by Wang et al. (2016). The study area is located in southwest China. A total of 541 basins make up the database and 12 debris flow initiation factors (basin relief ratio, gradient gradient in the initiation zone, drainage density, downward curvature of the main channel, vegetation cover, aspect of the main channel, topographic moisture index, Melton roughness number, lithology, annual precipitation, shape factor and curvature of the transport zone) were identified. The statistical approach uses PCA (Principal Component Analysis) and SOM (Self-organising map).
Another approach for mapping debris flow susceptibility at a regional scale is that proposed by Bertrand et al. (2017). In this case, the dominant flow type, debris flow or bedload transport, is predicted within the hydrographic network. Two predisposing factors are used for the analysis:

the slope of the channel section and the Melton index of its upstream catchment. The probability of debris flow occurrence is calculated using logistic regression. Debris flow susceptibility is defined by combining classes of morphometric probability of occurrence and classes of estimated sediment availability.

From this brief overview it is possible to say that many different parameters have been used in various studies around the world and over time. As a general assumption it is worth stating that the most used parameters are those that in some way explain or reflect the physical circumstances surrounding the formation, initiation and movement of the studied process.
In the field of torrential processes, key roles are played by factors describing sediment availability or production (lithological map, geological score, etc.), water supply (rainfall, channel overflow concentration, contributing area, etc.), and water quality (precipitation, channel overflow concentration, contributing area, etc.).

1.7 Principal Component Analysis

1.7.1 Introduction

In the wide field of Machine learning and Data mining procedures, the **Principal component analysis**, also known with the abbreviation **PCA**, is a big dataset simplification technique generally used in the multivariate statistic. The aim of this technique is to reduce the number of variables which describe a dataset, into a less numerous latent variables set. It examines the correlations between the original objects and condenses the information contained within these objects into a smaller group of components. This is performed ensuring the minor possible loss of information, thus maintainig the variance in the data (Tufféry, 2011). This procedure finds different applications, for example in the exploratory data analysis and for making predictive models.
The invention of the PCA dates to the 1901 by the mathematician and statistician Karl Pearson. In the 1930s, Harold Hotelling independently further developed the method (Wikipedia). Depending on the field of application this approach gets different names and definitions, passing from mechanical engineering to meteorological sciences and others.

1.7.2 General structure

The statistical procedure of the PCA converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables, using an orthogonal transformation. The original data are expressed in terms of these new variables, called **principal components**. Considered together, the new variables represent the same amount of information as the original variables, and in the same way also the same amount of variance in the data (Kuhn and Johnson, 2013).
According to various authors (Tufféry, 2011; Jolliffe, 2002; and others), the used orthogonal transformation is defined in such a way that the first principal component has the largest possible variance. This means that it accounts for as much of the variability in the data as possible. Each succeeding component in turn has the highest variance possible under the constraint that it is

orthogonal to the preceding components. The PCA procedure is sensitive to the relative scaling of the original data variables. For this reason a normalization step of the initial data is required. The used method to normalize each attribute is named mean centering and consists in subtracting each data value from the calculated variable's mean. In this way the empirical mean is zero. Possibly each variable's variance is normalized making it equal to 1.
The application of this method reveals the internal structure of big dataset in a way that best explains the data's variance. In other words the process gives a representation of the dataset viewed from its most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

1.7.3 Total variance and explained variance

As already mentioned, the principal components retain the same amount of information as the variables in the original data, and therefore the same amount of variance. The sum of the sample variances of all individual variables is called the total variance. However, the total variance of the data set is redistributed among the new principal components in the most unequal way (Kuhn and Johnson, 2013). To understand this statement, it is necessary to introduce the topic of **explained variance**. Given the **total variance** of a dataset, the fraction of variance explained by a variable is given by the ratio of the variance of that variable to the total variance.
When it comes to PCA, the first principal component not only explains the most variance among new variables, but also the most variance that a single principal component can explain (Kuhn and Johnson, 2013). It is also possible to calculate the fraction of variance explained by subsets of principal components by simply adding up their variances and dividing the results by the total variance of the dataset.

More generally speaking, the first k principal components (where k can be 1, 2, 3 etc.) explain the most variance any k variables can explain. In the same way the last k principal components explain the least variance any k variables can explain. This is a peculiar property of the PCA and permits to maximize the explained variance reducing at the same time the number of considered variables. In fact, the typical use of PCA is to keep only the first $k < p$ principal components. Because PCA is an orthogonal transformation, this corresponds to projecting the data from its original *p-dimensional* space to a *k-dimensional* subspace (Abdi and Williams, 2010). The remaining $p - k$ components, the variables with least explained variance, are lost in this projection.

Recapitulating, as highlighted by many authors, the PCA is designed to maximize the variance of the first k components, and minimize the variance of the last $p - k$ components, compared to all other orthogonal transformations. It is crucial to choose the first k components, and not just some k components, because they have the highest variance out of all principal components. In order to preserve all the important information of the analyzed database it is important to choose k big enough to make the lost information, the variance of the last $p - k$ components, sufficiently small.

1.8 Logistic regression

1.8.1 Introduction

The aim of the proposed model is to predict whether a certain basin belongs to the class of debris flows producing watersheds or not. All the basins will be classified either in the class *"debris flow prone"* or in the class *"not debris flow prone"*. This classification is performed using the calculated principal components obtained applying the PCA to the original features database.
The two classes of basins represent the *binary dependent variable* which will be predicted on the basis of many *predictors* or *independent variables* (the principal components). This prediction is performed using the **logistic regression**, which is a statistical model that in its basic form uses a **logistic function** to model the probability of a certain class or event as a binary dependent variable, or in other words a dichotomy (Wikipedia). The logistic regression belongs to the group of the *generalized linear model*, known with the acronym *GLM*.

1.8.2 Application

Binary variables can not be predicted properly using a common linear regression for two reasons: a) linear regression predicts probabilities outside the range of 0 to 1; b) the residuals will not be normally distributed around the predicted line since the prediction can only have one of the two possible values (Sayad).

Not only binary dependent variables can be modeled with this method. There are three different types of logistic regression:

- binary logistic regression (categorical response with only two possible outcomes, indicator variable is coded as 0 or 1);
- multinomial logistic regression (three or more categories without ordering);
- ordinal logistic regression (three or more categories with ordering).

For a general binomial dependent variable, the corresponding probability of the predicted class "1" can vary between 0 (certainly the value "0", referred to as a "unsuccess" or a "noncase") and 1 (certainly the value "1", referred to as a "success" or a "case"). The logistic regression itself models simply the probability of an event based on the predictors, and does not perform a statistical classification (Wikipedia). The model can be used as a binary classifier when coupling predicted probabilities and a *cutoff value*. Based upon this threshold, the obtained estimated probability is classified into classes. Elements with probability greater than the cutoff value are classified in the "1" class, while "0" class is given for smaller probabilities.
Also the predictor or dependent variable, one or more, can be of different types like nominal, ordinal, interval or ratio-level.
In the general case of a binary logistic regression, some assumptions are true:

1. dichotomous nature of the dependent variable (presence vs. absent);
2. no outliers are present in the data;

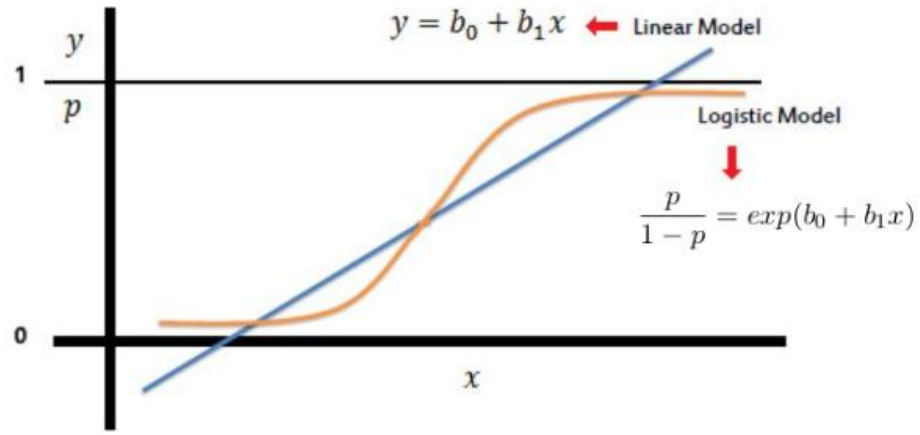


Figure 1.17: Sigmoid curve for the logistic function (modified from https://www.saedsayad.com)

3. no high correlations (multicollinearity) among the predictors.

As mentioned above, the core of the logistic regression is the **logistic function**. It is an S-shaped curve (sigmoid curve) that can take any real-valued number between the values 0 and 1 (never reaching exactly those limits). The visual representation of the logistic function is shown in Figure 1.17. This logistic curve, produced by the logistic regression, is limited to values between 0 and 1. The curve is constructed using the natural logarithm of the **odds** of the target variable, rather than the *probability* as usual for linear regression.

The odds is the ratio of the probability of a case (class 1) over the probability of a noncase (class 0), which can also be written as $p/(1-p)$. One common method for quantifying the predictive ability of a binary predictor is the odds ratio (Kuhn and Johnson, 2013).

The logistic regression equation can be written in terms of **odds** as shown in Equation (1.5). The constant b_0 moves the sigmoid curve left and right and the slope b_1 defines the steepness of the curve.

$$\frac{p}{1-p} = exp(b_0 + b_1 x) \tag{1.5}$$

Multiplying both sides of Equation (1.5) for the natural logarithm, is possible to write the same equation in terms of **log-odds**. The log of the odds is also known with the name of **logit**, which is a linear function of the predictors modeled by the logistic regression (Kuhn and Johnson, 2013).
The right hand side of Equation (1.6) represents a linear model, where the constant b_0 defines the intercept and the coefficient b_1 gives the variation of the logit or log-odds caused by a one unit change in x.

$$ln(\frac{p}{1-p}) = b_0 + b_1 x \tag{1.6}$$

In most applications, the base of the logarithm is usually taken to be e (Euler's number). However it can be used the logarithm in base 2 or base 10 (Wikipedia).

If needed the result can be expressed as probability that the prediction is a case (class equal

to 1). Equation (1.7), is formed by simple algebraic manipulation and calculates the probability that $Y = 1$.

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p)}} \tag{1.7}$$

As seen in the previous equations, the relations between predictors and dependent variable is given by the model coefficients, which are estimated using the **maximum likelihood estimation** method or **MLE**.

This technique can be used when we are willing to make assumptions about the probability distribution of the data. Based on the theoretical probability distribution and the observed data, the likelihood function is a probability statement that can be made about a particular set of parameter values (Kuhn and Johnson, 2013). The probability distribution that is most often used when there are two classes is the binomial distribution.

The set of parameters values with the larger likelihood would be deemed more consistent with the observed data.

Chapter 2

WORK'S OBJECTIVES

The objective of this thesis is to develop a method for the identification of watersheds subject to sediment transport, focusing on very intense processes such as debris flows, for small alpine catchments in the Autonomous Province of Trento.
Considering the limitations of the available data, a binary variable of the *likely-unlikely* type will be determined for each watershed-fan system.
For a practice-oriented application, the proposed method should be as automated and transparent as possible and the results obtained should be easy to interpret.

Chapter 3

STUDY AREA

The study area of this thesis is the territory of *Trentino*, also known with the official name of *Autonomous Province of Trento* (*Provincia autonoma di Trento* in Italian and *Autonome Provinz Trient* in German). Due to the presence of lenguage minority and to the wide use of dialects also other names are recognized: *Provinzia Autonoma de Trent* in Ladin, *Autonoma Provinz vo Tria* in Cimbrian and *Autonome Provinz va Trea't* in Mòcheno.

The Autonomous Province of Trento, together with the Autonomous Province of Bolzano (known with the name of *South Tirol*, forms the region *Trentino-AltoAdige/Südtirol*, one of the 20 regions of Italy. The provinces of Trento and Bolzano are the only ones in Italy, for which the Italian constitution directly acknowledge the status of administrative autonomy.

The province is located in the country's far north and is divided into 177 municipalities, in Italian named *comuni*. Its capital is the city of Trento, which is right in the center of the area (see Figure 3.1).
The administrative area covers an area of $6.207,12\ km^2$, with a total population of 537.976 inhabitants at the date 31^{st} August 2016. The population density presents strong variation within the province with an average value of about 86 inhabitants pro km^2 (Statistic service PAT).

3.1 Etimology

The name *Trentino* was introduced by the irredentist movement in the middle of the 19^{th} century and descends from the latin names of Trento: *Tridentum*. During the Habsburg dominion the whole area from the Brenner Pass to the boarder with Veneto was called *Südtirol*, sometimes with the specification of *Welschtirol* to identify to Italian speaking part (Tomasi 1997).
The istitutional use of the name *Südtirol* exclusively for the autonomous province of Bolzano is more recent and dates back to 1972 (Wikipedia: Trentino).

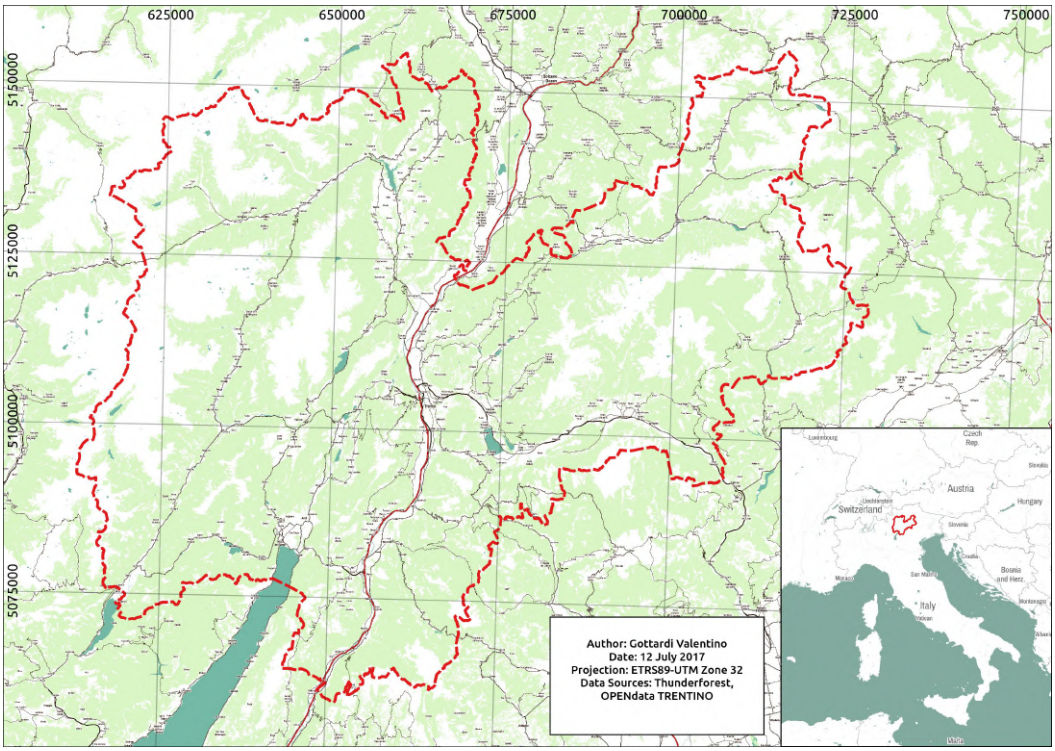


Figure 3.1: Geographical location of the Autonomous Province of Trento

3.2 Geography

3.2.1 Morphology

The most elavated peak of the province is the Cevedale with 3769 m above sea level, which is part of the Ortler-Cevedale massif and is located in the Rhaetian Alps. With only about 65 m above sea level,the lake shore of Torbole and Riva del Garda is the deepest point. This leads to an altitude difference of 3700 m in a few kilometres as the crow flies.
The general morphology of the provincial territory is regulated by the presence of the Adige valley which is the main valley cut. It has a North-South orientation and has been carved out by glaciers during the ice ages. The side valleys are sometimes suspended and connected by narrow gorges. Other important valleys are Val Rendena, Giudicarie, Valle dei Laghi and Alto Garda, Val di Non and Val di Sole, Valsugana, Primiero, Val di Fassa-Fiemme-Cembra.
The morphology is strongly influenced by the presence of extensive mountain ranges. Among the others are: Ortles-Cevedale Massif; Adamello-Presanella Massif with Presanella 3558 m above see level (the highest peak entirely included in Trentino) and Carè Alto 3465 m $a.s.l.$; Maddalene; Brenta Dolomites with Cima Tosa 3173 m $a.s.l.$; Western Dolomites with Latemar, Catinaccio-Rosengarten, Sasso Lungo-Langkofel, Sella, Marmolada, Cima Undici and the Pale di San Martino massif; Lagorai and Cima d'Asta ranges with Cima Cece 2754 m above see level and Cima d'Asta 2847 m above see level; Baldo-Bondone range and Ledro Alps.

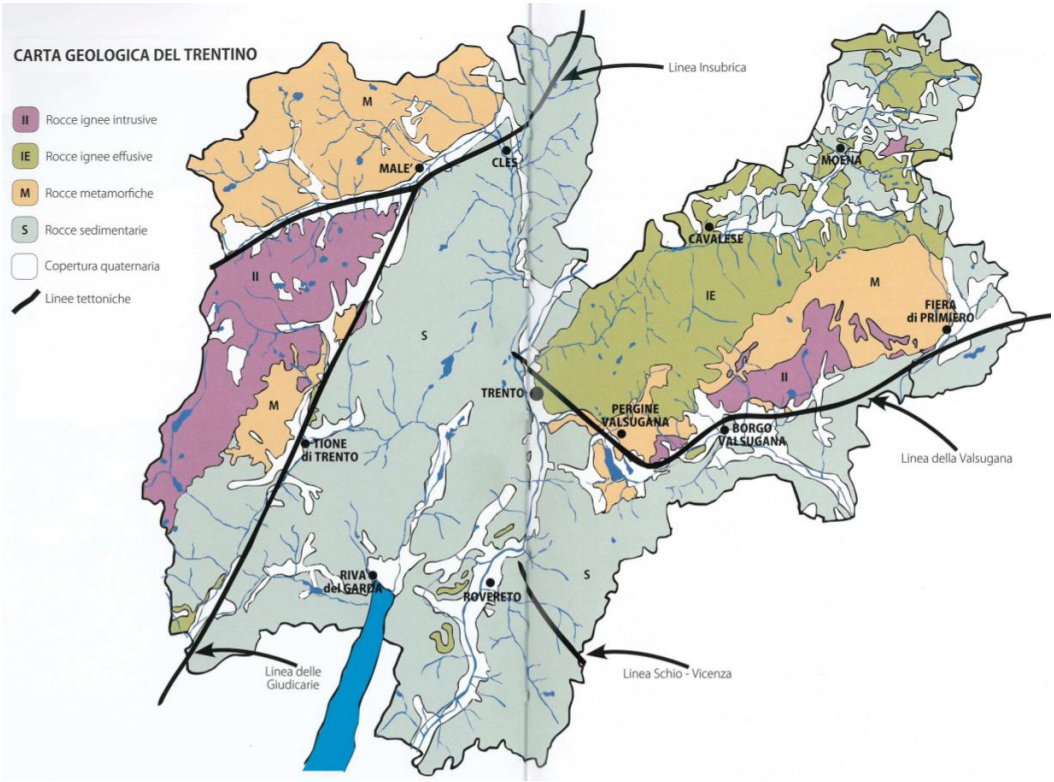


Figure 3.2: Geology of Trentino (from Caldonazzi and Avanzini 2011)

3.2.2 Geology

As in many other locations in the Alps, it is not easy to make generalisations about the geology and lithology of the region. This is mainly due to the various formation processes of the mountain ranges and valleys.

There are four main processes that contribute to the current geological structure: volcanism, sedimentary rock deposition, metamorphism and erosion (Caldonazzi and Avanzini 2011). Volcanism occurred in various time windows and forms the igneous rocks of the *Atesine Porphyric Platform* and the *Adamello and Cima d'Asta plutons*. Together with the deformations, compressions and heating generated by the forces of tectonic movements, volcanism led to the formation of metamorphic rocks that are widespread in many places in the territory (for example, the Austroalpine nappes of the Ortles-Cevedale massif). Rocks of sedimentary origin are widespread, most of them biochemical sedimentary rocks (e.g. limestone, dolomite) but there are also clastic rocks (e.g. conglomerates and sandstones) and chemical rocks (e.g. gypsum). All these formations have been continuously modified, mixed and eroded by tectonic movements, weathering, rivers and glaciers, also forming massive quaternary debris deposits, especially on the main valley floors.

Figure 3.2 shows a simplified distribution of the main lithological groups and fault systems on the Trentino territory. The purple coloured areas represent intrusive igneous rocks; the green areas show effusive igneous formations; the orange parts represent metamorphic rocks; the blue areas represent sedimentary rock formations and the white areas highlight quaternary deposits. The black lines show the main fault systems.

The lithology pattern varies strongly between the valleys and is of fundamental importance for the comprehension of torrential behaviours. In the following paragraphs, main valley's geology is broadly described.

Sole valley: presence of metamorphous rocks to the North and North-West (phyllites and basement's paragneiss) and instrusive igneous rocks (granite and tonalite) of the Presanella.

Non valley: dominant limestone and marly-limestone formations.

Giudicarie and Rendena: various formations, metamorphous rocks on the downriver, right bank of the Sarca (gneiss and micaschists); intrusive igneous rocks of Adamello (granite and granodiorite); dolomitic-limestone on Brenta Group and mixture in Val Daone (marl, sandstone and shale).

Upper Garda and Ledro: diffuse presence of limestone and dolomitic-limestone formations, marl and shale.

Vallagarina: limestone and dolomitic-limestone formations with local outcrops of volcanic rocks (porphyry, basalt) and magmatic rocks (granodiorite).

Adige valley: prevailing presence of limestone and dolomitic-limestone formations; to the North-East of Trento strong existence of effusive igneous rocks, expecially porphyry.

Upper Valsugana: various geology, South and South-West of the Valsugana fault diffused limestone and dolomitic-limestone formations; metamorphous rocks on the downriver, left bank of the Brenta (phyllites and micaschists) and effusive igneous rocks to the North (quartz-porphyry of the Lagorai range).

Lower Valsugana: really heterogeneus geology with limestone formations to the South and to the East; metamorphous rocks (phyllites and micaschists) to the North-West; effusive igneous rocks (quartz-porphyry) and intrusive igneous rocks (diorite and granite of the Cima d'Asta) to the North.

Cembra valley: prevailing quartz-porphyry presence on both Avisio banks.

Fiemme valley: the downriver, left bank of the Avisio is characterized by the Lagorai quartz-porphyry. On the other valley side dolomitic-limestone formations are present. The Predazzo district was in the past one of the main study area of the Alps for the important outcrops of intrusive (monzonite, sienite and granite) and effusive (basalt and tuff) igneous rocks.

Fassa valley: potent dolomitic and limestone formations placed on top of marly-limestones and sandstone. The underlying basement is build by igneous rocks of the Atesine Porphyric Platform. Basaltic outcrops are present in the Buffaure group.

Primiero: the West and North-West area is formed by mainly by metamorphous rocks, and lesser by igneous formations and sandstone; to the North is present the dolomitic formation of the Pale di S.Martino whereas to the South limestone is predominant.

3.2.3 Seismicity

Four main fault systems are active and significant for the province (see Figure 3.2):

- Insubric fault (NE-SW orientation);
- Giudicarie fault (NNE-SSW orientation);

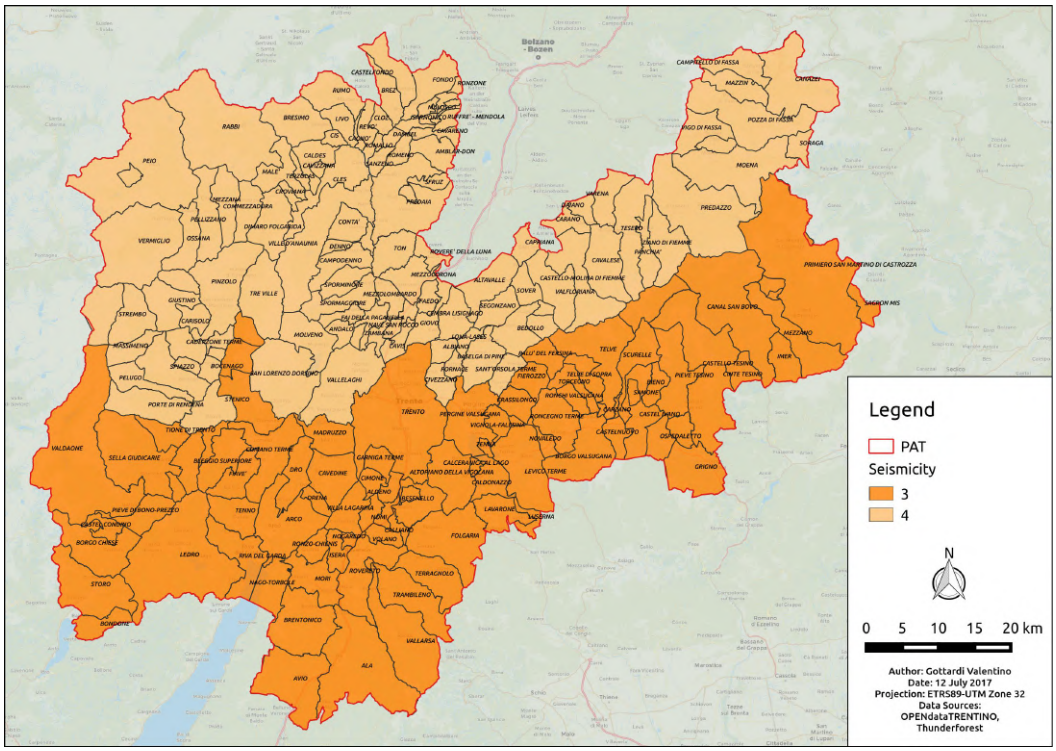


Figure 3.3: Seismic zones in Trentino. Source: Protezione Civile Trento

- Valsugana fault (ENE-WSW orientation);
- Schio–Vicenza fault (NW-SE orientation).

As displaied in Figure 3.3, southern Trentino is classified, according to the Italian seismic classification, in the 3^{rd} seismic zone (low seismicity), whereas the northern part is in the 4^{th} seismic zone (very low seismicity) (Civil protection PAT).

3.2.4 Hydrography

The main rivers and torrents existing on Trentino territory are listed in Table 3.1 and are ordered by catchment surface in the study area. Some of them flow entirely within the provincial boundaries, while the longer ones flow further towards the Po River or the Adriatic Sea. The Adige River is the most important water artery in Trentino. The water network count 3338 channels for a total length of 5760 km and 192 km^2 of public property.

Lakes are abundant in the province of Trento. Next to the natural lakes a large number of dams have been built to create totally new reservoirs or to improve the content capacity of natural Alpine lakes. All these water bodies represent a crucial source for drinking water as well as for irrigation and electricity production. Although only a small part of Lake Garda is included in the province, it is still the largest water body. Lake Caldonazzo is the largest lake entirely contained within the provincial area (Tomasi 1962). Based on water surface, the main lakes are listed in Table 3.2.

Table 3.1: Watersheds of Trentino

Watershed	**Surface** $[km^2]$ **in Trentino**	**Total length** $[km]$
Noce	1367	105
Sarca	1268	77
Adige	950	410
Avisio	937	89
Brenta	618	174
Chiese	410	160
Vanoi	237	25
Cismon	209	51
Fersina	170	37
Astico	84	53
Cordevole	44	79
Senaiga	44	15

Table 3.2: Lakes of Trentino

Lake	**Surface** $[ha]$
Lago di Garda	1421
Lago di Caldonazzo	528
Lago di S.Giustina	377
Lago di Molveno	322
Lago di Ledro	211
Lago di Malga Bissina	127
Lago di Levico	109
Lago di Cavedine	88
Lago di Forte Buso o Paneveggio	78
Lago di Toblino	70
Lago di Pian Palù	55
Lago artificiale di Fedaia	55

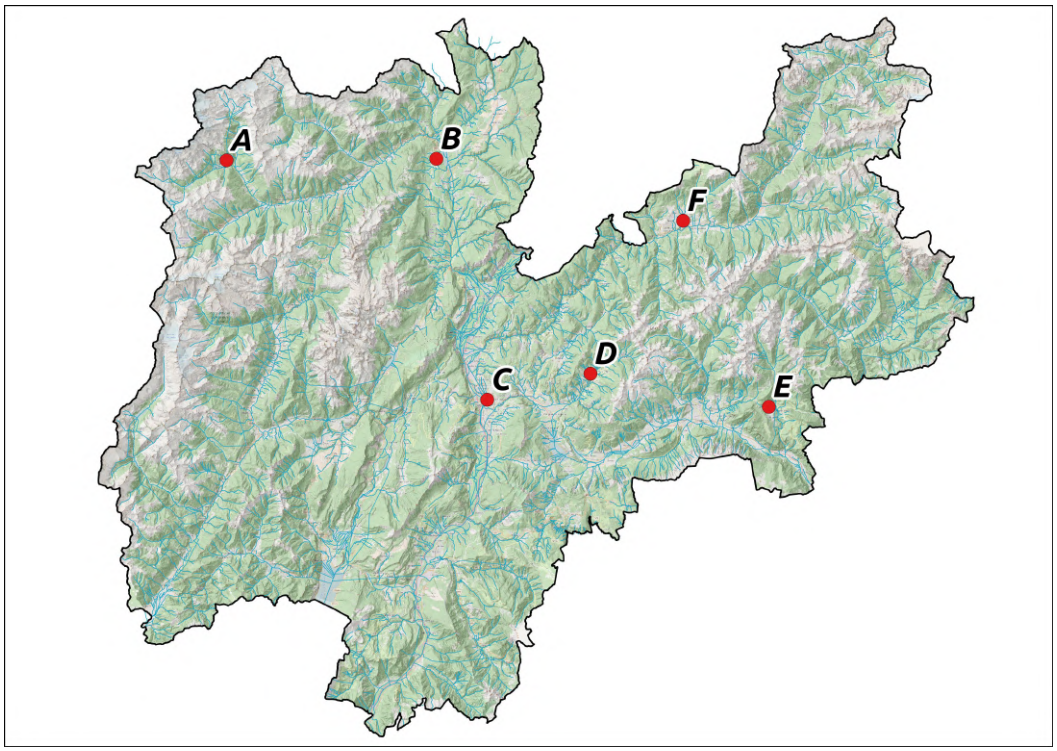


Figure 3.4: Weather stations location of the represented weather data

3.2.5 Climate

Due to the complex morphology of Trentino, characterized by differently sized and oriented valleys, mountain ranges, lakes, hollows and hills, an outstanding climatic variety is present.
Considering also the great elevation gradient, it is easy to understand how much difficult the classification of the climatic zones is. However, according to the thermal annual trend, the following main areas can be divided (Clima Trentino):

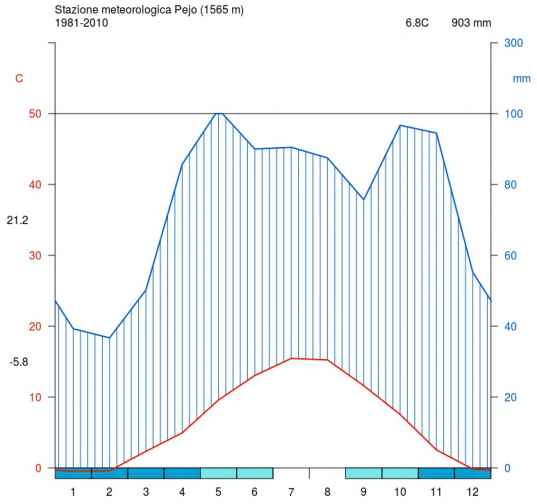
a. the lower zones, like the Upper Garda plain and the Adige valley, are characterized by relatively cold and snowy winters. Summers are hot and showery, often sultry during the day. Exception to this, is given in the afternoon by the *Ora* wind, which blows from the Garda lake to the north going beyond the Sarca valley and reaching Lavis in the Adige valley;

b. the lateral valleys, like Non valley and Valsugana, which have more moderate temperatures in summer and slightly colder in winter;

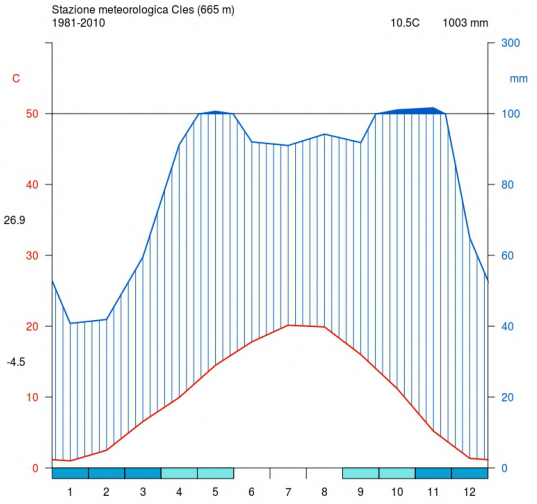
c. the cold hollows and depressions between 500 and 1000 m *a.s.l.*, like Bleggio and Fiemme valley floor have mild summers and frigid temperatures in winter;

d. the mountain zones over $1300 - 1600$ m *a.s.l.*, with the typical alpine climate characterized by fresh and rainy summers with frequent storms and cold and really snowy winters.

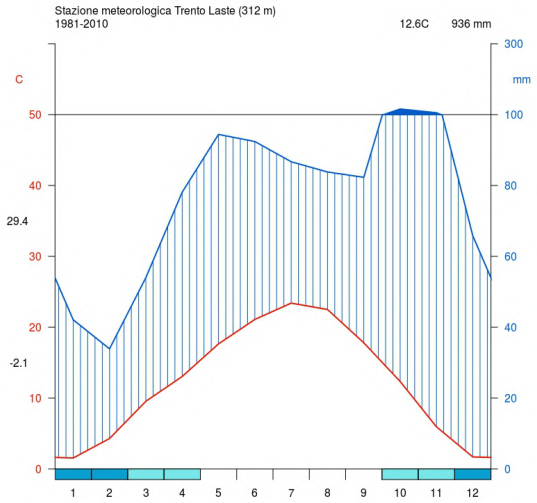
The precipitations are unevenly distributed between different areas: more rainy zones are located to the south and southwest because more exposed to moist air flows from the south. Most



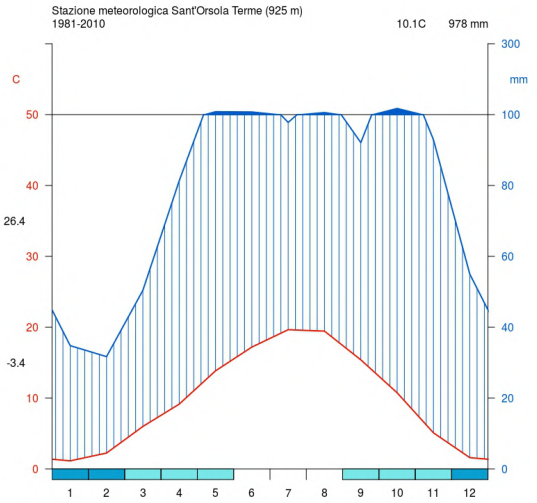
(a)



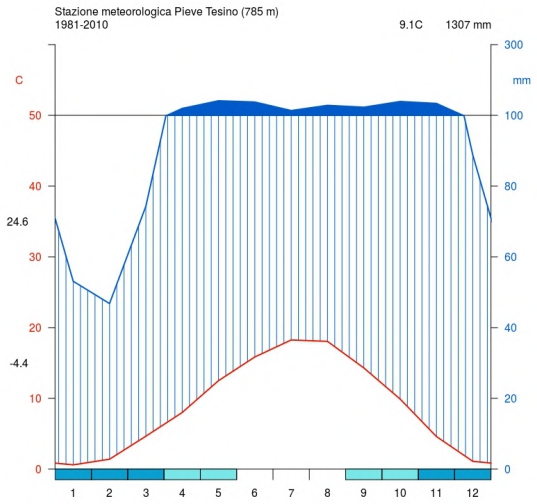
(b)



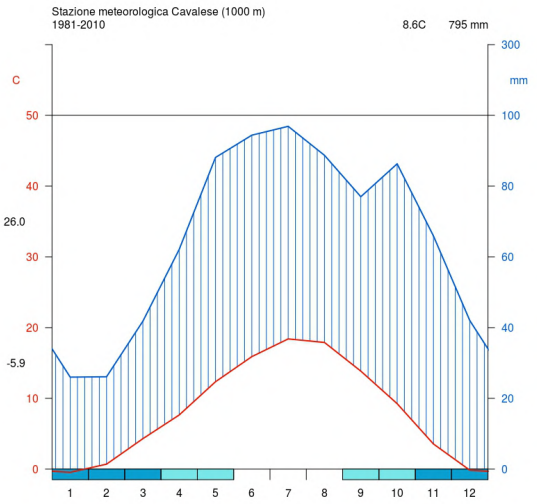
(c)



(d)



(e)



(f)

Figure 3.5: Climograms from *Walter & Lieth* of (a) Pejo, (b) Cles, (c) Trento Laste, (d) Sant'Orsola Terme, (e) Pieve Tesino and (f) Cavalese. Source: Meteotrentino

important cyclones generally come to the region from west and southwest, occasionly, but highly related with intense events, from southeast.
Relatively more closed zones are less exposed to south flows and receive a small amount of precipitation. For example the Avisio valleys, Fiemme and Fassa, represent this situation.

Important differences are present also in the rainfall regime, that diverges from the pre-alpine valleys to the interior valleys (see Figure 3.4 and Figure 3.5).
In the pre-Alpine sector of the Trentino Alps, represented by Valle dell'Adige, Valle del Chiese, Alto Garda and Valsugana, the annual rainfall pattern presents two precipitation maxima in spring and autumn and two precipitation minima in summer and especially in winter.
On the contrary, the areas far from the Po Valley and at high altitudes, show an alpine rainfall pattern with a seasonal precipitation maximum in summer due to the high frequency of thunderstorms.

Important to mention is the frequent presence of *föhn* winds. This wind is driven by pressure gradient between north and south of the Alpine divide. *Föhn* wind is characterized by dry and relatively warm air and by intense gusts of wind.
For Trentino, the appearance of *föhn* winds can occur when the air pressure difference (at sea level) between Innsbruck and Verona is about 10 hPa (Meteotrentino). If the *stau* occur on the austrian side of the Alps, than south directed *föhn* winds appear in many italian valleys.

3.3 Administration

3.3.1 Autonomy

The Autonomous Province of Trento, as well as the Autonomous Province of Bolzano, represents an unique example of autonomistic administration inside the country. This is due to the events that followed the end of 1^{st} and 2^{nd} World War.
After the 1^{st} World War the territory of Trentino and Südtirol were administrated together in the *Venezia Tridentina* region (Wikipedia). This institution was soon replaced by the advent of the Fascist, which ruled for more than 20 years. After this 2 decades, with the capitulation of Mussolini, Trento, Bolzano and Belluno were united and controlled by the Nazi Germany under the OZAV (*Operationszone Alpenvorland* in German, *Zona d'operazioni delle Prealpi* in Italian).
With the end of World War II and the fall of the dictatorship, the German-speaking part of the region wanted to be annexed to the Austrian Republic, while the Italian-speaking part demanded autonomous status for the region under Italian administration.

First autonomy statute

Accompanied by the relocation of some municipalities from Trentino to Südtirol, in 1946 the request for autonomy was satisfied and the *Autonomous Region Trentino-AltoAdige/Südtirol* was created.
On 5^{th} September, within the Treaty of Paris, the italian Minister of Foreign Affairs Alcide

Degasperi, born in Trentino and already member of the Chamber of Deputies in Vienna, signed with the austrian colleague Karl Gruber an agreement under the recognition of Italy and Austria for the protection of the languages minorities and thus the region's autonomy (*Accordo Degasperi-Gruber* in Italian, *Gruber-Degasperi-Abkommen* in German).
Considering this premises the first autonomy statute was signed in 1948.

Second autonomy statute

During a first calm period *Democrazia Cristiana-DC* and *Südtiroler Volkspartei-SVP*, the two main parties, worked together ensuring a good cooperation.
After the 1955 the tension between the speaking groups rise up leading to strong conflicts and even to terroristic attacks organized by the organization *Befreiungsausschuss Südtirol-BAS* (*Comitato per la liberazione del Sudtirolo* in Italian), which was founded in 1956.
After long negotiations, also with the Austrian government, in 1972 entered into force the second autonomy statute and large part of the region's competencies passed to the provinces with the result that they administrate separatly their own territory with legislative power. For this reason, the region begins to be only a connection organ within this freamwork.
Till nowadays new mandates are transfered from the state to the autonomous provinces like for example healthcare, education, work and transports. The financing of all this activity is made throught the withhold of the taxation's 90%.
The third autonomy statute is now under development by the provincial governments.

3.3.2 Internal Administration

To improve the financial efficiency is ongoing a procedure for the unification of small municipatilies all over the provincial territory. From 223 in 2009, the actual number of municipalities is reduced to 177, which are gathered in 15 valley communities (*Comunità di valle* in italian). The main municipalities are Trento (117.000 inhabitants), Rovereto (39.000 inhabitants), Pergine Valsugana (21.000 inhabitants), Arco (17.000 inhabitants) and Riva del Garda (16.000 inhabitants). Trento and other 3 municipalities are not part of any valley comunity.

The valley communities are listed in Table 3.3 and displayed in Figure 3.6.

3.3.3 EGTC Tyrol-South Tyrol-Trentino

Since the beginning of the 1990s, the two Italian provinces of Trentino and Südtirol have established relations with the Austrian Land Tirol with a view to setting up an international body.
The Madrid Framework Convention of 1980 allows cross-border cooperation between communities or territorial authorities. In 1991, the first joint session of the state parliaments of Tyrol, South Tyrol, Trentino and Vorarlberg took place in Innsbruck. After twenty years of cooperations and meetings called *Dreier Landtag*, in june 2011 the EGTC's charter was signed at Castel Thun (*European Grouping of Territorial Cooperation-EGTC*, *Gruppo Europeo di Cooperazione Territoriale-GECT* in Italian, *Europäische Verbund für territoriale Zusammenarbeit-EVTZ* in

Table 3.3: Valley communities of Trentino

n.	**Denomination**	**Municipalities**	**Inhabitants**	**Administrative center**
1	Comunità territoriale della Val di Fiemme	11	19.984	Cavalese
2	Comunità di Primiero	5	10.147	Tonadico
3	Comunità Valsugana e Tesino	19	27.439	Borgo Valsugana
4	Comunità Alta Valsugana e Bersntol	18	52.869	Pergine Valsugana
5	Comunità della Valle di Cembra	11	11.334	Faver
6	Comunità della Val di Non	29	39.464	Cles
7	Comunità della Valle di Sole	13	15.709	Malè
8	Comunità delle Giudicarie	25	37.775	Tione di Trento
9	Comunità Alto Garda e Ledro	7	48.899	Riva del Garda
10	Comunità della Vallagarina	17	89.288	Rovereto
11	Comun General de Fascia	7	9.982	Pozza di Fassa
12	Magnifica Comunità degli Altipiani cimbri	3	4.550	Lavarone
13	Comunità Rotaliana-Königsberg	8	29.229	Mezzocorona
14	Comunità della Paganella	5	4.914	Andalo
15	Comunità della Valle dei Laghi	3	10.595	Vezzano

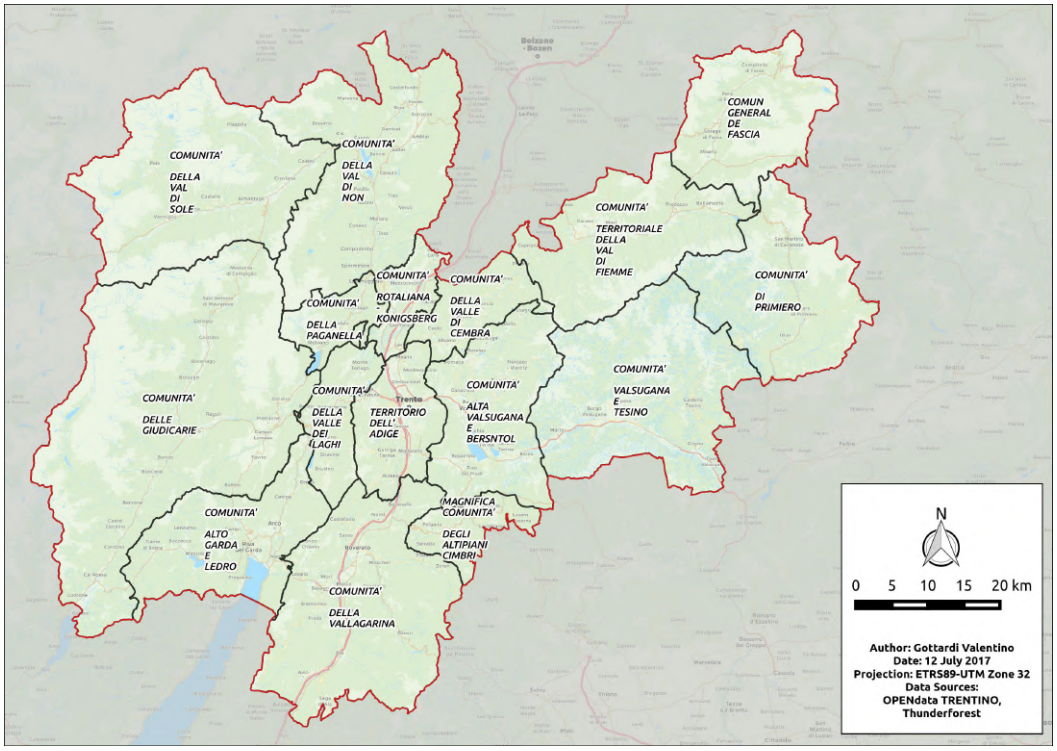


Figure 3.6: Valley communities of Trentino

German) posing the base for the european and national acknowledgement of the *European region Tyrol-South Tyrol-Trentino*, generally called with the Italian name *Euregio Tirolo-Alto Adige-Trentino* or with the German one *Europaregion Tirol-Südtirol-Trentino*.

The several activities of the European region are various, and goes from the direct collaboration with the European Commissions to the support of interregional Programs, passing by publications, sponsorship of student's activities and more (Europaregion).
Between many projects, quite important is now the coordination role of the future improvement of local and alpine mobility and transport politics. Other example is the ongoing creation of a common and multilingual avalanche forecast bulletin called ALBINA, founded by the Interreg Italia-Österreich European Regional Development Fund.

3.4 Economy

The economy in Trentino relates to the geographical position between north and south Europe, between German speaking and Latin speaking areas. This condition favours, nowadays and in the past, the trades.
In the Middle Ages the textile industry was essential, as well as timber trade and mining activity. Before the first World War, Arco, Levico and Roncegno become recognised as vacancy and thermal destinations between many austrian personality.
A particular feature is the widespread cooperative sector (with more than 100.000 associate in Trentino). This kind of companies are active in the credit management, transformation of agricultural products (e.g. dairying, wine production and fruit conservation) and distribution.
The '70s were the best economic period, thanks to the strong development of the Italian North-East and to the extension of the autonomy.

3.4.1 Primary sector

The primary or raw materials sector in Trentino is composed mainly of agriculture, forestry and mining. The spatial distribution of forests, pastures and agricultural lands is represented in Figure 3.7.

Fruit growing, particularly apple growing, is the most important form of agriculture, followed by vines and livestock farming. Less represented but still part of many farms are small fruits (strawberry, raspberry, blueberry and blackberry), cherry, horticulture, olive growing and cereal growing.
According to the 2010 agricultural census, the total cultivated area was 400.000 *ha*. Almost 110.000 *ha* as pasture and meadow, 22.000 *ha* as fruit and grapevine farming and 3.000 *ha* for cereals. The rest was forest. These areas were used by about 16.500 farms. About the 15% of the companies have livestock with 45.000 bovine, 27.000 ovine, 5.000 caprine and 3.000 equine and pigs (Agricultural service PAT).
Many products, such as wine and cheese, are protected by DOC, IGT, DOP and IGP designations.

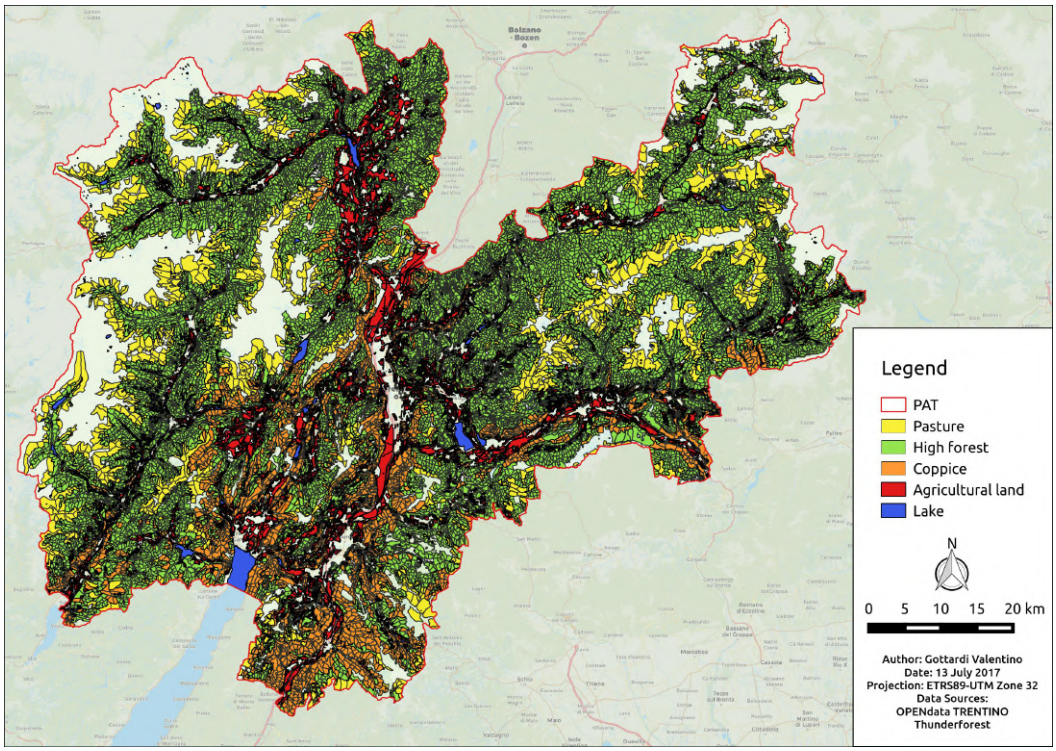


Figure 3.7: Primary sector in Trentino

Forest covers the 63% of the provincial area, thus silviculture is quite an important resource expecially for many public administrations (e.g. municipalities), considering that the 76% of the 390.463 *ha* of forest are public property (Forest and wildlife service PAT). Especially spruce forests are of excellent quality, like for example the population on Lagorai range in Fiemme valley.

Mining activity was widely spread in the past. Nowadays is concentrated in the porphyry extraction. Main use of this stone type is the paving of roads and squares, the roofing of typical houses and many new applications. The extraction zone is located on the western end of the Lagorai range, precisely on the territory of Albiano, Fornace, Cembra, Meano, Lona-Lases and Baselga di Pinè.

3.4.2 Secondary sector

The secondary sector or manufacturing is an important member of the economy, since about the 30% of the working population is active in this sector (Statistic service PAT) and almost one third of the provincial GDP (Gross Domestic Product) is attributable to the manufacturing.
The industries are characterized by small to medium dimension and are centered in the Adige valley, Vallagarina and Valsugana, but other industrial areas of local relevance are distributed all over the Trento province. The most represented working branches are textil, construction, mechanical engineering, wood and paper production chain and porphyry manufacturing. In many cases this sector is strongly supported by handicraft enterprises, distinguished by small dimension and familiar conduction.
Related to the cooperative form of many agricultural activities, is the presence of numerous food

manufacturing firms, which works mainly in the fields of conservation and distribution of apples and small fruits, wine transformation and trade, spirits production as well as fresh milk collection and cheese factory.

Thanks to the orography of the area and to the abundance of water, during the last two centuries, a large number of hydro-powerplants were constructed along the Adige river and all over the minor valleys. Just to focus the phenomenon, in 1889 the Ponte Cornicchio hydropower station by Trento was the first to be opened in the whole austro-hungarian Monarchy territory and the Serso hydropower station by Pergine Valsugana was the first three-phase alternating current in 1893 (Lappi 2008).
Expecially to mention is the Santa Giustina reservoir on the Noce river, with a high of $152,50\ m$ and a water volume of about 182 *million* m^3 (Progetto Dighe). Other wide systems of reservoirs and power stations are present in other valleys like for example the one on the upper reach of the Sarca torrent.

3.4.3 Tertiary sector

The tertiary or service sector is Trentino's most important occupation and source of income. This is because a considerable number of public services, normally under national administration, are managed directly by the Autonomous Province. For this reason almost 50.000 agents work, with various tasks (from administration, to healthcare and school), in the provincial offices (l'Adige, 29 June 2014).

One of the key activities is tourism. The tourism industry started at the end of the 19^{th} century. Thanks to the mild climate and the presence of thermal springs, many destinations experienced a boom period. At that time various infrastructures were built, first for military purposes, such as the *Great Dolomites Road* from Ora to Cortina d'Ampezzo.
A new revival of tourism began after the Second World War with the opening of many winter sports resorts. Today the tourist offer is very varied and ranges from active to cultural activities and is spread over the four seasons.

3.4.4 Quaternaty sector

The quaternary sector, also known as advanced tertiary sector, was recently joined into the economic hierarchy. This sector includes all the service enterprises with high added value and technology. The core business of this sector is based on know-how, intellectual services like research and development (R&D), training-education, consultation and ICT (information and communication technologies).
Thanks to the presence of the *University of Trento*, other research institutes like the *Fondazione Bruno Kessler-FBK*, the *Fondazione Edmund Mach-FEM* and *Polo Meccatronica* as well as start-up development centers like the *Progetto Manifattura* many new and innovative enterprises are encourage of move to Trentino.

3.5 Transport

Since centuries, the Brenner Pass as well as the Resia Pass, assumed a strong importance in the North-South trade and communication route. In both cases the Adige valley is the natural transport corridor used across different periods in time. Assuming this, the Austro-Hungarian Monarchy, began the construction of the *Brenner Railroad* (*Ferrovia del Brennero* in Italian, *Brennerbahn* in German) based on projects of the Tyrolian engineer Luigi Negrelli and German one Karl von Etzel. The three sections of the railroad were inaugurated in the period 1859-1867. The Brenner Railroad was the first to cross entirely the Alps passing one of the lowest alpine pass (only 1371 m *a.s.l.*).
In the province of Trento the total length of railroads is $204,7$ km (Europaregion). Beyond the main railroad, which runs along the Adige river, other two lines are present.
Along the Noce river was constructed and inaugurated in 1909 the *Tranvia Trento-Malè* now *Ferrovia Trento-Malè-Mezzana*, a 65 km long railway that cross Piana Rotaliana, Val di Non and Valle di Sole.
From Trento to Mestre runs the *Ferrovia della Valsugana* that follows the upper course of the Brenta river. This railroad is 157 km long and the stretch from Trento to Tezze di Grigno was inaugurated in 1896. The other section, from Mestre to the tyrolean boarder was then completed in 1910.

In the past the following railways were active:

- *Ferrovia Mori-Arco-Riva* or *Lokalbahn Moor-Arck-Reif* (from 1891 to 1936);
- *Tranvia Dermulo-Fondo-Mendola* or *Lokalbahn Dermulo-Mendel* (from 1909 to 1934);
- *Ferrovia della Val di Fiemme* or *Fleimstalbahn* (from 1917 to 1963).

The road network is even distributed over the provincial territory with different road classes, from highway to municipality road, passing by national and provincial roads. The total length of roads existing on the provincial area (except municipality road) is 2552 km (Europaregion).
The most important road directive is the $A22$ highway *Autostrada del Brennero* or *Brennerautobahn*, which together with the Austrian $A13$ highway connect the Po valley with Bavaria. This highway was completed in the 1974. On the same corridor runs also the national road *S.S. 12 dell'Abetone e del Brennero*.
Another important road is the *S.S. 47 Strada Statale della Valsugana*, often used because it ensures a shorter and cheaper connection between the Mestre harbour and Brenner Pass. For touristic purposes also the connection with Garda Lake and the dolomites are essential. In addition to this, the big amount of commuters that every day travel to the administrative centers imposes the renovation of the valley connections.
The forest road network is wide spread all over the forested lands, ensuring the possibility to reach remote and high elevated areas also for a touristic purposes.

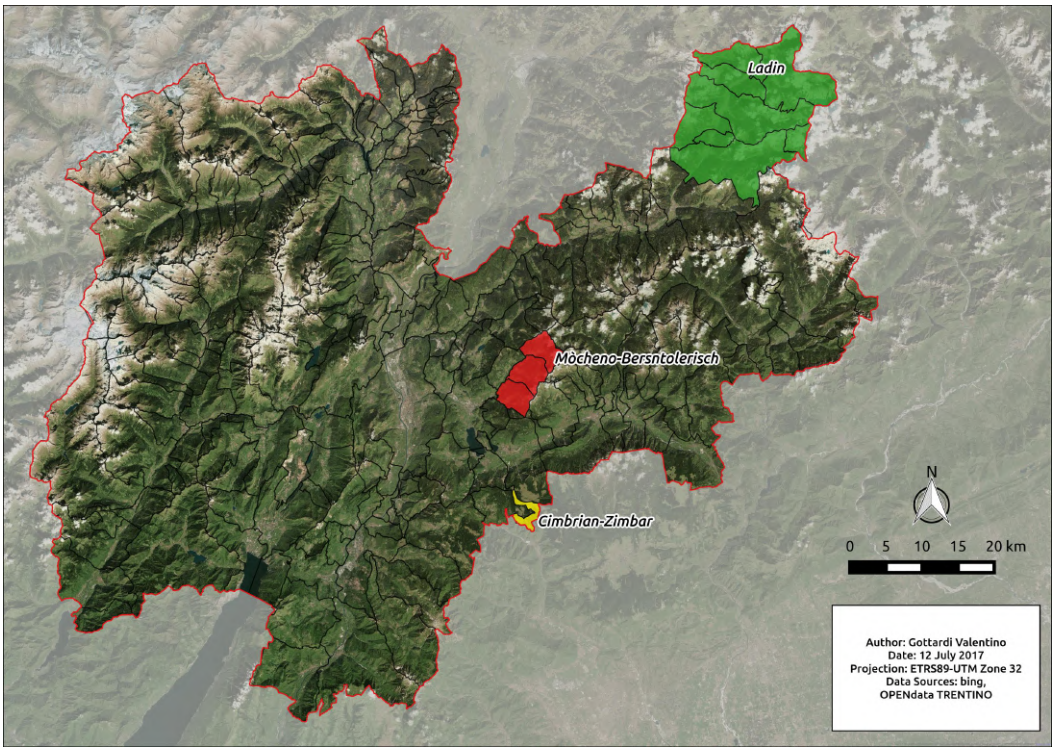


Figure 3.8: Linguistic minorities of Trentino

South of Trento, near Mattarello, the small commercial airport *Gianni Caproni* is located. The airfiel of Trento was constructed in 1914 north from the city. The structure was then moved to the current location in 1969. The nearest international airports are placed in Bolzano, Innsbruck, Verona and Bergamo.

3.6 Languages

According to the regional statute, every regional act and document has to be written in italian and german. The three recognised linguistic minorities in Trentino are Ladin, Cimbrian-Zimbar and Mòcheno-Bersntolerisch. Figure 3.8 displays the spatial extend of this languages on the provincial territory (Language minorities PAT).

Ladin

The five valleys around the Salla massif (Fassa in Trentino, Badia and Gardena in Su¨dtirol, Livinallongo and Ampezzo in Veneto) are home to the Dolomite Ladins.
Dolomite Ladin or Central Ladin is spoken by about 30,000 people and is part of the vast Rhaeto-Romanic language system, which also includes Romansh in the canton of Grisons (or Western Ladin, 40,000 speakers) and Friulian in the Italian Friuli region (or Eastern Ladin, 600,000 speakers in total, of which 420,000 are regular speakers).
The ancient area of spread of the Rhaeto-Romanic languages stretches from the source of the Rhine to the Adriatic Sea and has been progressively reduced by emigration and the linguistic

influence of the surrounding areas.
The dolomitic ladin spoken in Fassa valley is called *Fascian* and has three variants: *Moenat* spoken in Moena; *Brach* spoken in Soraga, Vich-Vigo, Poza-Pozza; *Cazet* spoken in Mazin-Mazzin, Ciampedel-Campitello, Cianacei-Canazei.
According to the 2011 census, on the 9909 Fassa valley inhabitants, 8092 declare themself as Ladins. To mention is also the affinity to the ladin of the dialects spoken in Non and Sole valleys. The common origin is acknowledged by linguists but not yet politically recognised.

Cimbrian-Zimbar

The Cimbrian-Zimbar is spoken in Luserna-Lusérn, a small village 40 *km* to the southeast of Trento. It still be used in some other municipalities of the provinces of Vicenza and Verona in Veneto. It's defined by linguists as the "more ancient periferic german parlance" and still regularly use by the 90% of the local population of Luserna.
The origin of this linguistic island dates back to the decade 1053-1063 A.D., when a famine forced the bavarian population of the Benediktbeuern convent's territory to move to the south of the Alps. The first stage took place in the today's 13 Comuni Veronesi on Lessini mountains. Than follows the expansion toward the 7 Comuni Vicentini. Only in 1216 A.D. also the plateaus of Folgaria and Lavarone were colonized with the permission of the princebishop of Trento Friedrich von Wangen and later the Luserna-Lusérn village was founded.
During the first World War Luserna-Lusérn was right on the fighting line and all the villagers, more than 1000 inhabitants, had to move to Bohemia.
1072 people claim themself as Cimbrian on the last demographic census of 2011.

Mòcheno-Bersntolerisch

The Mòcheno-Bersntolerisch language still spoken in some villages of the Fersina valley (known as *Mòcheni valley* or *Bersntol* in Mòcheno). More precisely in the municipalities of Vlarotz-Fierozzo, Garait-Frassilongo und Palai en Bersntol-Palù del Fersina.
With the approval of the local ecclesiastical hierarchy and the feudal nobility (with bavarian origin), bavarian population settled the Fersina valley from the 13^{th} century (some groups came also to part of the nearby Piné plateau, Pergine Valsugana and Roncegno in Valsugana, proved by existing surnames and village's arrangements). This germanic migration was common between the 12^{th} and the 14^{th} century in various part of Trentino and Veneto between the river Adige and Brenta.
The first colonists were farmers (called *roncatores*) while the second immigration wave, between the 15^{th} and the 16^{th} century, were miners (called *cànopi* or *knòppn*) and came to this area to exploit silver, copper and iron mines.
In the 18^{th} century this population was autorized to perform itinerant commerce all over the empire. This helped the conservation of the original language, thing that not happened for the other german colonies of eastern Trentino. The men traveled in summer to sell various goods,

they were called *krumern* or *clomer*.
From the 2011 census, the Bersentoler are 1660.

Chapter 4

DATA

All the presented analysis base themselves on the availability of various databases. They are provided by the SBM which continually fills in and revises them in order to have usable materials and information sources.
The available databases for this thesis work include characteristics of:

- **catchments**, geomorphological features, climatic information and other characteristics of torrent catchments;
- **fans**, geomorphological features of detected fan areas;
- **events**, informations about events occured in the provincial territory.

On the *Trentino OPENdata* homepage, a large number of freely usable database regarding the provincial area are available.

4.1 Catchment database

This database contains information, in the form of measurements and indices, on the geomorphological characteristics of catchment areas. In addition to these parameters, representative information on geology and precipitation is provided.
The latter, the geology and precipitation values, are derived from previous studies and analyses commissioned by various PAT departments. The geomorphological parameters, on the other hand, have recently been quantified directly by the SBM for this and other analyses.

The automate precess for the identification and description of watersheds is based on a DEM (*Digital Elevation Model*) of the PAT with 10 m resolution. A code, written in R, extracts the provincial water network and defines an univocal *torrent ID*. Using a GIS procedure, for each torrent, catchment and fan area are identified. In this way it is possible to create a *dual system catchment-fan* characterized by a common torrent ID. In the end, this gives the possibility to evaluate, separately or in a combined way, informations about geomorphological features of fan, basin and relative torrent.

An important note for this procedure is that the current scripting developed by the SBM only considers basins with a recognisable fan area. Streams with eroded fans, due to the large transport capacity of the valley river or a narrow valley floor, are therefore not detected by the automated procedure. Considering that these databases are constructed to describe the probability of generating intense torrential processes, i.e. debris flows, only small alpine basins are represented. For this reason, basins with a contributing area larger than 10 km^2 are also excluded because they are considered prone to generate only clear water outflows or bottom load transport events.
These exclusions lead to a partial representation of the whole provincial area. This fact is not problematic because the analysis is only focused on processes that potentially reach the fan area, where generally exposed objects and people are present. At the same time this misrepresentation of the actual territory can lead to difficulties and uncertainties in the creation of models.

The number of catchments detected by the SBM system is 1697. Magenta polygons in Figure 4.1 display the spatial distribution of the detected small alpine catchments of the PAT territory.

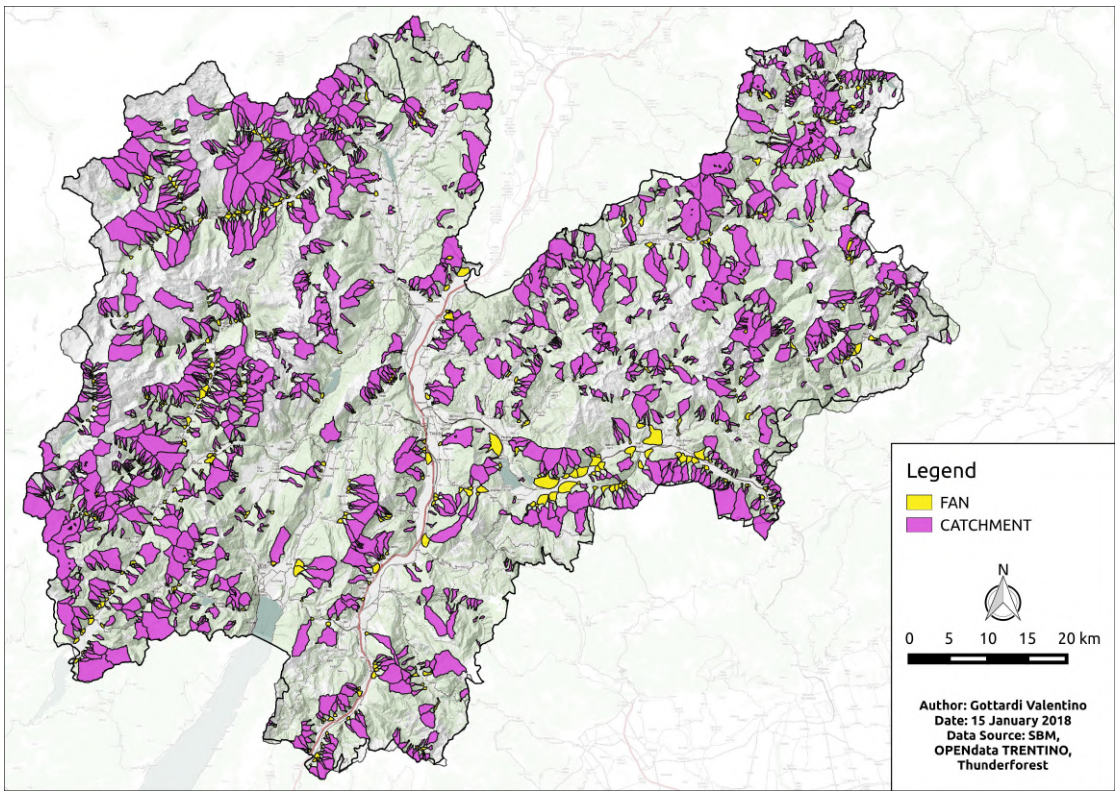


Figure 4.1: Considered catchments and fans in the Autonomous Province of Trento

20 basic arguments are available for the description of catchment's characteristics. They are listed in Table 4.2. Most of them are geomorphological information and indexes but also precipitation quantile and geological information are present.

The variable *classid* reports the univocal identification number (ID) of the basin. *barea* gives the plane surface of the catchment in square meters instead *bperim* reports the longitudinal extension of the watershed in meters.

Basic information about the vertical position of basins are given by the variables *bhmin*, *bhmax* and *bhmed*. They express, in meters above sea level, respectively the minimum elevation of the catchment (outlet), the higher point of the watershed and the average elevation.
Basin's morphology is displayed also with minimum, maximum and average slope angle expressed in degree (sexagesimal system) by the variables *bslo_min*, *bslo_max* and *bslo_med*.
Different geomorphologic indices can be used for the analysis of a watershed if its shape is taken into consideration. One of the most frequently used index is the Gravelius's index, which is defined as the relation between the perimeter of the considered watershed and the one of a circle having a surface equal to the watershed, see Equation (4.1). The more the index is close to the value of 1 and the more the shape of the watershed is circular. In the database this variable is named *bgravelius*.

$$bgravelius = \frac{bperim}{2 * \sqrt{\pi * barea}} \sim 0.28 * \frac{bperim}{\sqrt{barea}} \tag{4.1}$$

Another used index is the compact index, named in this work *bicompatt*. It is the ratio between the perimeter of the watershed and the diameter of a circle having a surface equal to the watershed, see Equation (4.2).

$$bicompatt = \frac{bperim}{d_{equivalent\ circle}} = \frac{bperim}{2 * \sqrt{\frac{barea}{\pi}}} \tag{4.2}$$

The variable *bmelton* reports the Melton's Ruggedness Number. It is a dimensionless measure proposed by Melton (1965). It is basically a dimensionless slope index for the catchment area. The relative relief of the basin is measured incorporating measures of travel distance and available relief. A valuable characteristic of the Melton's Number is its capability for discriminating basins with debris flow potential from basins where sediment transport processes are dominated by bedload. Moreover, a power function can quantify the relation between fan gradient and Melton's Ruggedness Number (Efe et al. 2011). According to De Scally and Owens (2004), debris flow fans are supplied by basins with a higher Melton's Number, while the fans are steeper, smaller and less concave.
bintipsom represents the basin's hypsometric integral or hypsometric curve. As stated by Willgoose et al. (1998) this index puts into relation area and elevation of a basin segment summaryzing the form of a drainage basin in one single value. Looking at the hypsometric curve gives a general synthetic view of the watershed relief and appears clearly the slope of the upper, medium and lower part of the catchment and from them it is possible to understand torrent's erosion and transport capacity. The *bstadio* variable represents the basin's stadium index which makes possible to evaluate if the basin is in a young, intermediate or adult stadium of evolution, allowing also to asses the relative amount of available sediment material.
To characterize the sediment availability for a debris flow event, the lithology distribution can play an important role. Quaternary depositions are important sediment sources, but their distribution is not represented in a specifical map. For this reason the variable *big* introduces a modified *Geological index*. Considering the lithology of a basin, D'Agostino (1996) originally define this index as the weighted average of the scores as a function of the areas pertinent to each rock type

Table 4.1: Geological Index values, from D'Agostino (1996)

Lithology	**IG value**
Quaternary deposits	5
Schists and phyllites	4
Marls, marly limestone, siltstone, etc.	3
Volcaniclastic rocks	2
Dolomite and limestone rocks	1
Massive igneous and metamorphic rocks	0
Intensely fractured and weathered rocks for all lithologies	3-5

(see Table 4.1).
To highlight the likelyhood to generate debris flow processes, the variable *big1045mea* represents the geological index calculated only for the areas where the initiation of a debris flow is likely to occure. In order to fulfill this assumption, only the Geological Index informations of the 10 m resolution DTM cells with a d8 slope between 10° and 45° are considered. These reference values are choosen on the basis of various literature sources. Marchi and D'Agostino (2004), from the study of 127 basins in the Eastern Italian Alps, shown debris flow inititations up to stream reach inclination of 90%. A confirmation comes from Coe et al. (2008), which reports initiations on a steep and small basin in Colorado between 14° and 45°. Cavalli et al. (2017), poses as braking and deposition threshold the value of 10°. The attachment to the article 10-5 (*Criteri e metodologia per la redazione e l'aggiornamento delle carte della pericolosità*) of the provincial regulations L.P. 1/7/2011 n.9 defines as source area for a *fall process* every cell with slope higher than 44°.

The parameter *bquantmm* is the 100 years precipitation quantile for a 1 hour duration rainfall. It quantifies the amount of precipitation in one hour that is reached or exceeded with a probability of 1 to 100 per year and should represents the debris flow triggering probability. Borga (2003) calculates for the entire provincial territory a set of precipitation curves (called *LSPP* from the italian denomination *Linee Segnalatrici di Probabilità Pluviometrica*) with a return period of 100 years and duration of 15, 30 and 45 minutes and 1, 3, 6, 12 and 24 hours. Although torrential processes are sometimes connected with really short and intense duration rainfalls, the 1 hour precipitation quantile can be considered appropriate to describe triggering condition on alpine catchments. Moreover, the precipitation duration of 1 hour is characterized by the higher consistency because it is calculated from a long time-frame and a large weather station number (SBM). The distribution of the precipitation is largely uneven over the province, with higher sum to the South and in Fassa Valley. The lower quantity is measured to the North-West part.
bdensdren represents the basin's drainage density. It is the relationship between length of the torrent network and catchment surface and is expressed with the unit km/km^2. The drainage density can be labeled as coarse (< 8), medium (between $8 - 20$), fine (between $20 - 2000$) and very fine (> 2000).
Considering the main torrent reach, the variable *blungasta* reports the length in meters instead *bslo_asta* the average slope in degrees (sexagesimal system).
The last used index is the form factor, variable *bffactor*, as reported in Dong et al. (2009) and

Table 4.2: Catchment's descriptive arguments

Argument	Unit	Description
classid	-	ID of the basin (same for the connected fan)
barea	m^2	Basin's surface
bperim	m	Basin's perimeter
bhmin	$m\ a.s.l.$	Basin's lowest elevation
bhmax	$m\ a.s.l.$	Basin's eighest elevation
bhmed	$m\ a.s.l.$	Basin's average elevation
bslo_min	°	Basin's lowest declivity
bslo_max	°	Basin's eighest declivity
bslo_med	°	Basin's average declivity
bgravelius	-	Basin's Gravelius index
bicompatt	-	Basin's compact index
bmelton	-	Basin's Melton index
bintipsom	-	Basin's hypsometric curve
bstadio	-	Basin's Stadium index
big	-	Basin's geology index (from D'Agostino 1996)
big1045mea	-	Basin's geology index of initiation zones
bquantmm	mm/h	100 years precipitation quantile for 1h duration
bdensdren	km/km^2	Basin's drainage density
blungasta	m	Torrent's length
bslo_asta	°	Average torrent declivity
bffactor	-	Basin's form factor

others. The form factor index is defined as the catchment area divided by the square of the creek length, see Equation (4.3).

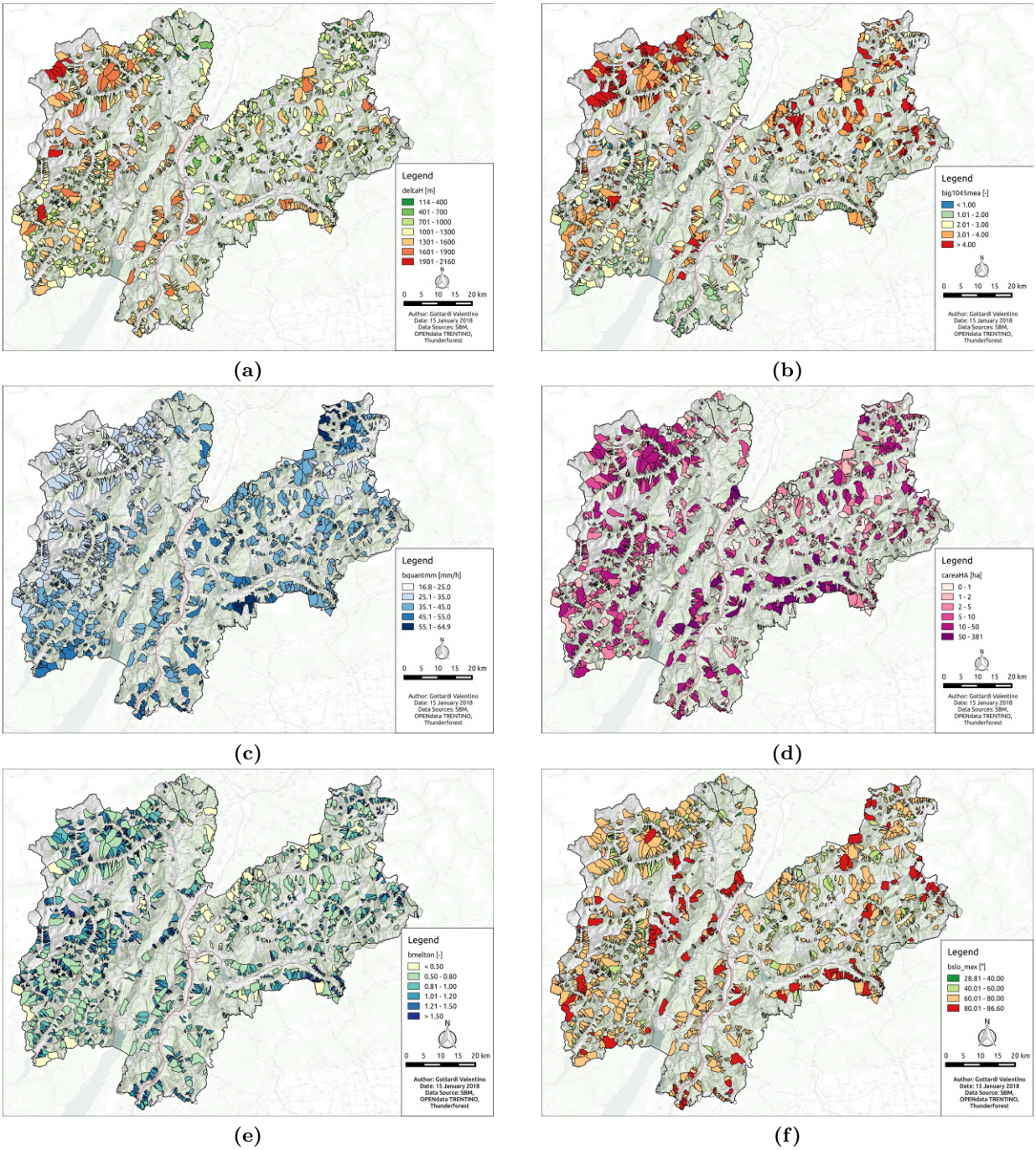
$$bffactor = \frac{barea}{L^2} \tag{4.3}$$

The statistical representation of the available catchment's parameters is represented by Table 4.3. Figure 4.2 shows, as example, the spatial distributions of some selected parameters from the entire watershed database. More precisely, spatial distributions of elevation difference, geological index from D'Agostino, 1 hour precipitation quantile, fan area, Melton's Number and maximum basin slope are displayed.

4.2 Fan database

The second database contains information on the geomorphological characteristics of the observed fan areas. The method already described used by the SBM to create the database identified a total number of 2584 fans. The identified fans are spatially distributed as shown by the yellow polygons in Figure 4.1.
Table 4.5 shows the basic arguments for the description of the fan characteristics. The 8 fan characteristics considered are described in broad terms in the following.

The variable *classid* reports the univocal identification number (ID) of the fan, which is the



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 4.2: Spatial distribution of some used parameters: (a) Difference in elevation, (b) Geological index, (c) Precipitation quantile, (d) Fan area, (e) Melton's Number, (f) Maximum basin slope

Table 4.3: Catchment parameter's statistics

Parameter	Unit	Min.	Max.	Range	Median	Mean	Standard deviation
barea	km^2	0.10	9.85	9.75	0.42	1.06	1.54
bperim	km	1.620	20.640	19.020	4.500	5.699	3.534
bhmin	$m\ a.s.l.$	77	2610	2533	1086	1114	518
bhmax	$m\ a.s.l.$	384	3694	3310	1998	2009	538
bhmed	$m\ a.s.l.$	341	3124	2783	1573	1592	509
bslo_min	°	0.0	32.6	32.6	3.8	6.3	6.8
bslo_max	°	28.9	86.6	57.7	67.8	66.3	11.4
bslo_med	°	8.3	65.6	57.3	34.5	34.6	7.8
bgravelius	-	1.27	3.60	2.33	1.83	1.91	0.32
bicompatt	-	4.54	12.87	8.33	6.54	6.81	1.14
bmelton	-	0.22	3.44	3.22	1.19	1.25	0.51
bintipsom	-	0.27	0.80	0.53	0.54	0.53	0.08
big	-	0.50	5.00	4.50	2.64	2.72	1.14
big1045mea	-	0.50	5.00	4.50	2.86	2.84	1.13
bquantmm	mm/h	16.77	64.89	48.12	37.72	37.59	8.25
bdensdren	km/km^2	0.19	21.24	21.05	4.75	5.25	2.81
blungasta	m	0.09	6541.54	6541.45	1091.70	1410.54	1047.19
bslo_asta	°	0.00	1.42	1.42	0.49	0.49	
bffactor	-	0.00	26.86	26.86	0.00	0.02	0.66

same of the connected basin, and permits to relate fan and watershed characteritics. *ambito_cnr* represents the hydrological pertinence to a given valley river contributing area. *id_opera* reports the identification number of present mitigation measures constructed on the fan area (a mitigation measure cadaster is available for the territory of the PAT but it is not considered in the present work).

danno_pote gives a broad information of the potentially damagable infrastructures present on the fan. This index is calculated with the multiplication of a relative value, which ranges from 0 to 1 and depends on the land use of every raster cell, and the corresponding cell's surface. Land use potential damage values are taken from the provincial civil protection regulation. Only land use with high importance and likelihood to be damaged by a torrential event are considered (see Table 4.4).

The argument *rasta* reports the identification number of the upstream torrent, which is different from the ID of watershed and fan. The geomophological characteristics of fans are described by 3 parameters. The first one is the *cslo_med*, which gives the average fan slope calculated with a GIS automated procedure (sexagesimal system).

The total plane area of the fan, in square meters, is reported by the variable *carea*. The area of the fan can be an important parameter to characterize the history of the catchment. Already Bull (1964) observed a morphometric relationship between fan area and drainage basin area in the form of a power function. May & Gresswell (2004) pointed out that the combined effects of drainage basin area, mainstream valley floor width and presence of debris flows in the last decades are positively related to the fan area. The same authors stated that fan area tend to increase with an increased valley floor width, by a nonlinear relationship. Erosion of the fan sediment takes

Table 4.4: Potential damage values of different land uses (PAT)

Land use	Relative value
Residential area	$1,00$
Primary road	$0,93$
Railway	$0,93$
Camping area	$0,90$

Table 4.5: Fan's descriptive arguments

Argument	Unit	Description
classid	-	ID of the fan (same for the connected basin)
ambito_cnr	-	Fluvial area of interest
id_opera	-	Eventual ID of presents mitigation measures
danno_pote	-	Potential damages (see below for description)
rasta	-	ID of upstream torrent
cslo_med	°	Average fan declivity
carea	m^2	Fan area
cvol	m^3	Fan volume

place expecially where the valley floors are narrow. In some cases, the deposits enter the main stream without forming a fan.
Ceriani et al. (2000), propose to use the volume of the fan in order to highlight the amount of sediment transported by the torrent during the centuries. The calculation of the fan volume can be performed using Equation (4.4) and is reported in the variable *cvol* in cubic meters:

$$cvol = \frac{1}{3} * carea * \Delta Hfan * \cos\left(Sfan * \frac{\pi}{180}\right) \tag{4.4}$$

where: *cvol* is the fan volume (in m^3), *carea* is the fan area (in m^2), $\Delta Hfan$ is the difference in alevation between the higher point of the depositional area of the fan (apex) and the lower point of fan toe and $Sfan$ is the mean fan slope (in m/m).

For the geomorphological parameters of the fan database, some basic statistics are performed and reported in Table 4.6. Only the fans directly connected to one of the 1697 catchments are taken into account to perform the statistics.

These fans are classified considering the type of the occured events. During the observation period, no events in the fan area have been reported for 1538 basins, which represent the 90.6%

Table 4.6: Fan parameter's statistics

Parameter	Unit	Min.	Max.	Range	Median	Mean	Standard deviation
cslo_med	°	1.8	45.9	44.1	15.6	16.2	6.46
cvol	m^3	2667	331409684	331407017	749677	4092216	16366396
carea	m^2	600	3809542	3808942	31140	79310	205229
danno_pote	–	0	1703909	1703909	7567	32594	98114

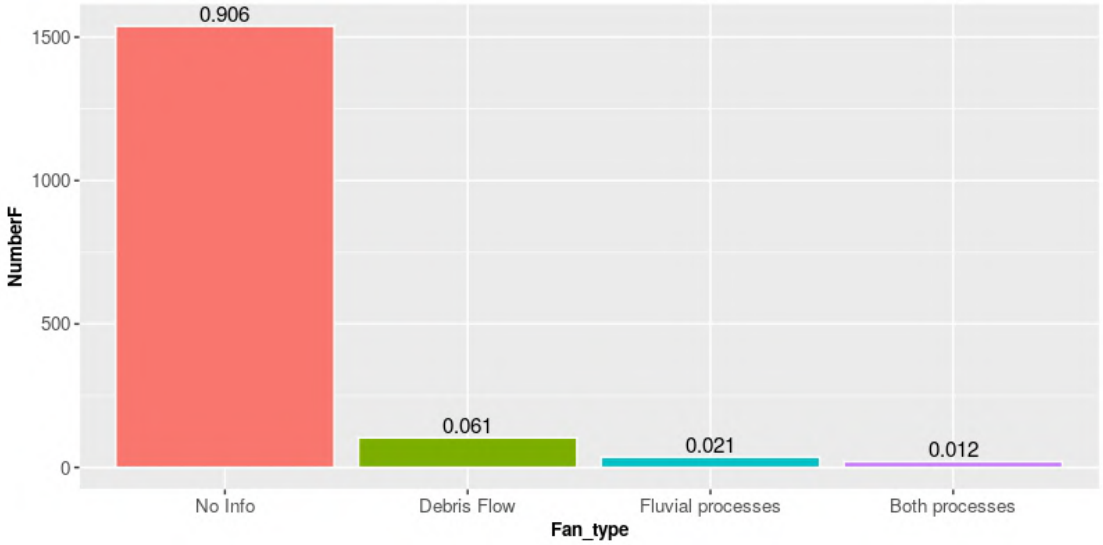


Figure 4.3: Fan classification depending on the reported events (numbers above the columns indicate the relative frequency), (PAT)

of the total (see Figure 4.3). 103 fans experienced only high intense processes as debris flows or debris floods. 36 basins generated only fluvial processes on the fans. These fans represent respectively the 6.1% and the 2.1% of the considered database. The remaining 20 fans (1.2%) shown a mixed behaviour with reports of both, high and low intensity processes. The fans with historical processes sum up to 159, representing the 9.4% of the total.

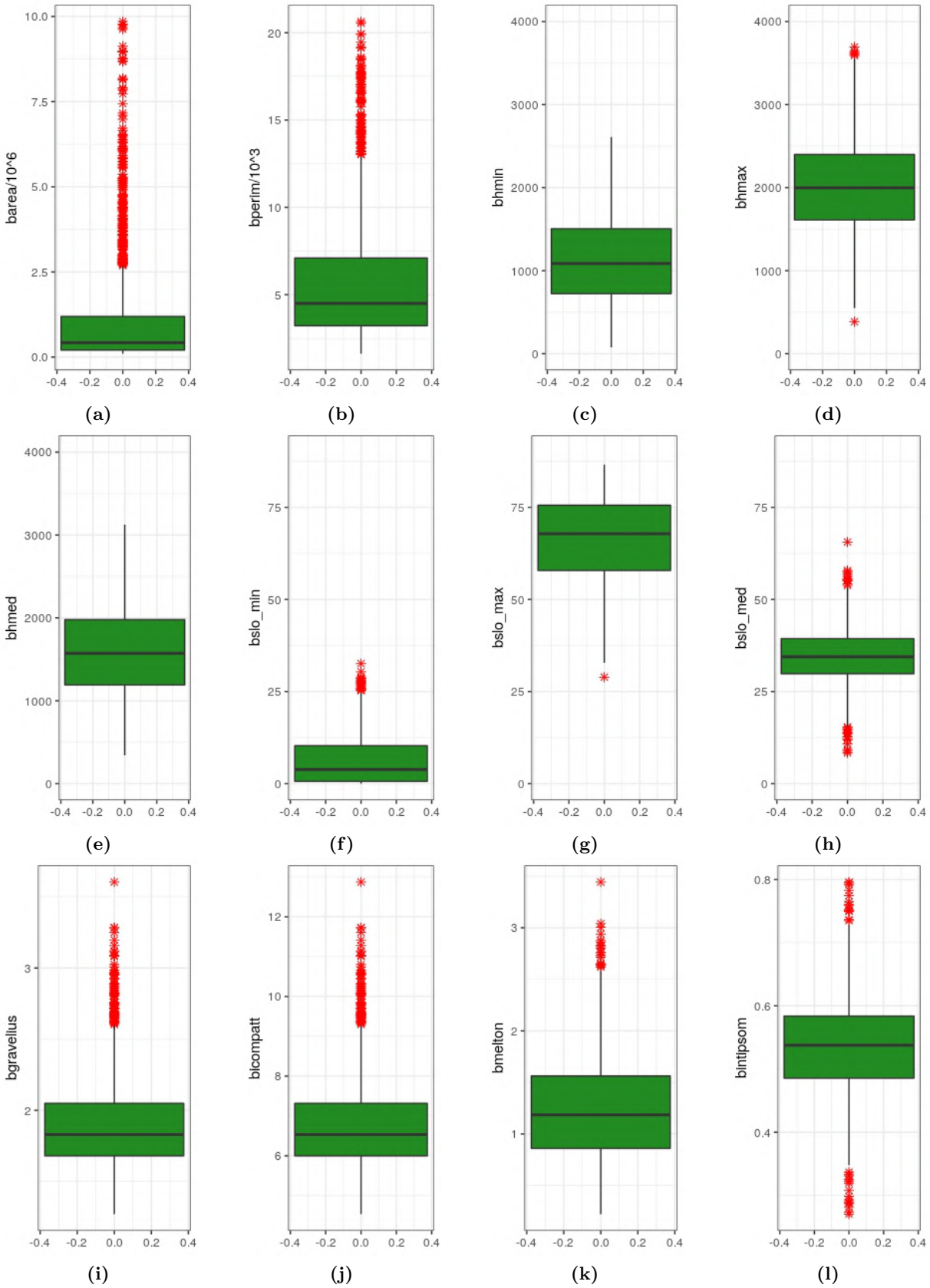


Figure 4.4: Boxplots of the used parameters: (a) Catchment area, (b) Catchment perimeter, (c) Minimum basin elevation, (d) Maximum basin elevation, (e) Mean basin elevation, (f) Minimum basin slope, (g) Maximum basin slope, (h) Mean basin slope, (i) Gravelius index, (j) Compact index, (k) Melton index, (l) Hypsometric curve.

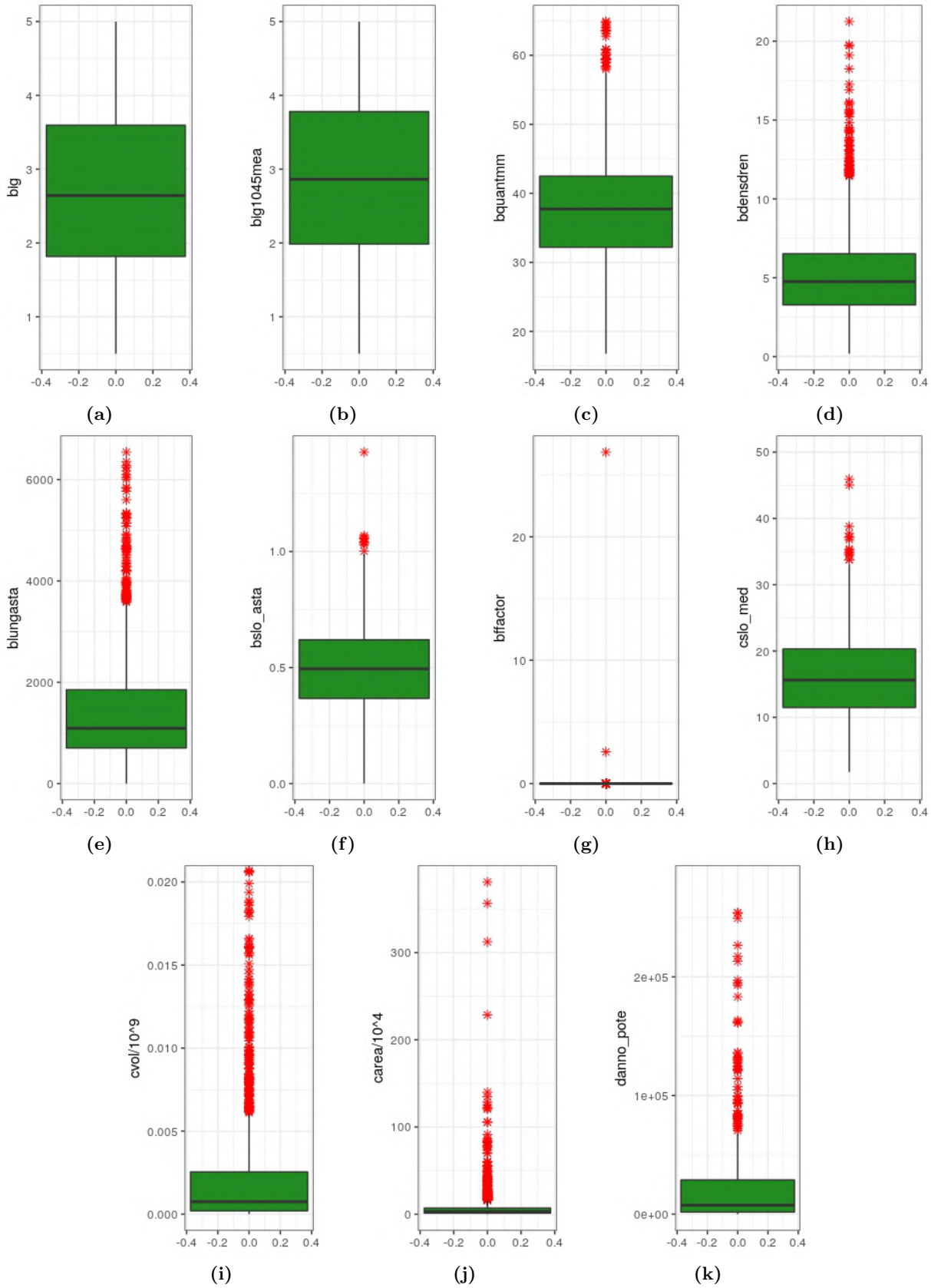


Figure 4.5: Boxplots of the used parameters: (a) Geology index, (b) Geology index initiation zone, (c) 100 years precipitation quantile for 1 hour duration, (d) Drainage density, (e) Torrent's length, (f) Mean torrent slope, (g) Basin form factor, (h) Mean fan slope, (i) Fan volume, (j) Fan area, (k) Potential damage.

4.3 Event cadaster

This database contains informations about the reported events, generated by all the water bodies on the provincial territory, from 1570 (mainly since 1750) to 2016. In total, the event cadaster contains 3522 events related to activity of torrents, rivers and lakes, happened on drainage basins and/or fans. It's easy to imagine that spatial resolution, as well as information completeness and reliability strongly depend on the considered time window and data source. Broadly, the uncertainty on the information veracity grows with the distance in time.

The available data are also characterized by various origin. The event cadaster was constructed and reviewed by the SBM a few years ago using several data sources. We can summarise them into the following groups:

- **event profile card**, event informations recorded by the SBM for each known event occured in the last 10-15 years;
- **ARCA database**, events described by the Geology Service of the PAT, excluding mass movements like landslides, rockfalls, deep-seated landslides, rock avalanches and other (many events are classified as landslides but probably related to torrential activity);
- **event chronicle**, descriptive recap of past events, included in the mitigation measure project for a given catchment;
- **articles** from newspapers and local bulletins;
- **other external data sources** like for example community and parochial registers and records.

The data source variability leads consequently to different precision and reliability of information.

Certainly, the completeness of the event series of a basin is not ensured or at least not for every single one. This is surely true for the catchments not occupied by important human activities (residential, industrial, tourist or agricultural use) or infrastructures (railways, roads, and others) like most of the high alpine basins or the watersheds completely covered by forests.
Also events of minor magnitude and intensity will be often not recorded, as well as all the events which stop in the transport zone and do not reach the fan area. In addition to unrecorded events, multiple counting of some of these events must also be considered, which generates damage in various locations of a fan.

A crucial limitation is also the definition of every process, which is in many cases not precise and univocal. This is mainly due to the unstady nature of phenomenon (e.g. continued transition from debris-flow to bedload transport, passing by hyperconcentrated flow, and conversely along a single torrent channel) and to the different knowledge and aims of reporters. Also the indication of a magnitude, when present, is not homogeneously reliable.

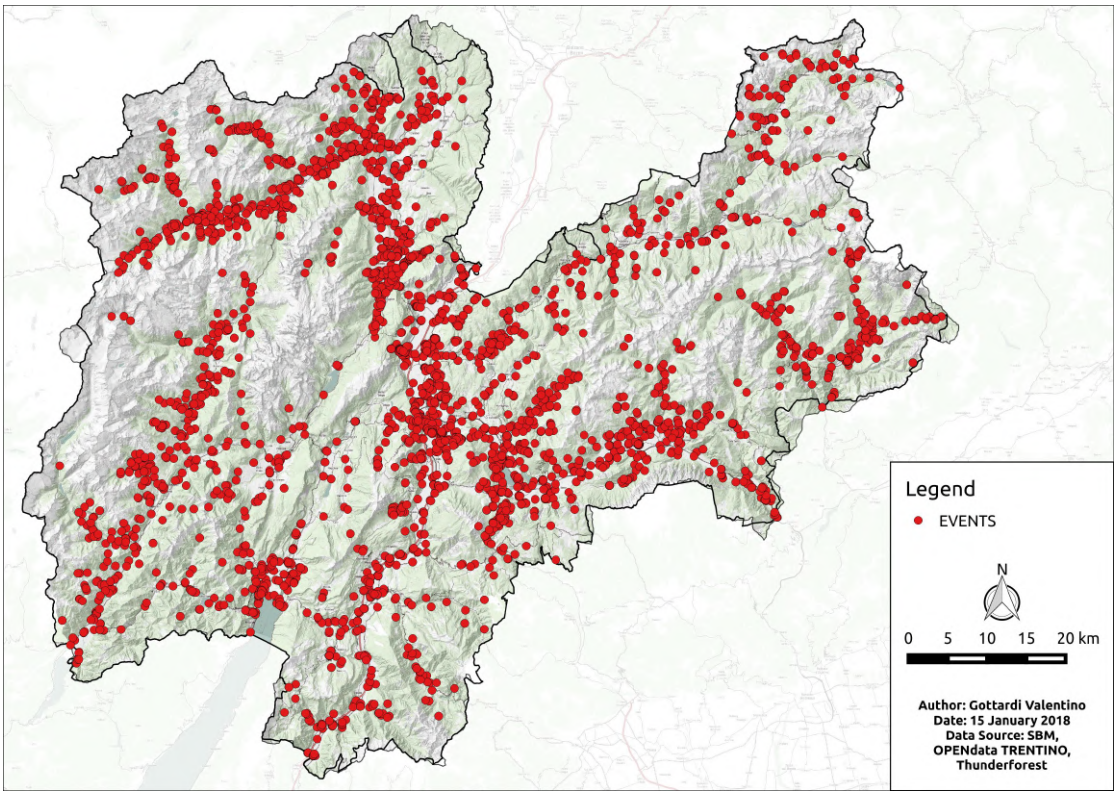


Figure 4.6: Reported events for the Autonomous Province of Trento

All this considerations about uncertainties imply the need of an incisive review and filtering of the events used for model building, fitting and validation.

A general overview of the spatial distribution of the reported events is visible in Figure 4.6. From this picture clearly appears that the signals are concentrated along the main valley's incisions.

For every single element presented in the Event Cadaster of the PAT, a wide number of informations are reported in form of a table database. In order to use only the arguments which give us interesting characteristics of the event, some of the table's columns were filtered out. The remaining important arguments are 13 and are broadly descripted in the following paragraph.
The variable *id_evsto* reports the univocal identification number (ID) of the event; *tipoevento* gives information about the reported type of event (this variable will be deeper described in the following); *att_tipoev* states if the information of the previous variable *tipoevento* is reliable or not, depending on information source etc.; *dataevento* reports the date of occurence of the event (for the more old events only month and year are present); *att_dataev* states the reliability of the date information; *localita* reports the place of occurence (different precision are present, from precise position of the damage to broadly the name of the valley where the event took place); *att_locali* states the reliability of the information about the place of occurence; *descrizion* reports the description of the considered event; *danni* describes the generated damages in term of casualties, damaged goods or monetary losses; *opere_post* tells if any mitigation measures where constructed after the event occurence; *fonte* reports the information source used to describe the

Table 4.7: Event cadaster's arguments

Argument	Description
id_evsto	ID of the event
tipoevento	Event's type (e.g. flood, debris flow, etc.)
att_tipoev	Reliability of the argument tipoevento
dataevento	Date of occurrence
att_dataev	Reliability of the argument dataevento
localita	Event's location
att_locali	Reliability of the argument localita
descrizion	Event's description
danni	Damage's description
opere_post	Mitigation measures constructed after the event
fonte	Information Source
corsoacqua	Name of the valley torrent
note	Notes

Table 4.8: Event's classification

Event's type	Number	Percentage
Debris flow	1180	33.5%
Landslide (torrent related event)	1104	31.3%
Flood	748	21.2%
Bedload transport	277	7.9%
Steady flood	98	2.8%
Flood in broad sense	85	2.4%
Debris flood	19	0.5%
Other	11	0.3%

event in the database; *corsoacqua* links the event to the generating watershed (using a certain torrent ID); if any annotation is present it is reported in the argument called *note*. Table 4.7 lists the descripted arguments used for the characterization of each reported event.

The 3522 events contained in the Event Cadaster are diveded into 8 different types of processes. As summarized in Table 4.8, the bigger number of recorded events refers to *debris flow* processes. They are 1180 representing one third of the total. Another big part, 1104 events representing the 31.3% of the database, is constituted by *landslides*. These processes are not properly described and they reliability is generarly low. They are probably not real landslides as described by Hungr et al. 2014, but more likely not well defined torrential processes or mixed events. An important source of uncertainties is also the lack of use of the technical name of debris flow in the Italian speaking part of the Alps. The database includes 748 *flooding* events, 277 *bedload transport* processes and 98 *steady flood* events (respectively 21.2%, 7.9% and 2.8% of the total). Other reported process types are *flood in broad sense* with 85 events or the 2.4% and *debris flood* with 19 events or the 0.5% of the total amount. Moreover 11 signals, equal to the 0.3%, are *undefined processes*. This distribution into process types is graphically displayed in Figure 4.7.

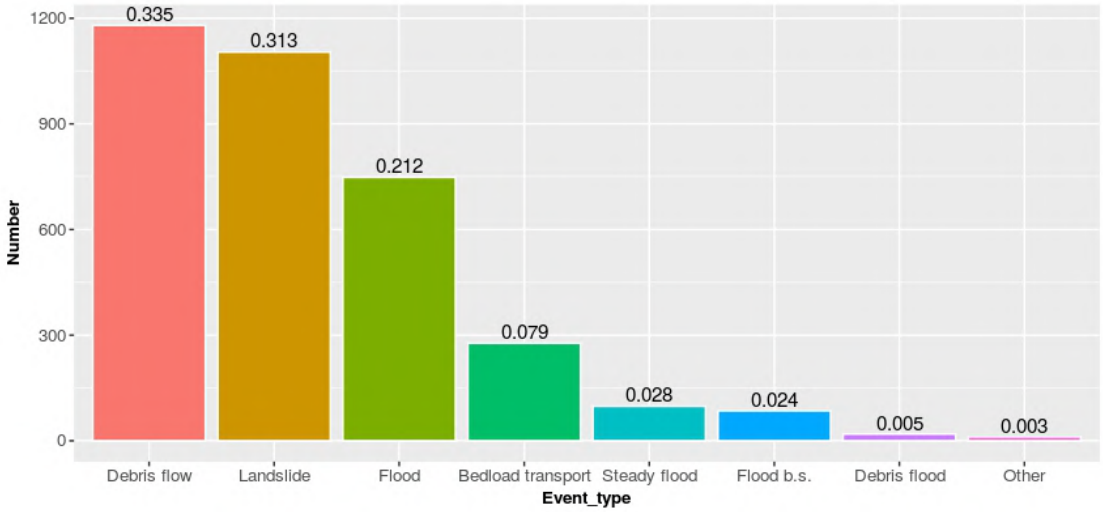


Figure 4.7: Classification distribution of the reported events in the Event cadaster (numbers above the columns indicate the relative frequency), (PAT)

Chapter 5

METHOD

5.1 Used softwares

All the analysis were performed mainly using the software ***R*** (more precisely the IDE ***RStudio***) and ***QGIS*** for the elaborations and to display the results.

5.1.1 Software *R*

Many interesting informations about the software *R* are available on `www.r-project.org` which is the official website of the *R* project. Basically, *R* is a programming language and software environment for statistical computing and graphics.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering and so on) and graphical techniques. Another important feature is the high extensibility of the functions thanks to the availability of many **packages**. A big variety of packages are available through the *CRAN (Comprehensive R Archive Network)* family of Internet sites covering a very wide range of modern statistics. In this way, *R* turns into an integrated suite of software facilities for data manipulation, calculation and graphical display and thus not only applicable to statistical tasks.

The software *R* is available as *Free Software* under the terms of the Free Software Foundation's *GNU General Public License* in **source code** form. It compiles and runs on a wide variety of *UNIX platforms* and similar systems (including *FreeBSD* and *Linux*), *Windows* and *MacOS*. The source code for the software envirnoment is written primarily in *C*, *Fortran* and *R*.
While R has a command line interface, there are several graphical front-ends available. Between the others widely used is ***RStudio***, which was used for this work.

R has many strengths. One of them is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Another one is that it is designed around a true computer language, and it allows users to add additional functionality by defining new functions. In any circumstance the user retains full control of the operations. These and other features improve the popularity of this software expecially among statisticians and data miners for developing statistical software and data analysis.

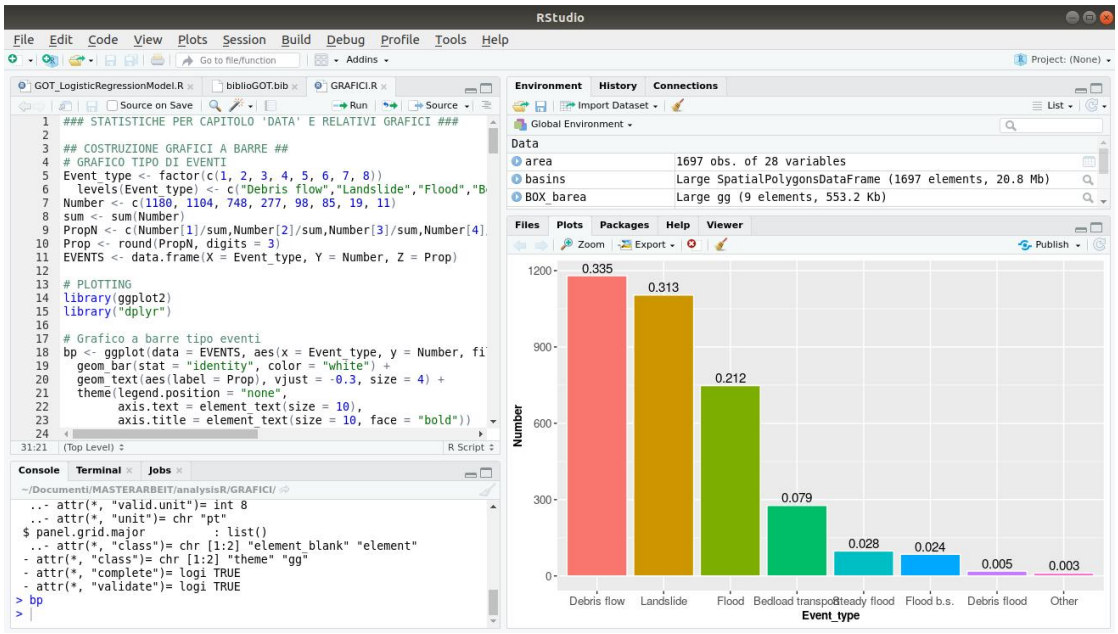


Figure 5.1: Rstudio's interface, version 1.2.5001

RStudio

The data alaborations were performed using *RStudio*, which is an *integrated development environment (IDE)* for *R*. This software combines an intuitive user interface with powerful coding tools and is ideal for beginners as well as for advanced users.

The interface, contains all the required tools to work with *R* like console, source, plots, workspace, history and an integrated help functionality as shown in Figure 5.1. The installation of new packages is easy and clear. An important feature is also the syntax highlighting editor (essential for a good code readability) with code completion. The code can be executed in various way and also directly from the source editor.
More precise informations can be found on the official website of the software on `www.rstudio.org`.

Utilized packages

The volunteer developers activity is clearly visible in the high number of available packages. The user-created packages are used to extend the capabilities of *R* with a variable number of functions. They allow specialized statistical techniques, graphical devices, import/export capabilities, reporting tools, etc.
The basic installation of *R* permits the utilization of a core set of packages. In addition to this, more than 11.000 packages are till now available at the *Comprehensive R Archive Network (CRAN)* and other repositories and easily usable after installation. The general source code to install and use a package is reported in the following script lines:

```
# install desired package (e.g. "ctv")
install.packages("ctv")
# recall package
library(ctv)
```

The available packages are visible at `https://cran.r-project.org/`. The *Task Views* section of the website offers an orderly distribution of packages into operational fields such as Finance, Genetics, Graphics, Machine Learning, Official Statistics, Spatial Statistics and others.
Informations about developer's community, package development support and different projects can be found on the mentioned website.

For the pourposes of this thesis, various packages were needed. Next paragraphs illustrate the main features and applicational fields of the used packages.

rgdal is the *R*'s interface to the *Geospatial Data Abstraction Library (GDAL)* which is used by other open source GIS packages such as *QGIS* and enables *R* to handle a broader range of spatial data formats. This package gives also access to the external library *PROJ.4* for projection/transformation operations.
Both raster and vector map data can be imported into and exported from *R* using respectively the appropriate **GDAL** and **OGR** functions.

proj4 provides a simple interface to the *PROJ.4* cartographic projections library. The package is used for lat/long projection and datum transformation. It is mainly used to transform geographic coordinates from one projection and/or datum to another.

corrplot contains tools to work with matrix and for the graphical display of correlation matrix and confidence interval.

dplyr is a tool for working with data frame like objects and is defined as a *Grammar of Data Manipulation*. It contains a lot of functions which correspond to the most common data manipulation tasks likes filter, mutation, selection, ranking, summarize and others also in combination.

caret is the short for *Classification And REgression Training* and is a set of functions for the creation of predictive models. The package contains tools for data splitting, pre-processing, feature selection, model tuning using resampling and variable importance estimation. Important feature is the uniform and standardized training and prediction operations for the different available models.

ggplot2 is a plotting system for *R* based on *The Grammar of Graphics* by Leland Wilkinson (2006). The user has to provide the data and tell the function how to treat and represent them. The drawing part is than automate but completely customizable by users. The functions of this package easily produce complex multi-layered graphics.

5.1.2 Software *QGIS*

As reported on the official website of the *QGIS* project `www.qgis.org`, this software is a professional *Geographic Information System (GIS)* and is a project of the *Open Source Geospatial Foundation (OSGeo)*.

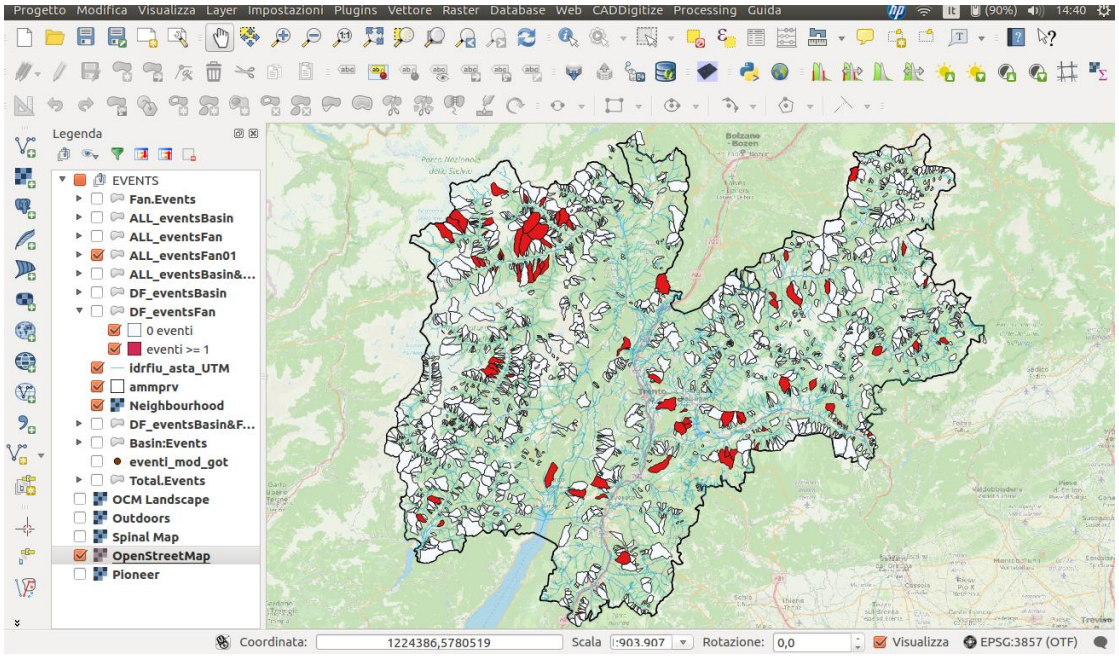


Figure 5.2: QGIS's interface, version 2.8.6-Wien

Its strength consist mainly to be a *Free and Open Source Software (FOSS)*. In this case "free" means that the users can freely run the software for any purposes (it is not necessary a *freeware*, which may be used without payment). Instead "open source" means that its *source code* is available with a license and everyone has the right to study, modify and distribute the software. This free software is released under the *GNU General Public License (GNU GPL)* regarding which more precise informations can be found on the website `www.gnu.org`.

It is a cross-platform software which runs on different operating systems like *Linux*, *Unix*, *Mac OSX*, *Windows* and *Android*.

QGIS is used to view, edit and analyse many geospatial data in various formats like vector (either point, line, or polygon features), raster (multiple formats are supported), and database. The user can also compose and export graphical maps and georeferenced images. The software's interface is divided into 5 areas as shown in Figure 5.2: a Menu bar, in the upper part, with standard hierarchical menu; a Toolbar gives quick access to some functions; a Map Legend where all the layers in the project are listed; a Map View where the maps are displayed; a Status bar, in the bottom part, with mouse position and scale of the map.

The functionality of *QGIS* is continuously maintained by a large group of volunteer developers who regularly release updates and bug fixes. They also extend the operability writing *Plugins* in the language *Python* or *C++*.

5.2 Application of the method

The available database is studied to obtain valuable information on the propensity of a given basin to generate intense torrential processes such as debris flows and debris floods. For this purpose, a principal component analysis followed by a logistic regression model is used. A model must not only be created but also tested. To accomplish these two tasks, reliable data must be filtered and the resulting database divided into separate subsets. A logistic regression model is trained and tested and then applied to the entire database.
A few of the common steps in model building are:

- preprocessing the predictor data;
- estimating model parameters;
- selecting predictors for the model;
- evaluating model performance;
- fine tuning class prediction rules.

The steps of the model formation and use are described in the following pages.

5.2.1 R packages and raw data import

The analysis is performed mainly using an R environment (as already described the software RStudio is used). As first step, the required libraries for the R packages have to be installed and loaded with the functions *install.packages()* and *library()*. In addition to the default functions already present in the R environment other tools are needed. For this reason the libraries *proj4*, *rgdal*, *corrplot*, and *caret* are loaded.
After this operation the raw data are imported into the *R* environment from Shapefiles using the packages *rgdal* and *proj4*. Spatial databases are normaly represented by a Shapefile which is a vector format for Geographical Information Systems. It describes the geometry of every element and the associated attributes. Two Shapefiles are used:

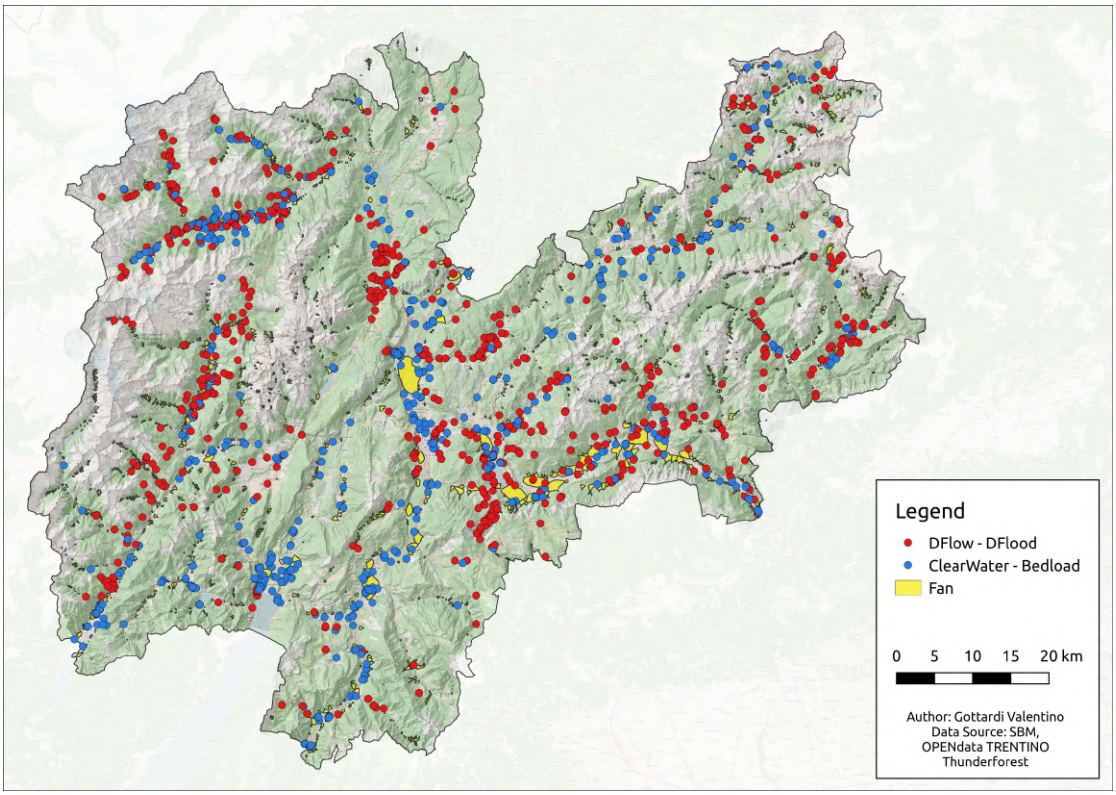
- **watershedMOD** represents drainage basin features (see Table 4.2);
- **DF_FAN** represents fan features (see Table 4.5).

Using the function *readOGR()* are imported the Shapefiles containing the geometry of catchment and fan. The two files are managed by R as "Large SpatialPolygonsDataFrame". This file class contains a *data* table, *polygons* geometry and *proj4string* reference coordinate. The projections of these spatial dataframe are set with the function *spTransform()*. The selected coordinate system is the ETRS89 / UTM zone 32N, (EPSG:25832). It is based on the *European Terrestrial Reference System 1989* datum and the *GRS 1980* ellipsoid.

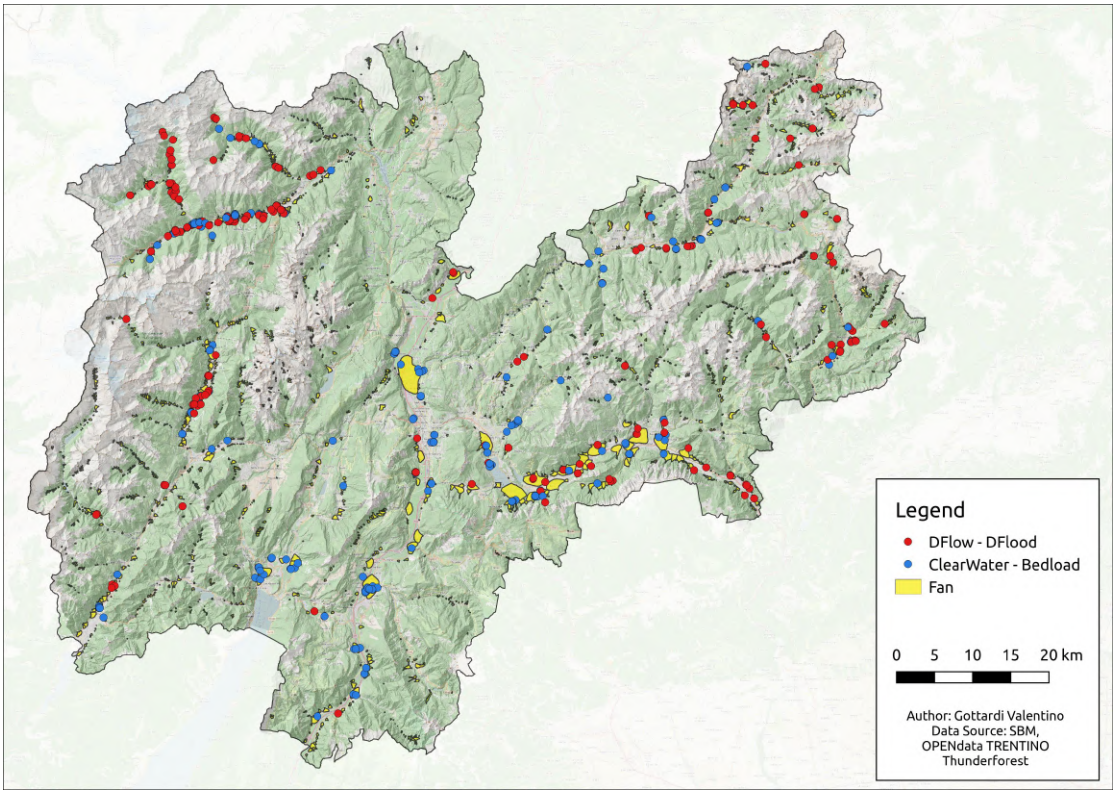
Data processing in GIS environment

The first stage of the analysis consists in filtering event data and it was performed in GIS environment. As exposed in chapter 4, the event cadaster presents various arguments.
In order to use only reliable informations, events with certain process type classification and location are filtered (using the variable *att_tipoev* and *att_locali*), obtaining a database of 2338 events (out of the 3522 events of the entire database).
This prudence have to be taken because, as stated by Marchi and Cavalli (2007), it should be remembered that at the time of the 1966 flood and in the years immediately afterwards, most debris flows were not recognised and analysed as such in Italy. The incorrect or complete omission of the classification of processes as debris flow in the Italian part of the Alps, while the term *Mure* already existed in Austrian literature, was pointed out by Castiglioni (1971).
Also the location of damages is sometimes unclear. Using only events with certain location are excluded general signals as widespread torrential activity on the entire water network.
We are interested in torrent related processes and for this reason we will not consider in the following analysis event observations connected with landslides, avalanches or river floods.

Given these premises, the events related to fluvial behaviour or low intense processes (516 clear water discharge events and 199 bedload transport events) are 715, whilst 824 intense processes (813 debris flow events and 11 debris flood events) are connected to torrential behaviour of the catchments. These 1539 fluvial selective transport and torrential mass transport processes with certain type identification and location are distributed along the main valleys of Trentino as shown in Figure 5.3a.
These events are not only localised on the basin fan. Since we want to predict the probability of debris flows reaching the fan area, we have to make a further selection of events, based on their location. Events that are reported in the basin but described as damaging infrastructure on the fan are relocated manually. Using the tool *Intersection* of the *Geoprocessing tools* in *QGIS*, the number of fluvial and torrential events on every single fan are counted. The first new column named *nDebris* counts the events related to debris flows and debris floods. The second new column named *nFluvial*, instead, counts low intense processes like bedload transport and clear water flood. Considering only observations within the perimeter of the fans (reported on the entire database which counts 2586 fans) they sum up to 406 events. 204 events of fluvial processes are detected (132 clear water discharge events and 72 bedload transport events). The number of torrential processes is 202, 198 of them are represented by debris flows and 4 by debris floods. The spatial distribution of these 406 events is depicted in Figure 5.3b. Blue dots represent fluvial processes (clear water discharge or bedload transport) instead red dots show torrential processes (debris flow or debris flood).



(a)



(b)

Figure 5.3: Spatial distribution of filtered fluvial selective transport (blue) and torrential mass transport processes (red): (a) events on the entire territory; (b) events only in fan area.

Table 5.1: Number of fans with different reported events. Rows: fluvial processes; columns: torrential processes

	0	**1**	**2**	**3**	**4**	**5**	**6**
0	1538	75	19	7	1	0	1
1	24	8	1	1	1	0	0
2	7	4	3	0	1	0	0
3	1	0	0	0	0	0	0
4	3	0	0	0	0	1	0
5	1	0	0	0	0	0	0

5.2.2 Data preprocessing

These raw data have to be prepared for the following steps. The statistical elaboration needs only numeric data. For this reason only the dataframes are extracted from the "Large SpatialPolygonsDataFrame". This means that the morphology of every single catchment and fan is considered only as numeric feature but not as geometrical information. To easily manage the entire set of information, the two dataframes are merged into an unique database. The *merge()* function of the *dplyr* package permits to define, for both database, a common field to drive the operation. Thanks to the univocal ID code the connection between numerical feature and geometry remains possible. After these passages a single database, called *dBase*, is ready.
Automatic generated or other unnecessary features, such as layer, date of polygon creation, information source, and other, are deleted. Other features are set to numeric vector in order to make the further calculation possible.
The variables of catchment and fan area and volume *barea*, *carea* and *cvol*, as well as variables *blungasta* and *bffactor* need to be expressed in their logarithm form to base 10 to better suit the statistic procedure.

Considering that the aim of the model is to predict the probability of a watershed to generate torrential events which traveled down to the fan, is essential to select basins with at least one reported event. The selection of active torrents is performed using the function *which()*. The result is the dataframe *dBaseActive*. 159 out of the 1697 basins have produced at least 1 event in their fan area with no distinction between fluvial or torrential processes.

A graphical display of events distribution can be created using the function *table()*. It is applied to the *dBase* dataframe. Rows of the table plot the number of the reported fluvial processes (bedload transport and clear water flood). The number of torrential processes (debris flow and debris flood) are instead listed by the columns (see Table 5.1). A total of 1538 fans didn't experienced any process during the considered time window. 75 fans reported one debris flow event but none fluvial processes, 19 fans experiences two debris flow events without any fluvial process, 24 fans counts one fluvial process without any torrential process, and so on. More in general 56 fans count at least one fluvial process and 123 fans at least one torrential process.

The feature *DF1* is added to the *dBaseActive* dataframe. This new feature is needed for the basin selection in the following passages. The feature gets the following values depending on the reported events:

- **DF1 = 1**, at least one debris flow or debris flood but none non-debris flow event (fluvial processes);
- **DF1 = 0**, at least one bedload transport event or clear water flood but none torrential processes;
- **DF1 = NA**, otherwise.

Over the total 159 catchments of the *dBaseActive* dataframe, 103 basins experienced only torrential processes and for them the feature *DF1* gets value 1. For 36 fans are reported only fluvial processes and the feature *DF1* gets value 0. To the remaining 20 basins with mixed behaviour, the value NA is assigned.
The following strings shown the used code.

```
if(dBaseActive$nDebris >= 1 & dBaseActive$nFluvial == 0){dBaseActive\$DF1 <- 1}
else if(dBaseActive$nDebris == 0 & dBaseActive$nFluvial >= 1){dBaseActive\$DF1 <- 0}
else{dBaseActive\$DF1 <- NA}
```

In the *dBaseActive* dataframe, one of the catchments shows NA values for the features *blungasta*, *bslo_asta* and *bffactor*. Considering the low number of available catchment appears important to get informations also about this basin. Therefore, the imputation of those values is performed by selection of a basin with the closest euclidian distance regarding all other features. Only numeric features are used to calculate the euclidean distance after scaling of the values. From the distance matrix the basin which has the lowest distance to the considered basin is selected. The missing features are copied from the nearest basin and attributed as shown in the script strings below.

```
# choose only numeric features
colidx <- seq(3, (ncol(dBaseActive)-4), 1)
colidx <- colidx[colidx !=15 & colidx != 24 & colidx != 23 & colidx != 22]

# reduce data set and standardize values
dBaseDistance <- dBaseActive[, colidx]
dBaseDistance <- scale(dBaseDistance)

# get distance matrix and that basin which has the lowest distance to the impute basin
rowID <- rownames(dBaseActive[is.na(dBaseActive\$bslo_asta), ]) # ID of the basin to impute
dmat <- as.matrix(dist(dBaseDistance, upper = TRUE)) # distance between basins
bidx <- which(rownames(dmat) == rowID) # get distance matrix row of the impute basin
nbidx <- names(sort(dmat[bidx, ])[2]) # get row ID of nearest basin

# get nearest basin data and impute
idxnb <- which(rownames(dBaseActive) == nbidx) # nearest basin
idxib <- which(rownames(dBaseActive) == rowID) # impute basin
dBaseActive[idxib, c(22:24)] <- dBaseActive[idxnb, c(22:24)] # impute
```

5.2.3 Feature selection

Not all the variables of the dataframe deliver useful information. Therefore, the reduction of the dataset to the most meaningful features is necessary. The idea is to search for variables which

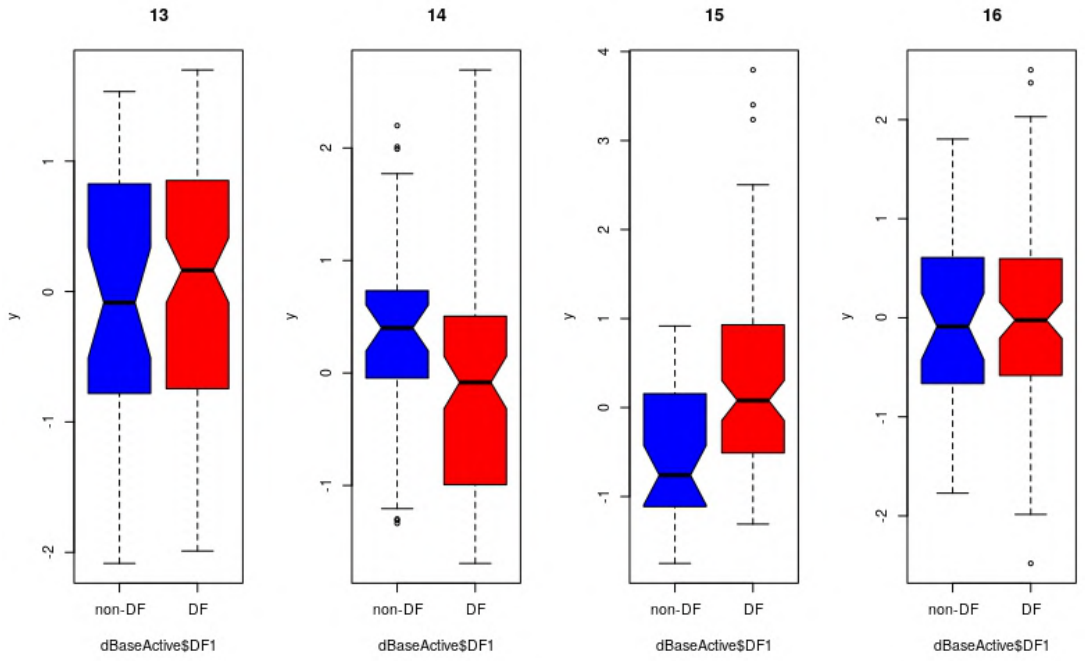


Figure 5.4: Feature selection using boxplot comparison: blue boxplot for fluvial processes class, red boxplot for torrential processes class.

show a certain difference between the class of basins prone to generate torrential processes from the class of basins prone to fluvial events. This class difference within the variables can be visually displayed by the boxplot comparison of the two classes. Only features characterized by a certain seperation ability in the one-dimensional case (difference between class "debris flow" and class "no debris flow") are mantained for the creation of the statistic model.

Variables are first centered and scaled in order to fit well into a boxplot for the separation ability check. The same variable is plotted twice on a single panel as shown in Figure 5.4. A blue boxplot represents the variable distribution for the basins classified as prone to fluvial processes (*non-DF class*). Torrential processes generating basins are represented by the red boxplot (*DF class*). As example, in Figure 5.4 four panels are drawn, corresponding to the variables named *big* (13), *bquantmm* (14), *cslo_med* (15) and *cvol* (16).

The pairs of boxplots are visually inspected to assess if the feature is a good predictor. The separation ability of a certain variable is given if the median values of the two classes are enough different. In other words if the median of one of the classes (black line at the center of the boxplot) lays out of the notch of the other boxplot (95% confidence interval of the median). As an example, the first and the fourth variables of Figure 5.4 are considered with poor separation ability and therefore not used for the model creation. In total, after the visual inspection, the following 9 variables are excluded: *bslo_max*, *bslo_med*, *bgravelius*, *bicompatt*, *bintipsom*, *big*, *cvol*, *bffactor* and *big1045mea*.

With the help of a correlation matrix, the reduced dataset is used to check for co-correlation

Table 5.2: Selected features for model building (*dBaseMod*)

Argument	Unit	Description
bhmed	$m\ a.s.l.$	Basin's average elevation
bslo_min	°	Basin's lowest declivity
bmelton	-	Basin's Melton index
bquantmm	mm/h	100 years precipitation quantile for 1h duration
cslo_med	°	Average fan declivity
carea	m^2	Fan area
bdensdren	km/km^2	Basin's drainage density
blungasta	m	Torrent's length
bslo_asta	°	Average torrent declivity

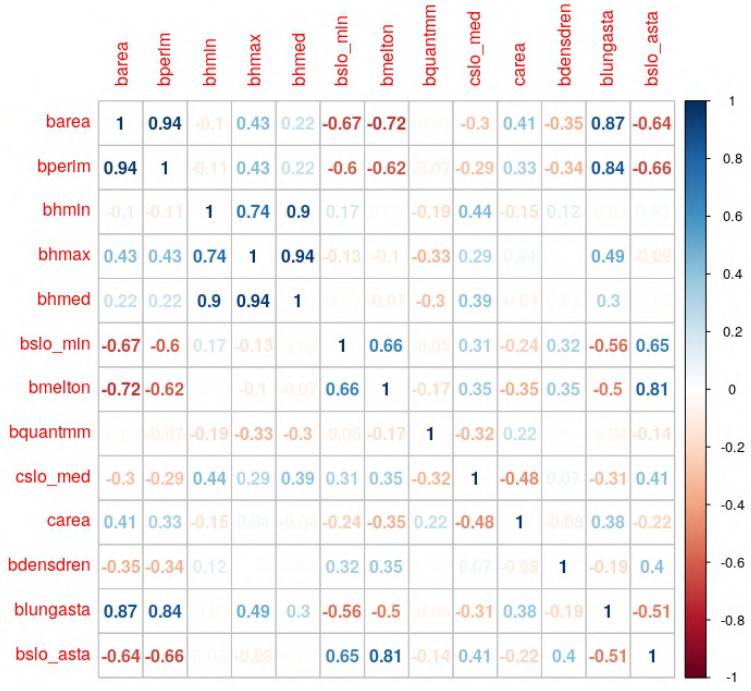
between the remaining 13 variables. Features with high co-correlation are then excluded. A correlation matrix is used to summarize correlation in data. It has the form of a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables.

The typical form of a correlation matrix presents the same variables shown in rows and columns. Each variable correlates perfectly with itself and this is shown by the value 1 on the main diagonal. This matrix is symmetrical, with the same correlation shown above and below the main diagonal.

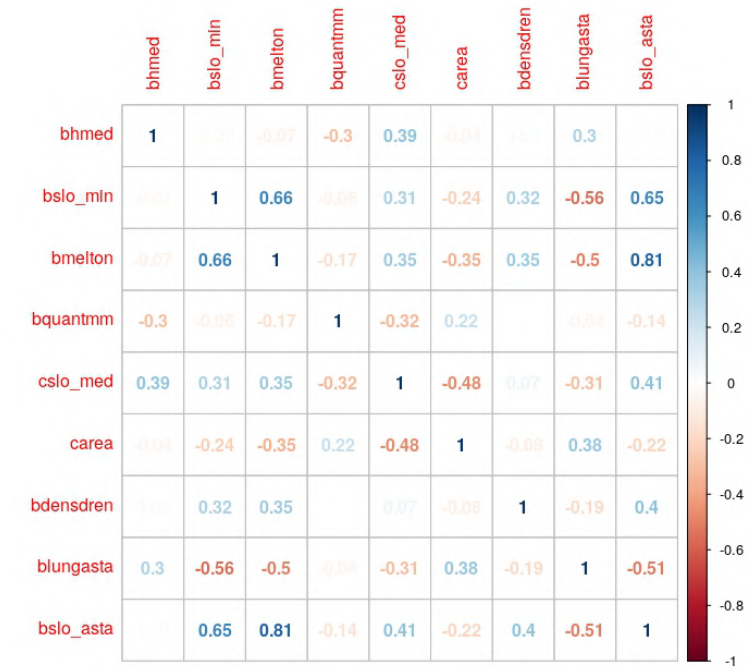
With the function *corrplot()* the correlation analysis between the 13 features characterized by good separation ability are calculated and plotted. The entire correlation matrix is shown in Figure 5.5a. High correlation values are those with values near to 1 or -1. The graph highlights positive correlation with the blue colored and negative correlation with red colored numbers. The high correlation between basin area and perimeter (value equal to 0.94) appears quite clear and logic. For this reason the variables *barea* and *bperim* are excluded. Basin area and perimeter are also correlated with stream length. High correlation exists also between the medium basin elevation and minimum and maximum elevations (correlation values equal to 0.90 and 0.94 respectively). Therefore the variable *bhmin* and *bhmax* are excluded.
Following script lines give an example of the plotting operations.

```
#plot boxplots
boxplot(y~dBaseActive\$DF1,
       main = paste(VariableName),
       notch = TRUE,
       col = c("blue", "red"),
       names = c("non-DF", "DF"))
# plot correlation matrix
corrplot(cor(dBaseSel[, c(7:9, 11:16)]), method = "number")
```

Figure 5.5b represents the correlation matrix of the 9 remaining variables listed in Table 5.2. These selected features show good separation ability in the one-dimensional case and are not strongly correlated. They are extracted from the dataframe forming the reduced *dBaseSel* database.



(a)



(b)

Figure 5.5: Correlation matrix of variables: (a) before and (b) after exclusion of high correlated variables

5.2.4 Principal component analysis

After that catchments and features have been selected begin the model building operations. In this case the model which will be fitted to the data have two different and consecutive approaches:

- principal components as inputs to overcome the problem of high collinearity between the predictors;
- logistic regression to separate between torrential processes dominated and fluvial processes dominated catchments.

As mentioned before, the Principal component analysis allows to deal with variables which are highly correlated. Considering the correlation between some of the available variables and the limitation deriving from the exclusion of features seems to be more indicated to apply the principal component analysis. This method delivers a simplified view of the entire dataset using a less numerous set of features, the principal components (PCs), which are linearly uncorrelated to each other. All the analysis is performed ensuring the minor possible loss of data variance.

To train and test the model only the active basins are used. As already explained, during the data preprocessing operations the variable *DF1* was attached to the database. Only catchments with well defined behaviour deliver useful information. In this case only the ones with *DF1* equal to 0 (only fluvial behaviour) or to 1 (only torrential behaviour) are selected. From the 159 active basins, the 20 which show a mixed behaviour are excluded. Only numerical features are used for the computation of the PCs.

The Principal component analysis is conducted applying the *prcomp()* function as shown below. This is present in the default package *stats* and its minimal call needs the indication of a data matrix from which are computed the PCs. Other arguments can be expressed like automatic centering or scaling of the input database (see script lines below). The result is an object of class *prcomp*.

```
1  # Conducting the Principal Component Analysis
2  # Estimate the PCs
3  PCAmodel <- prcomp(dBaseMod, center = TRUE, scale = TRUE)
4  # Predict the PCs
5  pcaData <- as.data.frame(predict(PCAmodel))
```

The application of the PCA on the 139 catchments with well defined behaviour results in the model named *PCAmodel*. As an example, Table 5.3 reports the first 10 calculated PCs. This table can be called using the function *summary()* and giving the name of the principal component analysis as data argument.

The PCA model has to be applied to the reduced database to predict the PCs. This operation delivers the dataframe called *pcaData* and is performed using the *predict()* function which is a generic function for predictions from the results of various model fitting functions.
The *pcaData* dataframe is a table containing all the observations (basins) with the relative 22 PCs replacing the original 23 features. To the *pcaData* dataframe is attached the behaviour group vector *DF1* obtaining the new database *dBaseModPCA*.

Table 5.3: First 10 PCs of the PCA (Active basins and whole predictors set)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	2.42	1.86	1.75	1.50	1.15	1.06	0.99	0.89	0.81	0.67
Proportion of Variance	0.28	0.17	0.15	0.11	0.06	0.05	0.05	0.04	0.03	0.02
Cumulative Proportion	0.28	0.44	0.59	0.70	0.76	0.82	0.86	0.90	0.93	0.95

5.2.5 Fitting the model

The 139 selected catchment-fan systems of the *dBaseModPCA* dataframe are used to train and test the model. A logistic regression will be performed using the PCs and applying bootstraping to asses its performance. The performance assessment is done based on the Youden's J statistic. Some explanations and definitions are now reported in order to clarify the fitting procedure which will be than described.

As already mentioned the logistic regression model is fitted on the active basins characterized by well defined behaviour. It is very important to decide which samples will be used to evaluate the performance of the model. Ideally, the model should be evaluated on samples that have not been used to build or fine-tune the model, in order to provide an unamplified sense of the model's effectiveness (Kuhn and Johnson, 2013). To assess how well the model works, it is important to have two different sets of samples: the first is used to create and train the model and the second is used to assess its ability to predict outcomes. They are called the *train subset*t and the *test subset* respectively.
Given the low number of basins to be submitted to the train subset and the test database, some approaches, called resampling techniques, are available to overcome this problem. The differences in the techniques usually lie in the various methods used to choose the subsamples. In each case a subset of samples is used to fit a model and the remaining samples are used to estimate the effectiveness of the model. The results of multiple repetition of this process are aggregated and summarised (Kuhn and Johnson, 2013).
In this case the *bootstrapping* method is used. It relies on random sampling *with replacement*. The term replacement means that if a data point is selected, it is still available for further selection. The obtained subset, called *bootstrap sample* is characterized by having the same size of the original data set. This necessary leads to the multiple representation of certain data points in this subset. Other data points are not selected at all and are referred to as the *out-of-bag samples*. The bootstrapping is performed multiple times, generally 1000 or 10000 times. The result of this iteration is a bootstrap sample used to built the model. The out-of-bag sample is predicted and used to estimate model performance.

Metrics as model overall accuracy rate, Kappa statistic, sensitivity, specificity and other are summarized by calling the function *confusionMatrix()*. The *confusion matrix* itself, is a widely used method to describe performance of a classification model. The form of the matrix is a cross-tabulation of the observed and predicted classes for the data.
The other performance statistics are listed in the console panel of *RStudio*. The overall accuracy rate reflects the agreement between the observed and predicted classes. For two classes other

statistics may be relevant when one class is interpreted as the event of interest (like for example debris flow event).

The *overall accuracy rate* is the simplest metric. This reflects the agreement between observed and predicted classes (Kuhn and Johnson, 2013). Accuracy rate does not make distinction between the type of errors and is therefore not really suitable for models with weighted errors. The overall accuracy have to be compared with the *no-information rate* which is the accuracy rate that can be achieved without a model. Alternatively it can be defined as the percentage of the largest class in training set. Models with accuracy rate greater than the no-information rate might be considered reasonable (Kuhn and Johnson, 2013).

The *Kappa statistic* or *Cohen's Kappa* was proposed by Cohen (1960). It takes into account the accuracy that would be generated simply by chance. It is calculated with the Equation (5.1). Where O is the observed accuracy and E is the accuracy based on the marginal totals of the confusion matrix.

$$Kappa = \frac{O - E}{1 - E} \tag{5.1}$$

The value of Kappa ranges between 1 and -1. When there is no agreement between observed and predicted classes the Kappa value is equal to 0, while value 1 means perfect concordance. Kappa values within $0,30$ to $0,50$ indicate reasonable agreement (Kuhn and Johnson, 2013).

The *sensitivity* of the model is the rate that the event of interest is predicted correctly for all samples having the event. Its meaning is more easily expressed by Equation (5.2).

$$Sensitivity = \frac{Samples\ reporting\ events\ and\ predicted\ as\ event\ prone}{Samples\ reporting\ events} \tag{5.2}$$

It can be seen also see as the *true-positive rate* because it is a representation of the accuracy between the population with events.

The *specificity* is defined as the rate that non-event samples are predicted as non-events. Equation (5.3) displays this definition. Like for the sensitivity, the specificity is sometimes called *true-negative rate* and is calculated as one minus the specificity.

$$Specifivity = \frac{Samples\ not\ reporting\ events\ and\ predicted\ as\ non-event\ prone}{Samples\ not\ reporting\ events} \tag{5.3}$$

In many cases it is interesting to have a single measure that reflects the false-positive and false-negative rates. 1950 W.J. Youden introduced the *Youden's J* statistic which summarize the magnitude of both types of errors. This parameter is calculated using Equation (5.4).

$$Youden's\ J = (Specificity + Sensitivity - 1) \tag{5.4}$$

The Youden's J index ranges from 0 through 1. It gets a zero value when the model is useless. A value of 1 indicates that there are no false positives or false negatives, therefore the test is perfect. Equal weight to false positive and false negative values are given by this index.

Resampling techniques

First step of the model fitting is the application of the resampling technique. As bootstrap repetition parameter the standard value of 1000 is used. The resampling is performed using the *sample()* function. This function permits to add the argument $replace = TRUE$ which turns the standard sampling into a random sampling with replacement, thus into the bootstrapping. From the 139 selected catchments, with already calculated PCs, the two following subsets are created:

- *datTrain*, dataframe of the bootstrap samples (according to the theoretical definition, it shows the same number of data as the one it is derived from, in this case 139 observations);
- *datTest*, dataframe of the out-of-bag samples.

The resampling code is reported below.

```
for(i in 1:m){
        idx <- sample(1:n, n, replace = TRUE)   # random selection of index (row n.)
        datTrain <- dBaseFit[idx, ]
        datTest <- dBaseFit[-idx, ]
```

Model training and test

When train and test subsets are available it is possible to train the model and evaluate its prediction. As already explained, the predictive model is based on a logistic regression which can be performed in *R* environment using the function *glm()*.
The term *glm* derives from the name *Generalized Linear Model* which collects a wide number of different models, from the basic linear regression to include models for non-normal distributions as the logistic regression (binomial data) and Poisson regression (count data). The used code is shown by the following script strings.

```
# Fit the model (6 most important PCs, 82% of Variance explained)
ModelTRAIN <- glm(DF1 ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6,
                  data = datTrain,
                  family = "binomial")
```

The model is named *ModelTRAIN* and is built using only the first 6 principal components in order to ensure the representation of as much as possible of the total variance as described in the section 1.7.3. Table 5.3 shows the result of the PCA and, more precisely, for every calculated PC the standard deviation, the proportion of variance explained by the PC on the total variance and the cumulative proportion of explained variance. It is easy to see that using the first 6 PCs, about 82% of the total variance can be explained. Representing more than 80% of the total variance using only 6 variables can be considered as a good compromise between information completeness and model simplification.
The amount of variation retained by each principal component is measured by the *eigenvalues*. As happes for the explained variation, the eigenvalues are large for the first PCs and small for the subsequent PCs. To determine the number of principal components to be retained after PCA is possible to consider the explained variance (eigenvalues) or the cumulative amount of total

variance. Kaiser (1961) indicated to retain PCs with eigenvalues bigger than 1 (is the same to say explained variance equal to 10%). This because such PCs account for more variance than accounted by one of the original variables in standardized data. The other approach is to choose the number of considered PCs, which ensure to cover a certain percentage of total variance, for example 75%. In the proposed model, the second method is used with the target to consider at least the 80% of the total variance.
The function needs to specify the class vector (*DF1*) and the variables used to predict the class of every observation (from PC1 to PC6). In addition, the used dataframe has to be indicated. In this case the model is built on the *datTrain* train subset. To force the generalized linear model function to performe a logistic regression, the argument *family = "binomial"* have to be added to the string.

After the creation of the model based on the train subset, it is applied to the test subset. The function *predict()* is used for this operation as shown in the following script.

```
# Predict test set and estimate
testPred <- predict(ModelTRAIN, newdata = datTest, type = "response")
testPred <- ifelse(testPred > 0.5, 1, 0)
```

The *type = "response"* option tells *R* to output calculated probabilities in the range $0 - 1$. Test subset dataframe is supplied by the argument *newdata = datTest*.
Utilizing the *ifelse()* function it is possible to fix a cut off value, default set to 0.50, for the estimation of propension classes. Basins with predicted debris flow propensity higher than 0.50 are assigned to the class 1, otherwise class 0 is given.

The performance of the trained model is evaluated using the statistics sensitivity, specificity and Youden's J. The function *confusionMatrix()* delivers these metrics and needs the specification of the vectors to compare. The argument *data* links the vector of predicted classes. Observed classes are reported by the argument *reference*. Also the positive class value needs to be indicated as shown by the below reported script.

```
confusionMatrix(data = factor(testPred, levels = 0:1),
                reference = factor(datTest\$DF1, levels = 0:1),
                positive = "1")
```

ROC curve and alternative cutoff

The diagnostic ability of a binary classifier model, as the used logistic regression, can be visually illustrated and investigated using the *ROC curve*. The curve depicts the tradeoff between sensitivity and specificity varing the discrimination threshold (between class 0 and class 1). The best cutoff, alternative to the default threshold of $0,50$, can be found using different methods.

The ROC curve is created and stored in the *ModelROC* object using the *roc()* function (see script strings below). It needs the response vector (with the observed classes *DF1*) and the predictor vector (with the predicted probabilities for every basin of the test subset). Once the ROC curve is calculated it can be plotted for the visual inspection.
Also many other numerical indexes can be taken from the ROC curve, such as the *area under the*

curve, or *AUC*. It is a performance measurement for classification problem, shortly, the higher the AUC, the better the model is in predicting event as event and non-event as non-event.

```
# create the ROC
ModelROC <- roc(response = datTest$DF1, predictor = datTest$Prob)
# alternative cutoff - optimal threshold
cutoff <- coords(roc = ModelROC,
                 x = "best",
                 best.method = "youden",
                 transpose = TRUE)

# Predict using new cutoff
testPredNew <- ifelse(testProb > 0.5222838, 1, 0)
```

Using this curve, an appropriate balance between sensitivity and specificity can be determined (Kuhn and Johnson, 2013). This problem often occurs when the default threshold equal to $0,50$ is not appropriate for the data. A new cutoff value can be fitted with the help of the ROC curve. In this case the threshold which maximizes the Youden's J index is choosen (which measures the proportion of correctly predicted samples for both the event and non-event group).
The alternative cutoff value is calculated calling the *coords()* function. First of all, this function needs the indication of the ROC object. Than, the favorite method is selected. As already mentioned, the Youden's J index is choosen. The script needs the specification of the argument *best.method = "youden"* and *x = "best"* in order calculate the optimal cutoff which maximizes the Youden's J index of the model.
The new probability threshold, equal to $0,5222838$ is used to classify again the test subset forming the new variable *TestPredNew*. This new class variable is feeded to the *confusionMatrix()* function in order to get new matrix for the numerical performance evaluation.

5.2.6 Class prediction with literature methods

The performance of the trained model can be compared with other methods found in the literature. To get a simple overview of the methods, two different approaches are used. The first one uses only parameters concerning the morphology of the catchment. The second is based on the morphometry of both catchment and fan.
The selection of the methods used for comparison is done by trying to use only the features already available. In this way the results of both, the proposed model and the available models, are well comparable.

Prediction based on catchment parameters

Between the classification systems based only on the catchment's morphometric parameters, the works of D'Agostino (1996) and of De Scally and Owens (2004) are selected. These authors proposed a similar approach based on the Melton ruggedness number (MRN), which differs only by the fixed threshold (see Table 5.4).
Following previous authors as for example Melton (1965), Jackson et al. (1987) and Marchi et al. (1993), D'Agostino (1996), analysed 62 historical events occured on small and medium sized

Table 5.4: Transport process classification, based only on Melton ruggedness number (MRN)

Author	Geographic area	Water floods	Debris flows
D'Agostino (1996)	Italian Alps	$MRN < 0,50$	$MRN > 0,50$
De Scally and Owens (2004)	New Zealand Alps	$MRN < 0,75$	$MRN > 0,75$

basins in the oriental part of Trentino. He used the Melton ruggedness number to differentiate between water floods (such as water discharge and bedload transport) and debris flows (included debris floods). The proposed Melton index threshold is equal to $0,50$. Basins with lower Melton index are referred to produce only fluvial processes, instead high intense processes are expected for basins with higher index.
The same parameter is used also in the work of De Scally and Owens (2004). This analysis is performed considering catchments of the Alps range of New Zealand. The same procedure as already explained is used but the proposed threshold is fixed to $0,75$.

Prediction based on catchment and fan parameters

D'Agostino (1996) and of De Scally and Owens (2004) also proposed an approach based not only on the morphology of the watershed but also on the geometry of the fan. In addition to the Melton ruggedness number is also considered mean slope of the fan. Table 5.5 summarizes the fixed thresholds.
For watershed with morphologically well defined fan area, D'Agostino (1996) proposed the Equation (5.5) which sets the mean fan slope (Sf) against a multiplied value of MRN. The Melton index has to be multiplied for a number which ranges between 7 and 14. The same author used the value 14 to discriminate between bedload transports and debris flows. This is a quite conservative value in the optic of torrent management, considering that the basins predicted as debris flow prone have a really low probability of missclassification. This means that many basins are classify as bedload transport dominated. Very few catchments, characterized by really steep fans or high MRN are predicted in class 1. With lower values the likelihood to predict a bedload transport event instead of a debris flow will be smaller. A basin is classified as debris flow prone if the mean fan slope (value expressed in grad) is bigger than the so multiplied MRN.

$$Sf > 7 - 14 * MRN \tag{5.5}$$

Using Melton ruggedness number and mean fan slope, De Scally and Owens (2004) predicted the most likely event type of torrential catchments of the New Zealand Alps. Differently from D'Agostino (1996), which compares mean fan slope against the MRN, these authors stated that a basin is prone to generate debris flow if at the same time, both MRN and mean fan slope, exceed given thresholds. As expressed by Equation (5.6), debris flows and debris floods are expected in catchments with Melton ruggedness number bigger than $0,75$ and mean fan slope steeper than 7,5°. If these two requirements are not respected only bedload transport processes are expected.

$$MRN > 0,75 \; and \; Sf > 7,5 \tag{5.6}$$

Table 5.5: Transport process classification, based on MRN and fan mean slope (Sf)

Author	Water floods	Debris flows
D'Agostino (1996)	$Sf < 7 - 14 * MRN$	$Sf > 7 - 14 * MRN$
De Scally and Owens (2004)	$MRN < 0,75$ and or $Sf < 7,5°$	$MRN > 0,75$ & $Sf > 7,5°$

In *R*, the just explained approaches are applied to the selected basins (*dBaseMod* database), which are also use to train and test the proposed model. The following script lines shown this procedure.

```
# From D'Agostino (1996) - Melton Ruggedness Number
dBaseMod$Dago1 <- as.factor(ifelse(dBaseMod$bmelton > 0.5, 1, 0))
pcaData$Dago1 <- dBaseMod$Dago1

# From DeScally and Owens (2004) - Melton Ruggedness Number
dBaseMod$DeSc1 <- as.factor(ifelse(dBaseMod$bmelton > 0.75, 1, 0))
pcaData$DeSc1 <- dBaseMod$DeSc1

# From D'Agostino (1996) - Fan's medium slope against Melton Ruggedness Number
dBaseMod$Dago2 <- as.factor(ifelse(dBaseMod$cslo_med > (dBaseMod$bmelton * 14), 1, 0))
pcaData$Dago2 <- dBaseMod$Dago2

# From DeScally and Owens (2004) - Melton Ruggedness Number and Fan's mean slope
dBaseMod$DeSc2 <- as.factor(ifelse(dBaseMod$bmelton > 0.75 &
                                   dBaseMod$cslo_med > 7.5, 1, 0))
pcaData$DeSc2 <- dBaseMod$DeSc2
```

The models and the relative predictions, based on the work of D'Agostino (1996) are named *Dago1* and *Dago2*. The first is based only on the MRN and the second on mean fan slope and MRN. The same is valid for the work of DeScally and Owens (2004) but models are respectively named *DeSc1* and *DeSc2*.

5.2.7 Application of the model

After the model has been built and assessed on the train and test subsets, it is necessary to build a new version which has to be used on the entire database. The new model is named *ModelALL*. It considers the first six PCs (from the *PCAmodel*, which was already computed before training operations) and uses the same options as for the training model. A new logistic regression is created modelling the selected watersheds of the *dBaseModPCA* dataframe (composed by 139 active basins with well defined behaviour) specifying it through the *data* = *dBaseModPCA* argument of the *glm()* function.

```
# use 6 most important PCs, original selected watersheds
ModelALL <- glm(DF1 ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6,
          data = dBaseModPCA,
          family = "binomial")
```

Some observations of the entire catchment database *dBase* are affected by missing values. More in detail, 13 observations present NA values for the variables *blungasta*, *bslo_asta* and *bffac-*

tor. The NA values are attributed using the procedure already explained in section 5.2.2. The variables *cslo_med* and *cvol* shown NaN values for other 22 basins. Also for these observations, the values are attributed with the same approach, after converting the NaN values into NA (see script in the R-code annex).

From the entire database, containing 1697 catchments and fans, non numerical features are removed. With this operation is formed the *dBasePred* dataframe. The data are now ready for the PCs computation. Differently from the procedure used before the model training (explained in section 5.2.4), the PCA is conducted without using the *prcomp()* function but directly predicting the PCs calling the already calculated PCA for the selected basins. This expedient allows to use as reference the *PCAmodel*, maintaining all the setting parameters fitted on the selected and well known basins as shown in the script lines below.

```
# Modelling data set (only numerical features)
dBasePred <- dBase[, c(-1, -2, -15, -26, -27, -28)]
# Conducting the Principal Component Analysis (using "PCAmodel")
dBasePred <- as.data.frame(predict(PCAmodel, newdata = dBasePred))
```

Following step is the prediction of debris flow occurence likelihood down to the fan, for every single catchment. It is performed using the *predict()* function. In this case the *ModelALL* logistic regression, which is based only on selected basins, is used to predict the entire database. This is set with the *object* = *ModelALL* and the *newdata* = *dBasePred* options. The calculated probability ranges from 0 to 1 and is contained into the *dfProb* vector, which is added directly to the database of the *Large SpatialPolygonDataFrame* called *basins*. This allows to export as a Shapefile the result of the model. Also the vector *DFclass* is added to the *basins* dataframe. This new variable is created using the function *ifelse()* and using the alternative cutoff value set to $0,5222838$ as visible in the following strings.

```
# Predict with the model
dfProb <- predict(object = ModelALL,
            newdata = dBasePred,
            type = "response")
# Add prediction vector of debris flow probability to the database
basins@data\$DFprob <- dfProb
# Add class prediction vector of debris flow probability to the database
basins@data\$DFclass <- ifelse(basins@data\$DFprob > 0.5222838, 1, 0)
```

5.2.8 Save and export the results

The results of the model can be visualized in form of a table or, thanks to the informations of polygon geometry connected to the database, as a map. In *R* environment it is possible to draw a map using the *plot()* function. A wide number of plotting options are available to create maps which are nice looking and easy readable.
In many cases it is more interesting to work in a separated GIS environment such as *QGIS*. In effect, sometimes a manual check of single basins is required or the data is needed as input of other softwares. In these cases is necessary to export the shapefiles. A folder for the shapefiles

is created in the directory of the *R* project using the function *dir.create()*. In the package *rgdal* is present the function *writeOGR* which make possible to perform the export of the file. It is necessary to specify some options such as object to be exported, saving directory, name of the exported shapefile and file format as shown in the next script strings.

```
# folder creation, dataset to shapefile
dir.create("RShapefiles")
# export shapefile
writeOGR(obj = basins,
       dsn = "RShapefiles",
       layer = "ResultsModellingPCA",
       driver = "ESRI␣Shapefile")
```

Chapter 6

RESULTS

This chapter summarises the results obtained by applying principal component analysis and logistic regression to the available database. Both the spatial distribution and the tables are used to visualise the results.

6.1 Events and basins selection

Considering the reported events occured in fan area, active basins are selected as explained in sections 5.2.1 and 5.2.2.

A total of 269 events are considered to classify the basins into active and inactive. Figure 6.1 displays the distribution of these events on the provincial territory, dividing them into two classes: torrential processes (red points) and fluvial processes (blue points). The processes are reported on one of the 1697 fans connected to the considered catchments. They are also characterized by certain event type and location and can be a debris flow, a debris flood, a bedload transport process or a clear water discharge process.
Table 6.1 summarizes the subdivision between the different process types of the 269 events. Most of the events are represented by debris flows ($65,4\%$ of the total). With the addition of the few reported debris floods the torrential processes sum up to the $66,9\%$ of the selected events. About one third of the total, instead, is represented by fluvial processes with the clear water discharge slightly more reported as the bedload transport (52 events against 37).

Table 6.1: Process type subdivision of the selected events

Process type	**Number**	% **on total**
Debris flow	176	65,4
Debris flood	4	1,5
Torrential	**180**	**66,9**
Bedload transport	37	13,8
Clear water discharge	52	19,3
Fluvial	**89**	**33,1**

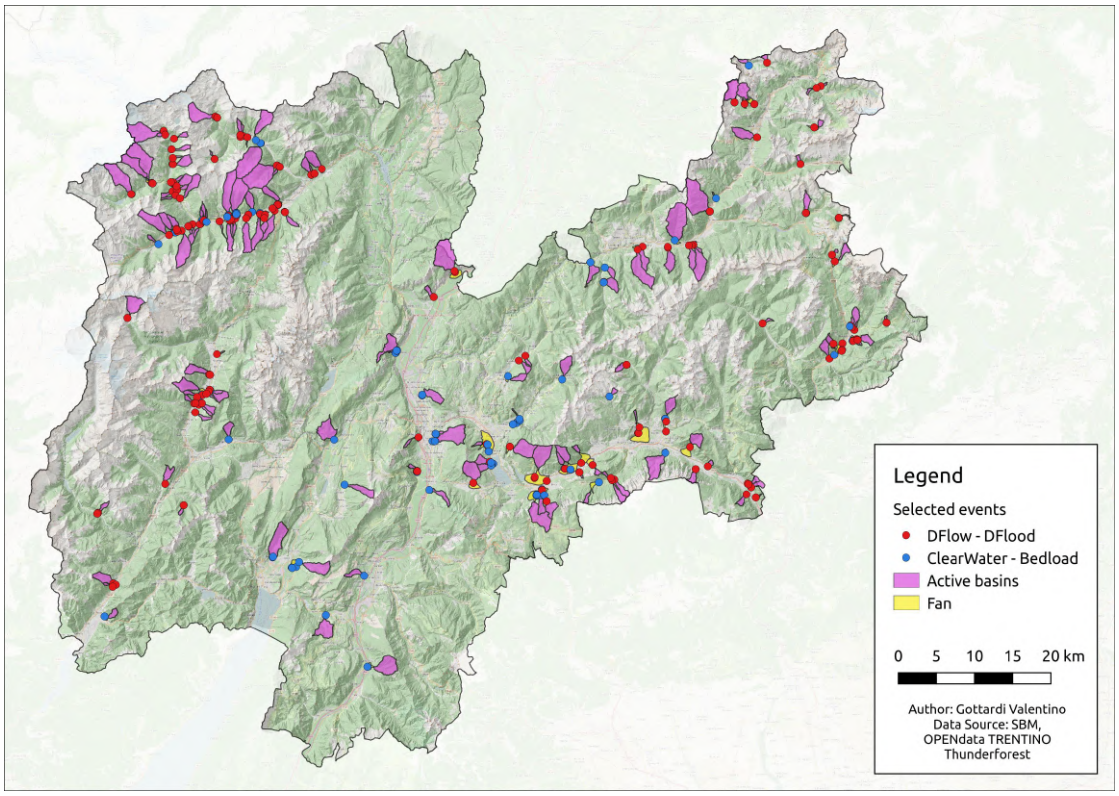


Figure 6.1: Spatial distribution of the selected events for model building. Red points show torrential processes, instead blue points display fluvial processes. Magenta and yellow polygons represent active catchments and fans respectively.

Table 6.2: Active basins selection

Process type	**Number**	**% on 159 basins**	**DF1 class**	**% on 139 basins**
Only torrential	103	64,8	1	74,1
Only fluvial	36	22,6	0	25,9
Mixed behaviour	20	12,6	NA	-
Total	159	100,0		100,0

Catchments with fans reporting at least one of the selected events are considered as active and collected into the *dBaseActive* database. 159 watershed-fan systems are selected and their spatial distribution is shown in Figure 6.2. The color of the polygons discriminates between dominant process type. Red polygons represent basins with dominant torrential behaviour, instead, blue color shows basins with fluvial behaviour. Mixed behaviour catchments are displayed by green polygons and fans by yellow ones.

The dominant behaviour of a catchment is given by the variable *DF1*, which has been attributed considering the reported process types as explained. The result is summarized in Table 6.2. The $12,6\%$ of the active basins have a mixed behaviour and are not really interesting for model building (20 basins). 103 catchments, corresponding to the $74,1\%$ of the remaining 139 catchments are represented by basins with a dominant torrential behaviour reporting only debris flows and debris floods. Only bedload transport events or water floods are reported on the other $25,9\%$ of the active basins (36 basins).

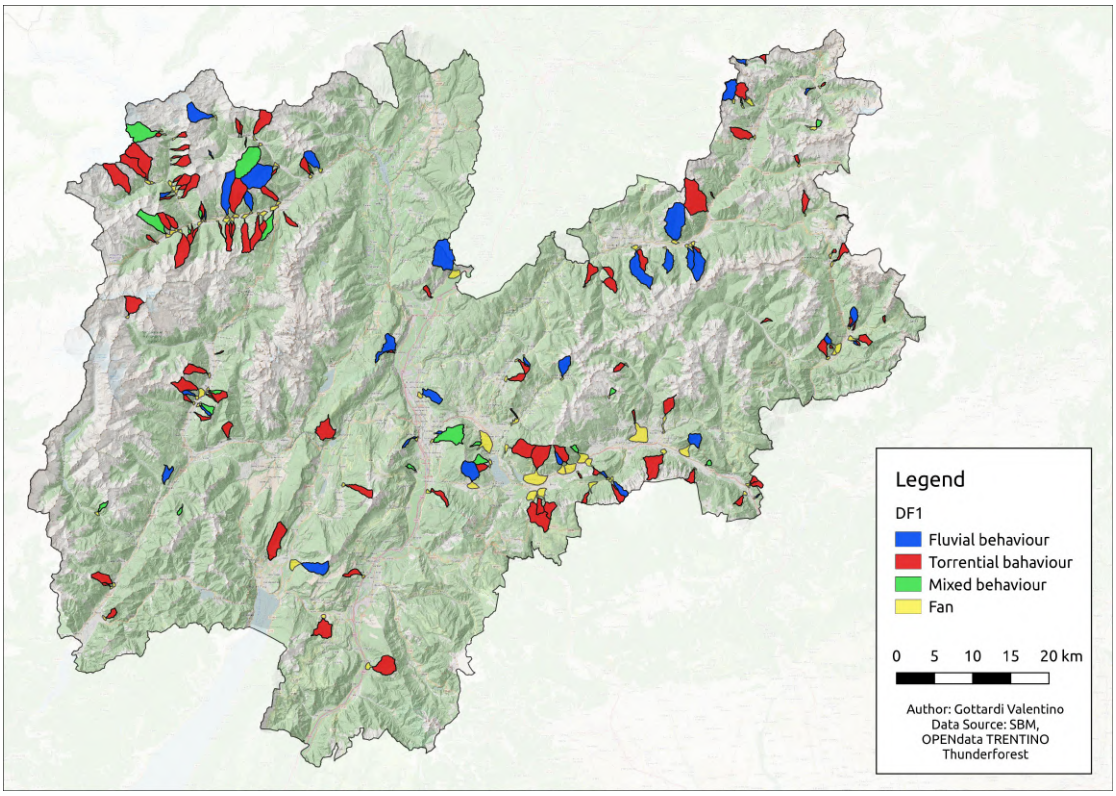


Figure 6.2: Spatial distribution of the active watershed-fan systems (*dBaseActive*). Red polygons show basins with dominant torrential behaviour; blue polygons are basins with fluvial behaviour; green polygons represent catchments with mixed behaviour; fans are displayed by yellow polygons.

Table 6.3: Reported events on the active fans

Process type	Number	Event number	% on total	Event/Fan
Only torrential	103	144	53,5	1,39
Only fluvial	36	58	21,6	1,61
Mixed behaviour	20	67	24,9	3,35
Total	159	269	100,0	1,69

Some of the active catchments reported more than one event on the relative fans. As visible in Table 6.3, 144 processes occured in basins with a torrential behaviour, which corresponds to $53,5\%$ of the total. Fluvial behaviour basins experienced 58 events ($21,6\%$) and the catchments with mixed behaviour count 67 events ($24,9\%$ of the total). Another interesting value is the calculated mean number of events occured on an active catchment-fan system. For torrential processes the mean is equal to $1,39$ debris flows or debris floods per fan. Basins with fluvial behaviour show a slightly more elevated mean value equal to $1,61$. Basins with mixed behaviour are generally more active and they produce about $3,35$ events per fan. A partial explanation of this high value is that some of the catchments without a dominant process type show densely infrastructured fans. This leads to a more precise and complete report of all events occured on the fan, also for the one with small magnitude and intensity, which is not the case for the majority of the basins.

6.2 Feature selection

The original idea was to search among the many features for some of the most significant and strongly separating features between the two groups of basins. Two steps constitute the first approach.
The first step is the visual analysis of the feature boxplots. For each individual variable, boxplots of the debris flow and non-debris flow basins are created and compared as shown in Figure 5.4. This first analysis leads to the exclusion of 9 variables.
The remaining 13 features are further inspected looking for co-correlation. This second step is performed applying a correlation matrix and deleting variables which show high correlation (values of calculated correlation near 1 or -1). This operation selects 9 variables characterized by low correlation (see Figure 5.5b) deleting 4 further features.

Deleting a considerable number of features leads to a consistent loss of information contained in the available databases. In addition to this, the proposed method needs to use the Principal component analysis, which allows to deal with highly correlated data. The choice which has been done is to process the entire database, without selecting variables for separation ability and co-correlation, with the PCA procedure. The principal components are constitutively linearly uncorrelated to each other and using an appropriate number of PCs will ensure an acceptable loss of data variance and a significant dataset simplification.

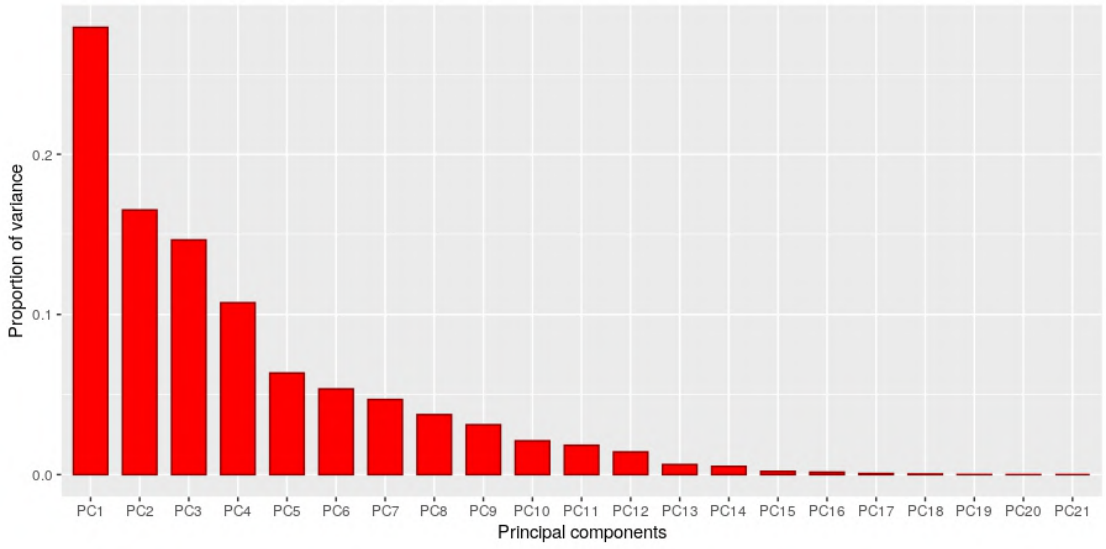
6.3 Model building and fitting

6.3.1 Principal component analysis

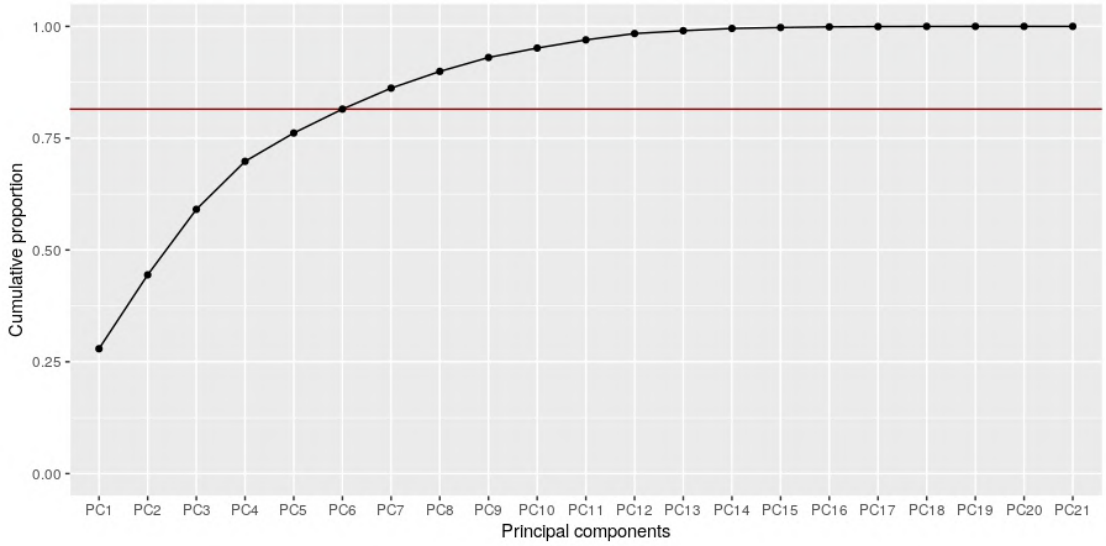
The Principal component analysis is applied to the database *dBaseMod* which collects the active watersheds. After centering and scaling, the variables are used to calculate the principal components. Table 6.4 reports for every calculated PC the relative standard deviation, the proportion of explained variance and the cumulative proportion. As already explained the first PC represents the biggest part of the total variance and the following PCs have decreasing importance. This table is formed feeding the principal component model to the *summary()* function.

Table 6.4: Calculated Principal components (*dBaseMod* database)

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Standard deviation	2.421	1.862	1.754	1.501	1.154	1.060	0.992
Proportion of Variance	0.279	0.165	0.147	0.107	0.063	0.053	0.047
Cumulative Proportion	0.279	0.444	0.591	0.698	0.762	0.815	0.862
	PC8	**PC9**	**PC10**	**PC11**	**PC12**	**PC13**	**PC14**
Standard deviation	0.887	0.808	0.665	0.620	0.545	0.364	0.328
Proportion of Variance	0.037	0.031	0.021	0.018	0.014	0.006	0.005
Cumulative Proportion	0.899	0.930	0.952	0.970	0.984	0.990	0.995
	PC15	**PC16**	**PC17**	**PC18**	**PC19**	**PC20**	**PC21**
Standard deviation	0.206	0.177	0.118	0.085	0.033	0.016	7.8e-16
Proportion of Variance	0.002	0.002	0.001	0.000	0.000	0.000	0.000
Cumulative Proportion	0.997	0.999	1.00	1.00	1.00	1.00	1.00



(a)



(b)

Figure 6.3: Principal component analysis on active basins (*dBaseMod* database): (a) PCA scree plot, (b) cumulative proportion of explained variance.

Figure 6.3 represents graphically the importance of the calculated components. The upper panel shows a scree plot. A scree plot displays how much variation each principal component captures from the data. The lower panel represents the cumulative proportion of explained variance with increasing PCs number. The black line shows the cumulated amount of variance while the red line indicates the cut-off point fixed choosing the first 6 PCs (the explained variance sums up to $81,5\%$ of the total variance).

The Principal component analysis model can be investigated calling the *rotation* argument ($PCAmodel\$rotation$). The result is a matrix of rotation values indicating the variable loadings. Loadings are interpreted as the coefficients of the linear combination of the original variables from which the principal components are constructed. From a numerical point of view, the loadings are

equal to the coordinates of the variables divided by the square root of the eigenvalue associated with the component. Trying to summarize it in the simplest way, the loading values relates original variables with the calculated PCs giving informations about how much every original variable counts for a single PC.

The calculated loadings for the first 6 PCs are listed in Table 6.5. Assuming that loadings are the key to understanding the underlying nature of a particular factor, is possible to make some statements for the first PCs.

For example, considering calculated loadings for PC1 is possible to see that some variables are associated with positive values and other ones with negative values. Variables *bmelton* and *bslo_asta* show the highest positive loadings, respectively equal to $0,3459$ and $0,3152$. Negative loadings are maximum for variables *barea* ($-0,3901$), *bperim* ($-0,3650$) and *blungasta* ($-0,3467$). These values give the impression of the importance of the cited variables for the formation of the first PC. In other words, PC1 is positively sensitive to the value of the *bmelton* and *bslo_asta* (high MRN and steep main channels deliver strong advices of the likelihood of debris flow formation and transport). On the other hand, big values of the variables *barea*, *bperim* and *blungasta* represent catchments with a general larger shape and total channel length which are dominated by bedload transport processes. The other variables are less meaningful for the formation of PC1 but can play a roll for the subsequent PCs. It is important to remember that PC2 (and the sequent PCs) have a decreasing variance explanation capacity. For this reason, it is possible that also significant high loadings end up with low explanation capacity, or at least lower than the one of PC1.

Table 6.5: PCA loadings, first 6 PCs

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
barea	-0.3901	0.0485	0.0252	0.0475	-0.1332	0.0169
bperim	-0.3650	0.0233	0.0635	0.1533	-0.1010	0.0103
bhmin	0.0413	0.3606	0.2984	-0.1298	0.2415	0.1236
bhmax	-0.1272	0.4330	0.2456	0.0529	0.0951	0.0555
bhmed	-0.0592	0.4393	0.2847	-0.0384	0.0958	0.0166
bslo_min	0.3006	0.1190	-0.1261	-0.0880	0.1599	0.0040
bslo_max	-0.1485	0.2496	-0.2967	0.1526	-0.2092	0.1184
bslo_med	0.1432	0.3335	-0.3500	0.0494	-0.0639	0.0773
bgravelius	0.1549	-0.1317	0.1770	0.5185	0.0880	-0.1577
bicompatt	0.1549	-0.1317	0.1770	0.5185	0.0880	-0.1577
bmelton	0.3459	0.1222	-0.0739	0.2511	0.0265	-0.1123
bintipsom	-0.0213	0.1549	0.0535	-0.0921	-0.4788	-0.5169
big	-0.0402	-0.1553	0.3507	-0.1144	0.2921	0.0432
bquantmm	-0.0665	-0.0992	-0.3070	-0.0447	-0.0178	0.2306
cslo_med	0.1855	0.2221	0.2282	-0.1574	-0.1941	-0.1555
cvol	-0.2008	0.1426	-0.2424	-0.0352	0.4531	-0.4367
carea	-0.2470	0.0422	-0.2825	0.0139	0.4304	-0.3476
bdensdren	0.1646	0.1575	-0.1560	0.0706	0.1981	0.4143
blungasta	-0.3467	0.1298	0.0020	0.2539	-0.1186	0.0776
bslo_asta	0.3152	0.2054	-0.1847	-0.0039	-0.0931	-0.2054
bffactor	0.0614	-0.1899	0.0344	-0.4469	0.0297	-0.1441

6.3.2 Logistic regression

The first 6 calculated PCs are used to feed the logistic regression which has to predict the dominant behaviour of each catchment. The logistic regression uses these independent variables to model a binary dependent variable (*DF1*). The calculated probability ranges between 0 and 1. The classification into active or inactive basins uses the default threshold equal to $0,50$.

To get details about the logistic regression, named *ModelTRAIN*, is used the *summary()* function. Figure 6.4a represents the console output of this operation. It reports the original call of the *glm()* function and a summary of the deviance residuals (which seems to be good since they are close to being centered on 0 and are roughly symmetrical excluding important outliers). An important part of the panel is occupied by the coefficients. The column *Estimate* lists, like for a linear model, y-axis intercept and multiplication value for each PC as well as relative P-values and statistical significance. Moreover, null deviance, residual deviance and Akaike Information Criterion can be used to compare different models and compute different matrix.

The performance of the logistic regression can be graphically expressed plotting the predicted probability and observations. Watersheds are sorted on the basis of increasing predicted probability forming a rank. Each point is plotted respecting the ranking and colored with reference to the observations (variable *DF1*). Figure 6.4b depicts a graph that shows the predicted probabilities that each catchment will generate torrential processes along with the reported dominant behaviour. The graph is constructed on the 139 basins of the train subset. In this case, most of the catchments with reported torrential behaviour (points colored in turquoise), are predicted to have a high probability to generate debris flows or debris floods (predicted probability higher than the $0,50$ threshold, shown by the red solid line). Moderate good separation ability is shown for the fluvial dominated watersheds (colored in salmon), which are predicted to have both low and high probability of having torrential processes (below and above the $0,50$ predicted probability threshold).

Considering the numbers, 94 of the 139 basins in the train subset are characterised by torrential behaviour. The logistic regression correctly predicts probabilities higher than the threshold for 84 catchments, only 10 are predicted with low probability. The remaining 45 catchments show only fluvial behaviour and their predicted probabilities are almost even distributed above and below the threshold value. 22 basins have probabilities bigger than $0,50$, while 23 basins have smaller predicted probabilities.

Also with some missclassification, the performance on the train subset is good, with erroneous classification equal to $30,3\%$ (10/33) in the low probability class (predicted probability below the $0,50$ threshold) and to $20,8\%$ (22/106) in the high probability class.

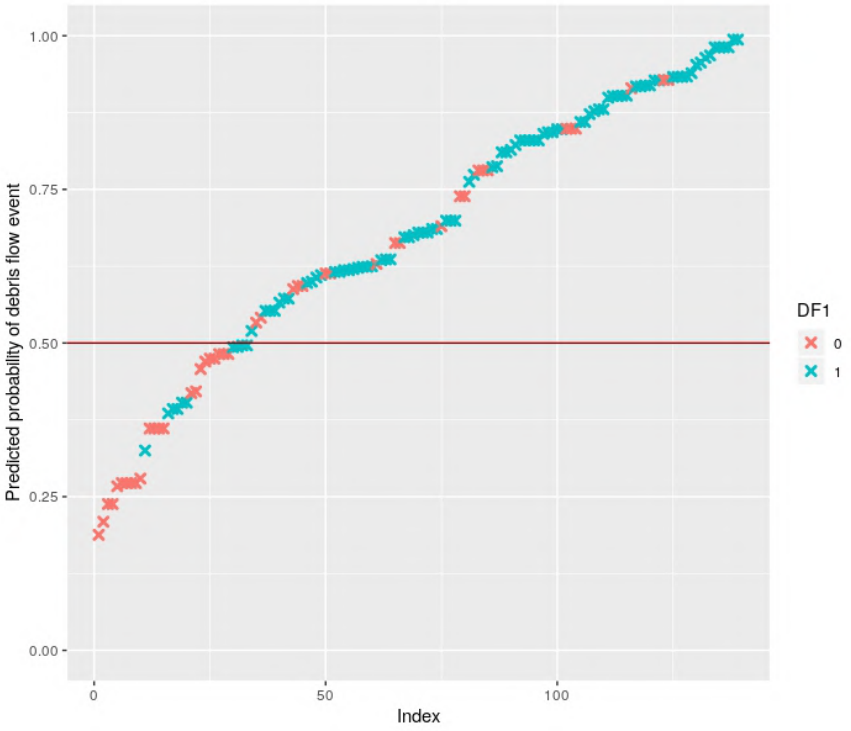
The same approach can be performed on the test subset (composed by 54 basins). In this case, the error rate is higher for the low probability class, and is equal to $45,5\%$ (5/11). Instead, for the high probability class the error rate goes down to the $11,7\%$ (5/43).

6.3.3 Receiver operating characteristic (ROC) curve

The *receiver operating characteristic curve*, generally known as *ROC curve*, is a graphical plot. It is widely used to illustrate the diagnostic ability of a binary classifier system as its discrimination

```
Console   Terminal   Jobs
~/Documenti/MASTERARBEIT/aa_NEWMODEL/GOT/
> summary(ModelTRAIN)

Call:
glm(formula = DF1 ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6, family = "binomial",
    data = datTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2937  -0.8779   0.4547   0.8886   1.4991

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.93394    0.22724   4.110 3.96e-05 ***
PC1          0.27628    0.09825   2.812  0.00492 **
PC2          0.20217    0.12918   1.565  0.11759
PC3          0.26486    0.11908   2.224  0.02613 *
PC4          0.31364    0.19544   1.605  0.10854
PC5          0.32700    0.20133   1.624  0.10432
PC6         -0.47234    0.23352  -2.023  0.04311 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 175.04  on 138  degrees of freedom
Residual deviance: 143.90  on 132  degrees of freedom
AIC: 157.9

Number of Fisher Scoring iterations: 5

>
```

(a)



(b)

Figure 6.4: Logistic regression (train subset): (a) summary *ModelTRAIN*, (b) sorted calculated probabilities (turquoise points basins with torrential behaviour, salmon points basins with fluvial behaviour, red solid line $0,50$ predicted probability threshold).

Table 6.6: Confusion matrix proposed model

Threshold 0,5000	**DF1 = 0**	**DF1 = 1**
Pred = 0	5	6
Pred = 1	5	38

Threshold 0,5223	**DF1 = 0**	**DF1 = 1**
Pred = 0	6	8
Pred = 1	4	36

Table 6.7: Statistics for the used cutoff values

Index	***Threshold 0,5000***	***Threshold 0,5223***
Accuracy rate	0,7963	0,7778
No-information rate	0,8148	0,8148
Kappa statistic	0,3501	0,3622
Sensitivity	0,8636	0,8182
Specificity	0,5000	0,6000
Youden's J	**0,3636**	**0,4182**

threshold is varied (Wikipedia). It is created plotting sensitivity against 1-specificity.

After the model have been trained on the train subset it is used to predict the elements of the test subset. The results are gained using a *confusion matrix*. This is a simple cross-tabulation of the observed and the predicted classes for the data (in this case the test subset). Diagonal cells denote cases where the classes are correctly predicted while the off-diagonals illustrate the number of errors for each possible case (Kuhn and Johnson, 2013).

The ROC curve of the model is depicted by the blue line in Figure 6.5. The ROC curve represents the sensitivity and specificity across a continuum of cutoffs. The AUC is also highlighted. The proposed model presents an AUC equal to $69,1\%$, with a confidence intervall between $46,5\%$ and $91,7\%$.
As already explained in section 5.2.5, a new cutoff value can be identified to define an appropriate balance between sensitivity and specificity. In this case the new cutoff is fixed to $0,5222838$.

Predictions of the two different threshold values are shown in form of confusion matrix in Table 6.6. It is easy to see that increasing the threshold value leads to a greater number of predicted basins in the 0 class (which pass from 11 to 14). In the mean time, the basins predicted as active decrease from 43 to 40. However, the general distribution in the matrix remains almost the same. Other statistic indicators are helpful in order to assess model performances. Table 6.7 lists some statistical indexes for both models. Quite interesting are the differences between sensitivity and specificity indexes. With the optimized cutoff the sensitivity slightly decrease (more active basins are predicted as inactive in comparison with the default threshold) and the specificity increase (for the opposite reason). The new cutoff value can be considered as better, looking at the Youden's J index. This last metric measures the overall missclassification and did not consider different weights for the various types of errors.

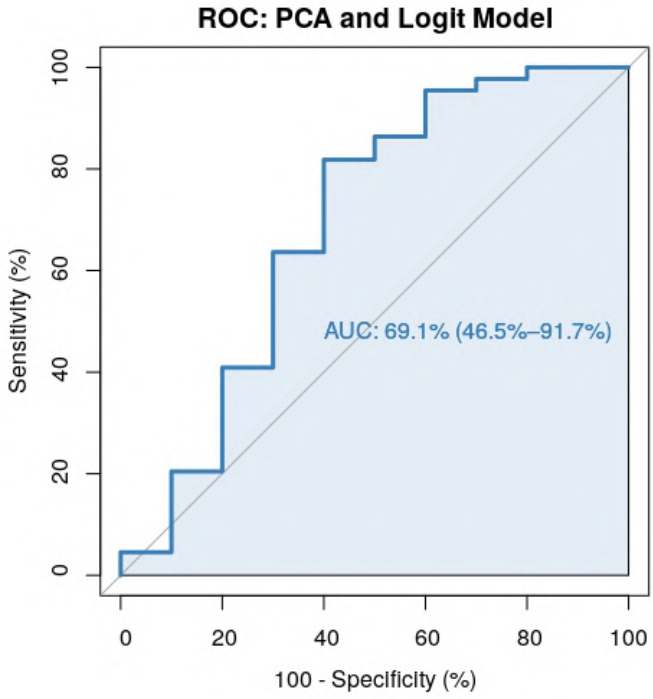


Figure 6.5: ROC curve of the proposed model (calculated on the *datTest* database).

6.3.4 Comparison with other methods

The prediction performances of the proposed model are tested against the observations of the test subset. For a better description of the discrimination ability between dominant fluvial or torrential behaviour, the models proposed by D'Agostino (1996) and DeScally and Owens (2004) as explained in Section 5.2.6 are used.

Figure 6.6 is composed by 6 panels. The active basins are represented by black outlines. Colored polygons are used only for catchments which are part of the test subset. Classification of the 54 watersheds based on observed events (variable *DF1*) is shown by panel 6.6a. Orange areas indicate basins with reported debris flows or debris floods, green with only bedload transport processes.
Panel 6.6b represents the results of the proposed model. Predicted classes are colored in blue, in the case of predicted fluvial behaviour (class = 0), or in red, in the case of predicted torrential behaviour (class = 1). The same colors are used to depict results of the other four used methods. Panel 6.6c and panel 6.6d report the classification produced considering only the Melton ruggedness number (MRN threshold equal to 0,50 and to 0,75 respectively). Classes predicted with the application of the Equations 5.5 and 5.6 which combine mean fan slope and Melton ruggedness number are shown in panel 6.6e and 6.6f.

Additionally, to the spatial distribution of the test subset, is interesting to make a comparison between predicted classes and observations. This operation is commonly performed in a clever way using the *confusion matrix*. The confusion matrix of the proposed method, together with the other four methods in literature are listed in Table 6.8. The 54 test basins are not evenly

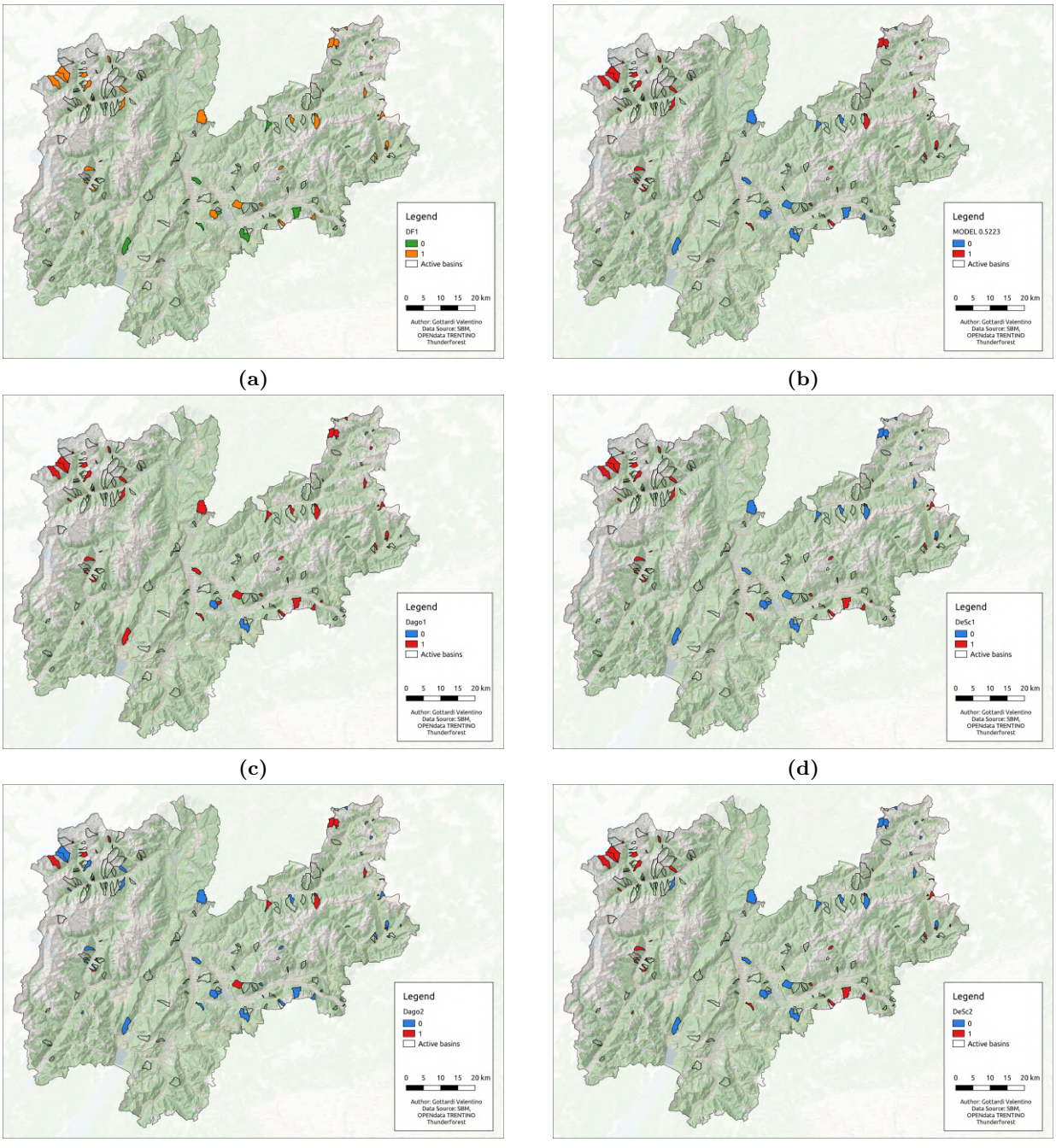


Figure 6.6: Class predictions of the used methods: (a) observation (variable *DF1*), (b) Model prediction, (c) model *Dago1*, (d) model *DeSc1*, (e) model *Dago2*, (f) model *DeSc2*. Models *Dago1* and *Dago2* from D'Agostino (1996); models *DeSc1* and *DeSc2* from DeScally and Owens (2004). Class 1 represents basins predicted as debris flow prone.

Table 6.8: Confusion matrix for the used models

MODEL	**DF1 = 0**	**DF1 = 1**
Pred = 0	6	8
Pred = 1	4	36

Dago1	**DF1 = 0**	**DF1 = 1**
Pred = 0	1	1
Pred = 1	9	43

DeSc1	**DF1 = 0**	**DF1 = 1**
Pred = 0	5	11
Pred = 1	5	33

Dago2	**DF1 = 0**	**DF1 = 1**
Pred = 0	9	32
Pred = 1	1	12

DeSc2	**DF1 = 0**	**DF1 = 1**
Pred = 0	5	13
Pred = 1	5	31

Table 6.9: Class prediction: number and percentage (test subset)

Class	***DF1***	***MODEL***	***Dago1***	***DeSc1***	***Dago2***	***DeSc2***
Class 0 [n]	10	14	2	16	41	18
Class 0 [%]	*18,5*	*25,9*	*3,7*	*29,6*	*75,9*	*33,3*
Class 1 [n]	44	40	52	38	13	36
Class 1 [%]	*81,5*	*74,1*	*96,3*	*70,4*	*24,1*	*66,7*

distributed between fluvial and torrential behaviour as visible from the column *DF1* of Table 6.9. Only 10 basins, corresponding to $18,5\%$, are characterized by dominant fluvial behaviour, while 44 have generated only debris flows. The ratio between class 0 and class 1 can only be approximated predicting with the proposed model, althought only 6 of the 10 class 0 predicted basins are corrently defined (see confusion matrix). The calculated class ratios of the other four methods, differ strongly from the observed events.

Some of the statistics obtained with the *confusionMatrix()* for the models are listed in Table 6.10. For the description of the reported indexes see Section 5.2.5. The considered base statistics are the overall accuracy rate, the no-information rate, the Kappa statistic, the sensitivity, the specificity and the Youden's J.

The no-information rate is the same for the five models because of the common train and test database. Its value is the prevalence of class 1 basins in the test subset. To complete the information, 94 out of the 139 train subset basins present the variable *DF1* equal to 1 (representing $67,6\%$ of the total).

The accuracy rate varies strongly between the models. It is high for the *Dago1* model and low

Table 6.10: Statistics for the used models

Index	***MODEL***	***Dago1***	***DeSc1***	***Dago2***	***DeSc2***
Accuracy rate	0,7778	0,8148	0,7037	0,3889	0,6667
No-information rate	0,8148	0,8148	0,8148	0,8148	0,8148
Kappa statistic	0,3622	0,1118	0,2030	0,0786	0,1562
Sensitivity	0,8182	0,9773	0,7500	0,2727	0,7045
Specificity	0,6000	0,1000	0,5000	0,9000	0,5000
Youden's J	**0,4182**	**0,0773**	**0,2500**	**0,1727**	**0,2045**

for the *Dago2* model (also because the high multiplication value choosen for its constitutive equation). Considering the proposed model it is important to highlight that the overall accuracy rate is slightly lower than the no-information rate. The strongly uneven class frequence is responsible for much of this effect. For this reason other metrics have to be used in order to assess model performance to take into account class distributions of training and test subsets.

Kappa statistic considers the accuracy that would be generated simply by chance (Kuhn and Johnson, 2013). Values near 0 indicate no agreement between observation and prediction. Low values are achived by the *Dago1*, *Dago2* and *DeSc2* methods. For this parameter, the proposed logistic regression gets $0,3622$, which is a positive rating because values between $0,30$ and $0,50$ are evaluated as "moderate to good agreement".

Simplifying, the sensitivity is the ability to detect the positive class ($DF1 = 1$), while the specificity is the ability to detect negative class ($DF1 = 0$). An extremely high sensitivity $(0,9773)$ is obtained by the model *Dago1*. Thanks to the low MRN threshold, most of the test basins are predicted as debris flow prone, leading to the almost complete detection of the basins with observed torrential processes. The model *Dago2*, instead, has a considerable specificity $(0,9000)$ due to the high multiplication value (set to 14) of the applied equation. On the other hand, the first shown a really low specificity $(0,1000)$ and the second a suboptimal sensitivity $(0,2727)$. The proposed model, instead, is characterized by a good sensitivity equal to $0,8182$ (confirming its good ability to predict as debris flow prone catchments which experienced torrential events) and an average specificity of $0,6000$ (with a certain missclassification of basins with fluvial behaviour into the class of torrential behaviour).

When the model did not use weights or penalties for the two different types of errors, it is possible to have a single measure that reflects the false-positive and the false-negative rates. Youden's J index summarizes the magnitude of both types of errors measuring the proportions of correctly predicted samples for both event (class = 1) and nonevent (class = 0) groups. Between the tested methods the highest Youden's J value is achived by the proposed model $(0,4182)$ followed by the models of DeScally and Owens (2004) and from the approach of D'Agostino (1996). This index condense the informations from the sensitivity and the specificity and therefore suffers from the same problems described above.

Table 6.11: Predicted probability statistics

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	NA's
0,1067	0,6665	0,8235	0,7720	0,9239	0,9951	5

6.4 Application to the entire database

6.4.1 Logistic regression and prediction

After the model have been built and trained, it is applied to the whole catchment and fan database as explained in section 5.2.7. First step is the creation of the *ModelALL* logistic regression using all the active and selected basins and the already calculated PCs. The results of this new model can be recalled, as already done for the training model, calling a function in *R*. Figure 6.7a represents the console output of this operation. The same evaluations concerning deviance residuals, coefficients, deviance and AIC, which were made for the model test, can be carried out in this situation. Important to mention is the good significance of the first two PC considering the P-value. This can be seen in the coefficients panel, highlighted by the two and three stars (P-value minor than $0,005$).

With the argument *type* = "*response*" of the function *predict()*, the probability to generate and transport a debris flow or a debris flood down to the fan of every single basin of the database is predicted. The distribution of the predicted probability is graphically displayed in Figure 6.7b. All the 1697 watersheds are ranked with increasing predicted probability and plotted in form of black rings. The red solid line represents the used cutoff value (fitted to $0,5222838$) to separate between catchments with predicted dominant fluvial behaviour and catchments with predicted dominant torrential behaviour.
Theoretically the debris flow likelihood ranges between 0 and 1. The statistic description of its distribution is listed in Table 6.11 and graphically displayed using a boxplot by Figure 6.8a. The predicted probability ranges between $0,1067$ and $0,9951$ with a mean value equal to $0,7720$.

As already introduced, the predicted probability to generate debris flows of a basin, can vary between 0 (low probability) to 1 (high probability). The used cutoff value is slightly higher than the default $0,50$ in order to maximize the Youden's J index of the model. The results, in terms of class numerosity is shown in Figure 6.8b. Only 5 out of the total 1697 basins, which represent $0,29\%$, have missing values in at least one of the necessary variables. For these few basins the probability can't be estimated and the prediction is not performed. Bedload transport and clear water discharge processes are considered dominant for 210 basins. These catchments with low predicted debris flow likelihood represent the $12,38\%$ of the entire database. A torrential behaviour, with debris flows and debris floods have to be expected for the remaining 1482 catchments of the provincial territory. This number represents $87,33\%$ of the considered small alpine catchments.
It is interesting to note that only 159 basins, $9,37\%$ of the total, have reported events in the relative fan area during the considered time window.

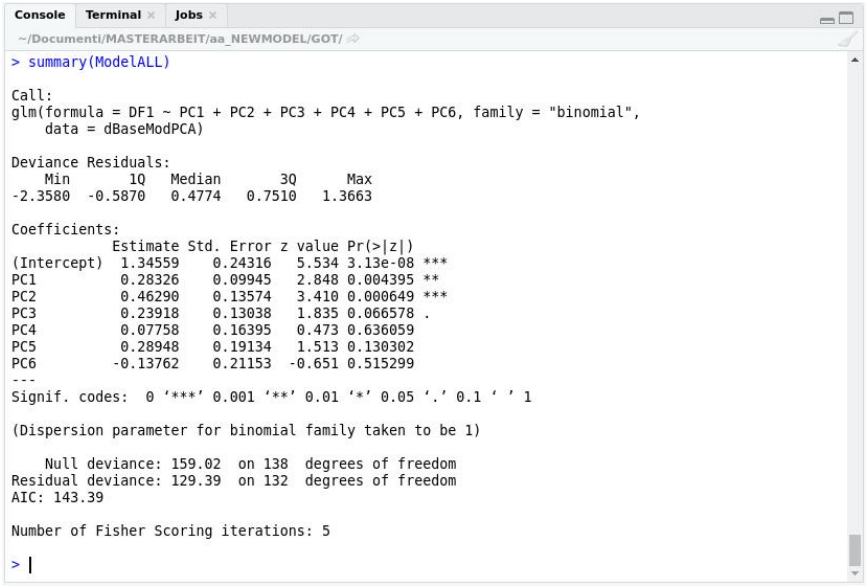
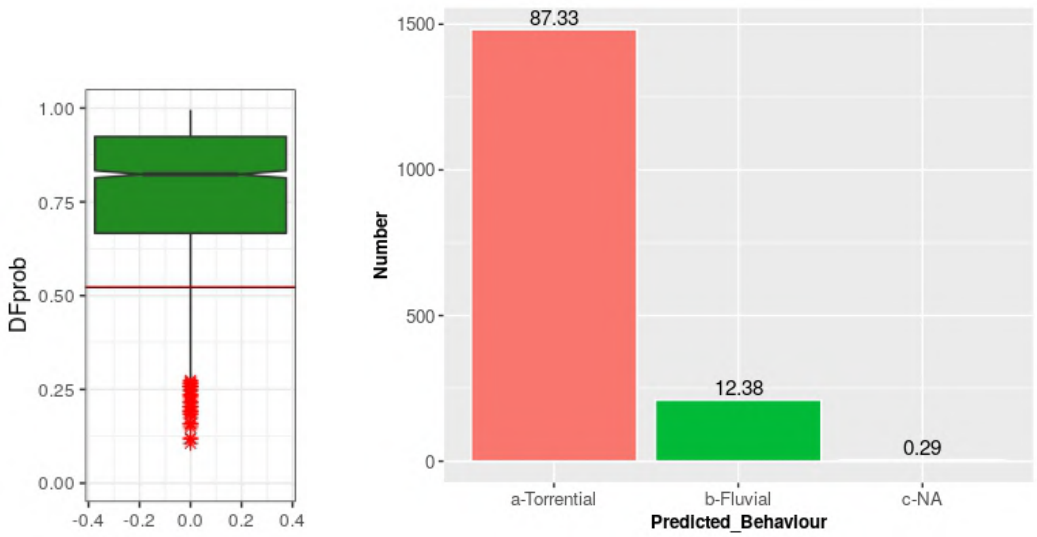


Figure 6.7: Logistic regression (entire database): (a) summary *ModelALL*, (b) sorted calculated probabilities (red solid line $0,5222838$ predicted probability threshold).



(a) (b)

Figure 6.8: Predicted probabilty and basins classification (entire database): (a) boxplot predicted probability (red solid line $0,5222838$ predicted probability threshold), (b) basins classification based on the proposed model.

6.4.2 Plotting the results

In addition to the numerical description of the model results, maps can be plotted for a visual interpretation. At the end of the computation procedure, the shapefile containing all the basins describing variables and the predicted debris flow probability and class distinction is exported from the *R* environment. The plotting operation is performed using *QGIS* because of the simple and fast needed procedure.

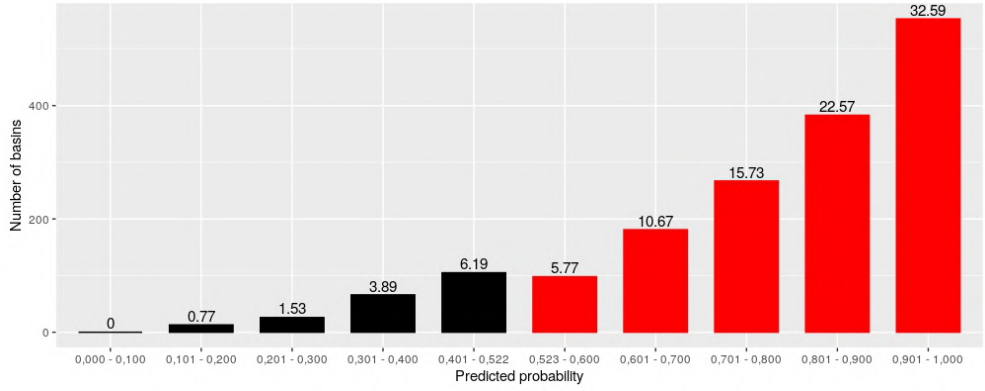
The entire catchment database is plotted in Figure 6.11. The upper panel of the figure represents the basins sorted into 9 classes on the basis of the predicted probability to generate and transport debris flows or debris flood down to the fan area (probability between $0,000$ and $0,200$ are summed up in one group). The polygons are coloured using a color palette which goes from blue to red. Low predicted probabilities are displayed by cold colors (blue) while high predicted probabilities are represented by warm colors (red). Intermediate probabilities are symbolized by different color shades. The important step, between probabilities below and above the used cutoff is depicted by the transition between the light blue and the light yellow polygons.

The numbers of the catchments represented by every probability class of Figure 6.11a are reported in Table 6.12. The values in the table are displayed in Figure 6.9, where the upper panel shows a histogram of the number of basins counted for each probability class. The lower panel shows the cumulative number of basins with increasing predicted probability. Both the panels use black color for probability classes predicted as dominated by fluvial processes and red color for dominant torrential processes.

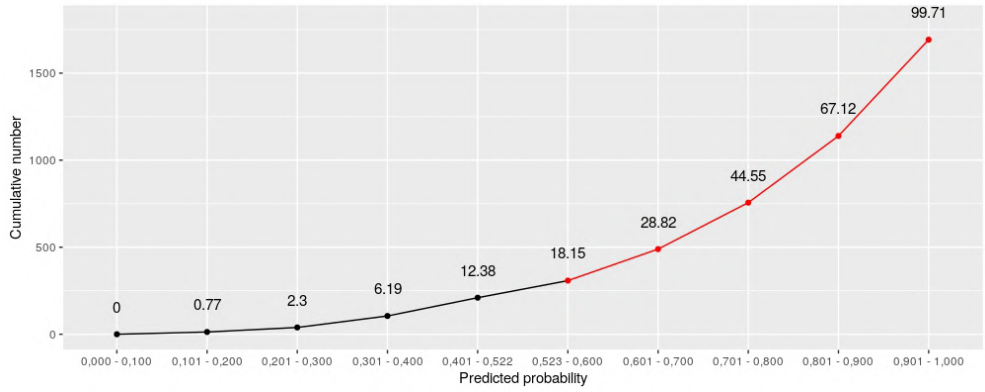
With the adoption of the alternative probability cutoff equal to $0,5222838$, some basins included between the default threshold (equal to $0,50$) and the new cutoff, are predicted into the class 0 group instead of class 1 group. A total 25 watersheds are affected by this change representing $1,47\%$ of the entire database. Their spatial distribution is shown in Figure 6.10. The 25

Table 6.12: Predicted probability, distribution into the classes displayed in Figure 6.11a (5 NA's)

Probability class	Number	Num. %	Cumulate	Cum. %
0,000 - 0,100	0	0,00	0	0,00
0,101 - 0,200	13	0,77	13	0,77
0,201 - 0,300	26	1,53	39	2,30
0,301 - 0,400	66	3,89	105	6,19
0,401 - 0,522	105	6,19	210	12,38
0,523 - 0,600	98	5,77	308	18,15
0,601 - 0,700	181	10,67	489	28,82
0,701 - 0,800	267	15,73	756	44,55
0,801 - 0,900	383	22,57	1139	67,12
0,901 - 1,000	553	32,59	1692	99,71



(a)



(b)

Figure 6.9: Predicted probability, distribution into the classes displayed in Figure 6.11a (5 NA's): (a) number of basins for the predicted probability classes, (b) cumulative number of basins for the predicted probability classes. In black are represented basins with predicted fluvial dominated behaviour, while in red with torrential dominated behaviour.

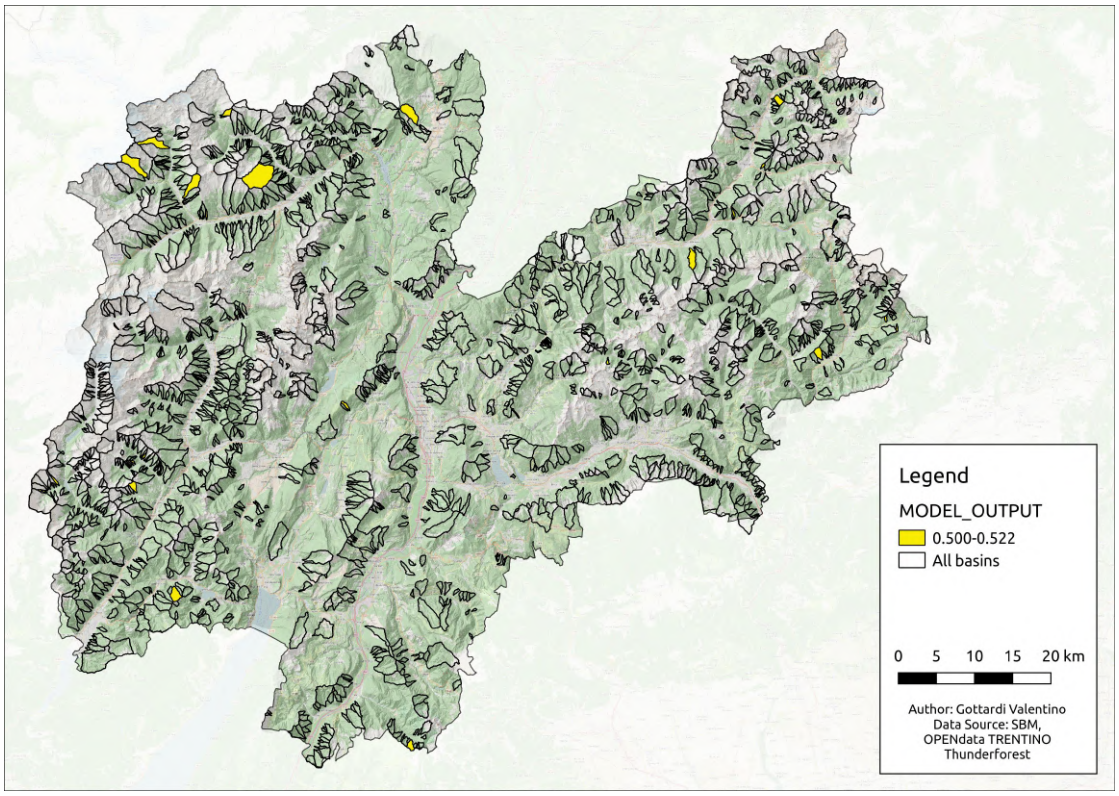
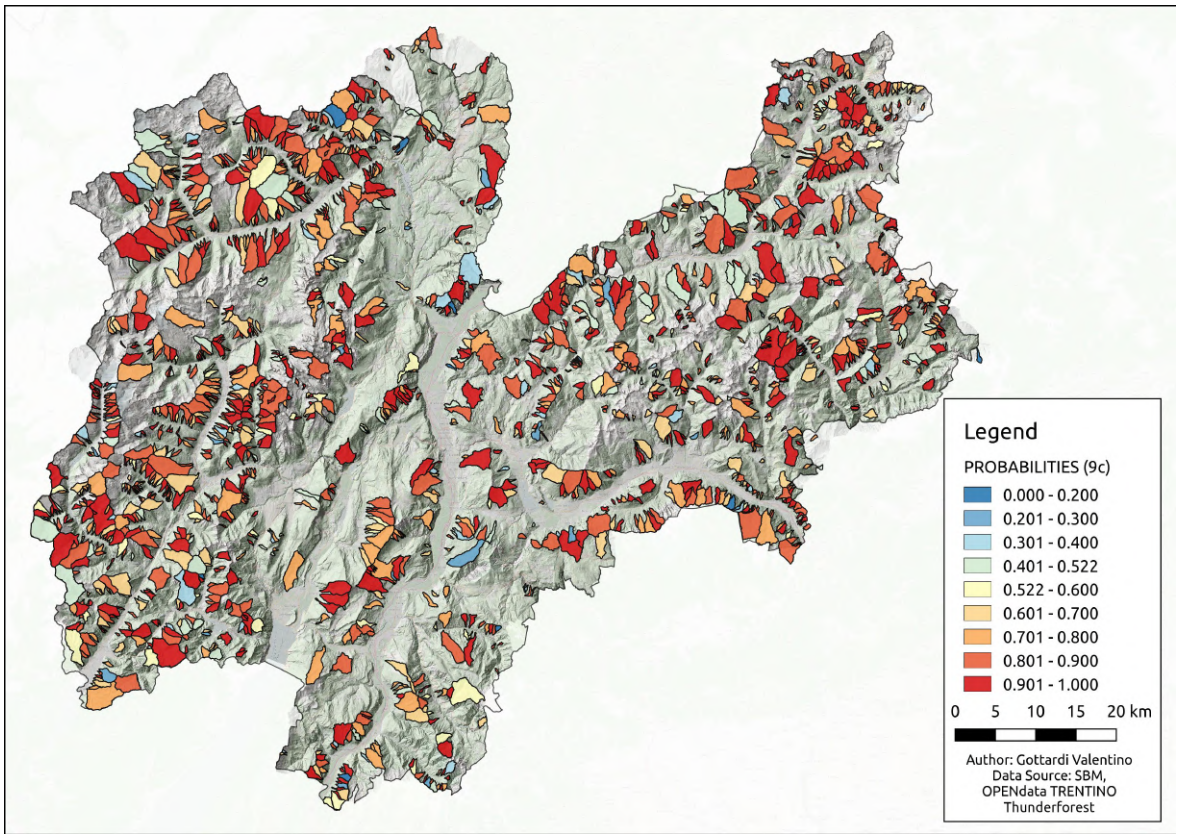


Figure 6.10: Basins with predicted likelihood between default threshold and alternative cutoff.

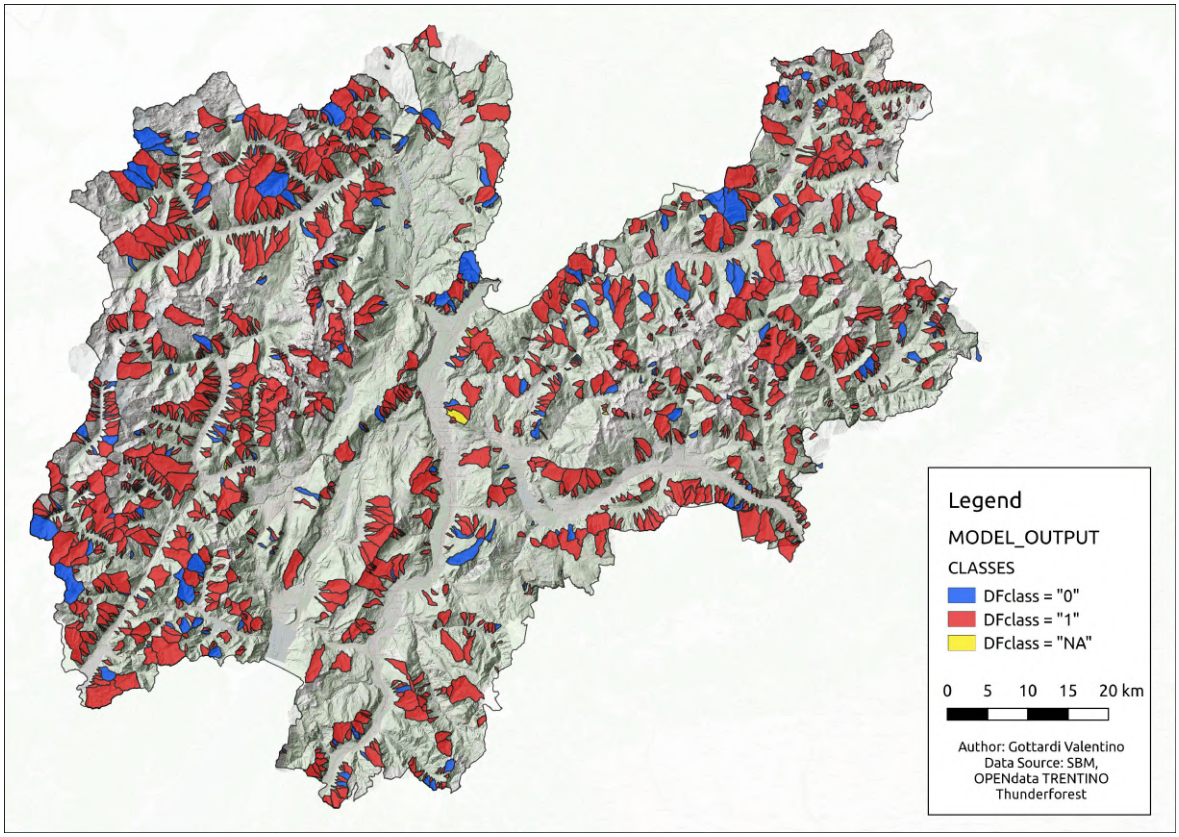
basins are plotted with yellow polygons, together with the boundaries of the other catchments in black.

The distribution of the catchments, classified into $DFclass$ = "0" and $DFclass$ = "1", is shown in Figure 6.11b. As already explained the classification is based on the predicted probabilities and the calculated alternative cutoff value. Blue polygons represent basins with predicted debris flow probabilities below the cutoff value and considered as dominated by fluvial processes. Red polygons shown the catchments with predicted torrential behaviour. The 5 basins without prediction are depicted by yellow polygons.

In order to produce an easily readable basin classification, the provincial territory is divided into four areas. Each of these zones is plotted separatly in the following pages.
Basins located in the area of the Sole and Non valleys, as well as the upper reaches of Sarca torrent and Adige river are represented in Figure 6.12. The lower part of the Adige river, the lower reach of Sarca torrent and the Chiese and Leno torrents are depicted in Figure 6.13. the tributary area of the Avisio torrent is shown in Figure 6.14 while catchments of Fersina and Vanoi torrents and Brenta river are represented in Figure 6.15. The other small watersheds are included but not mentioned in the description.

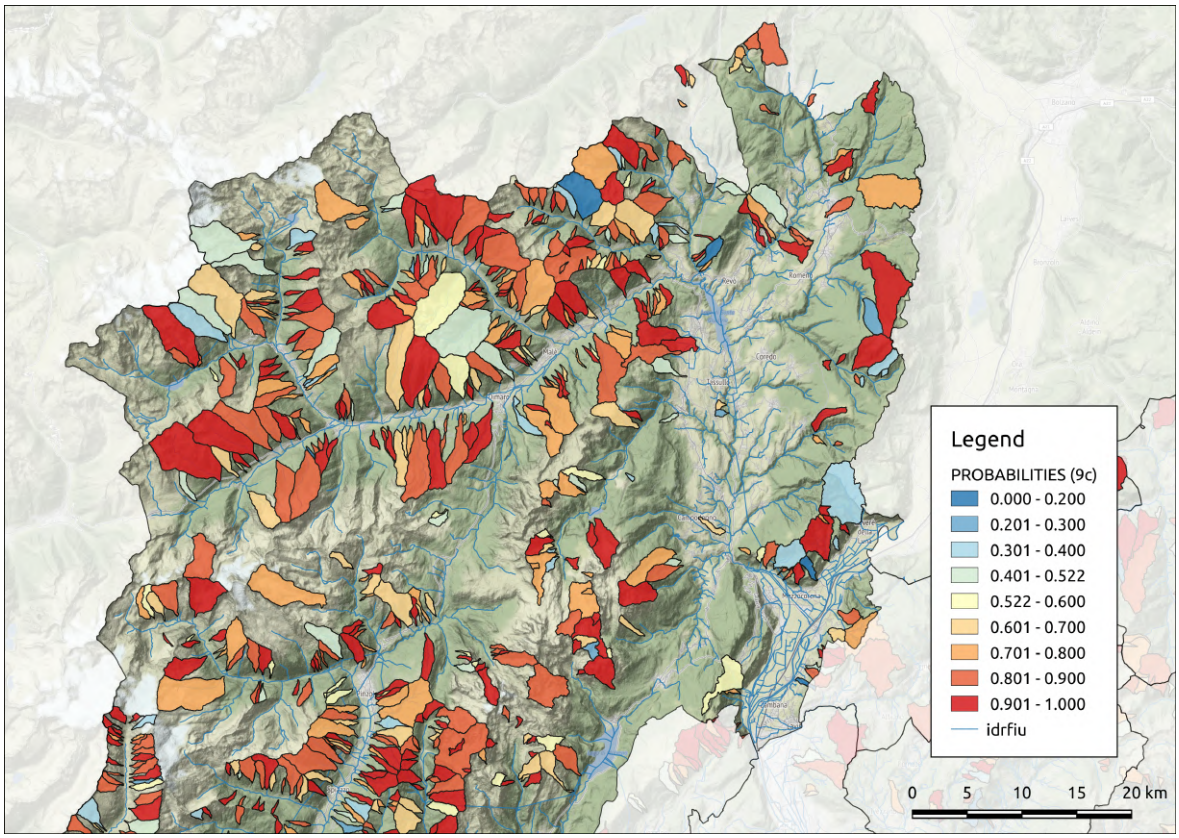


(a)

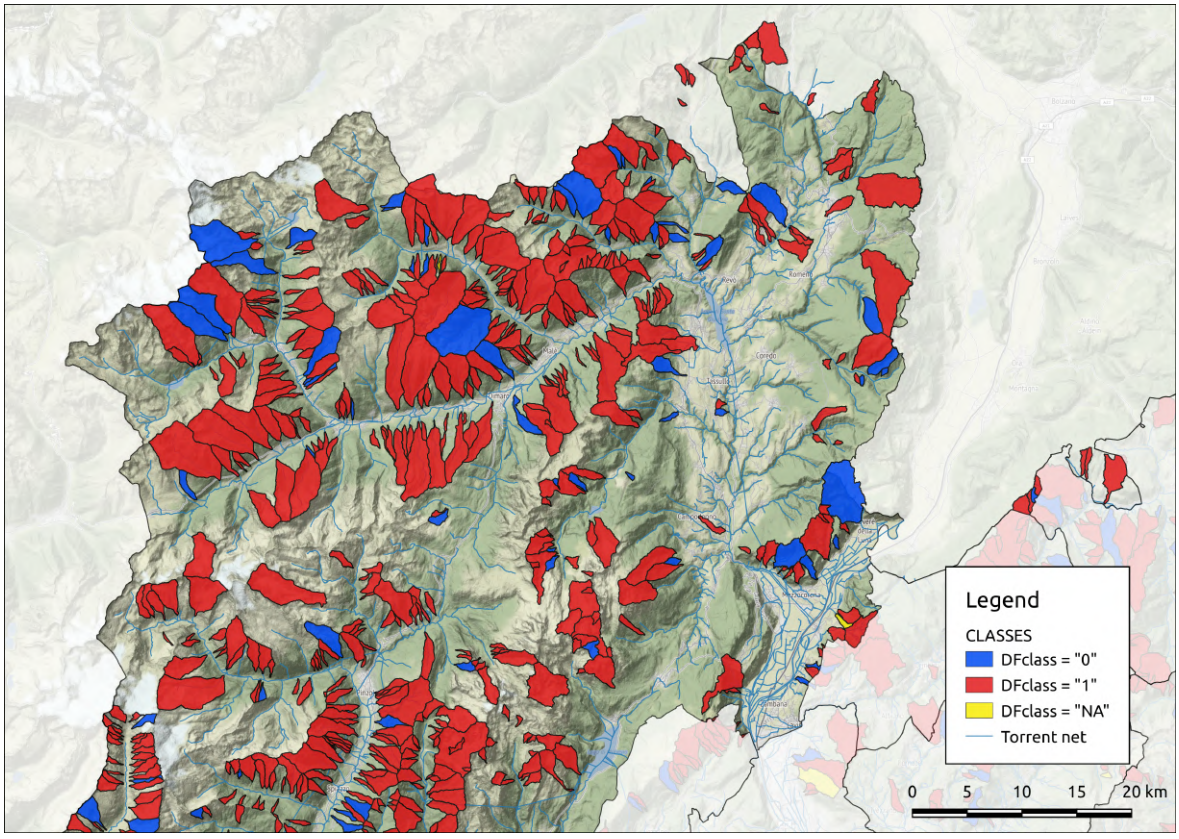


(b)

Figure 6.11: Modeling results: (a) predicted probability divided into 9 classes, (b) predicted classes.

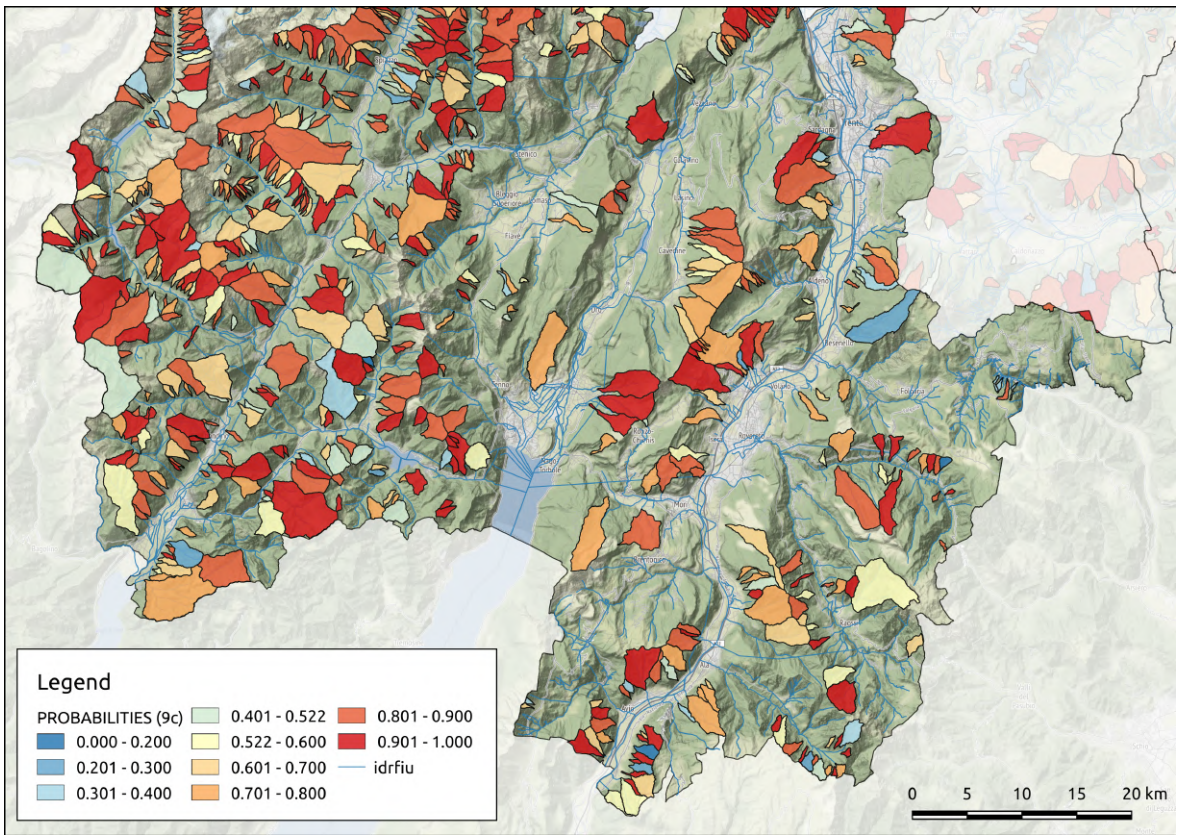


(a)

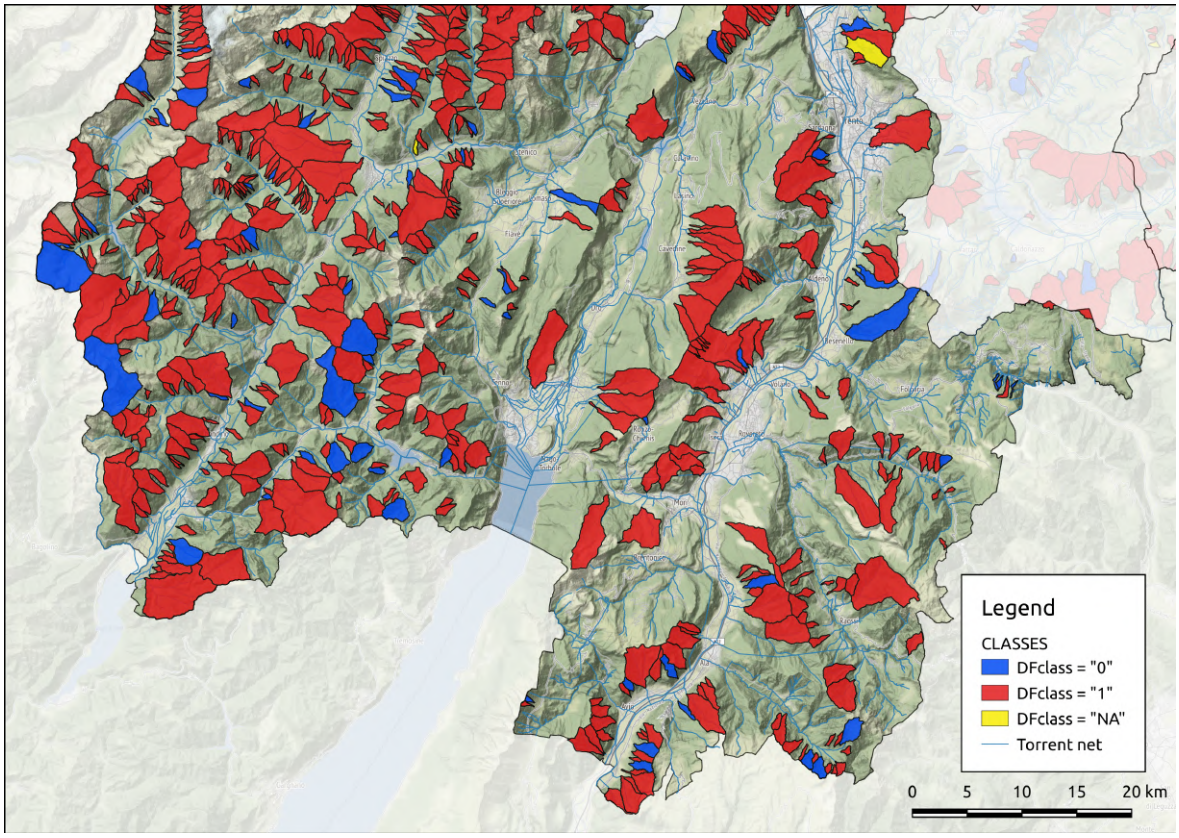


(b)

Figure 6.12: Noce torrent, upper reaches Sarca and Adige: (a) predicted probability, (b) predicted classes.

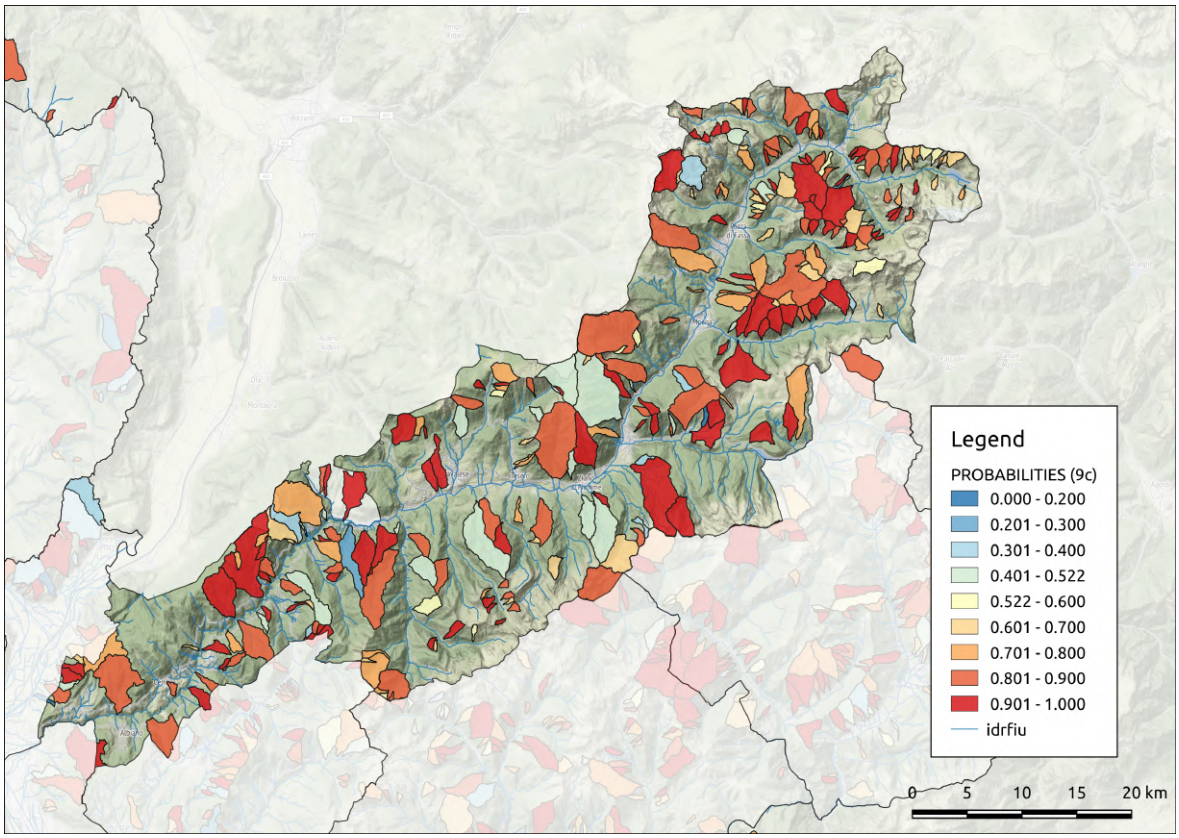


(a)

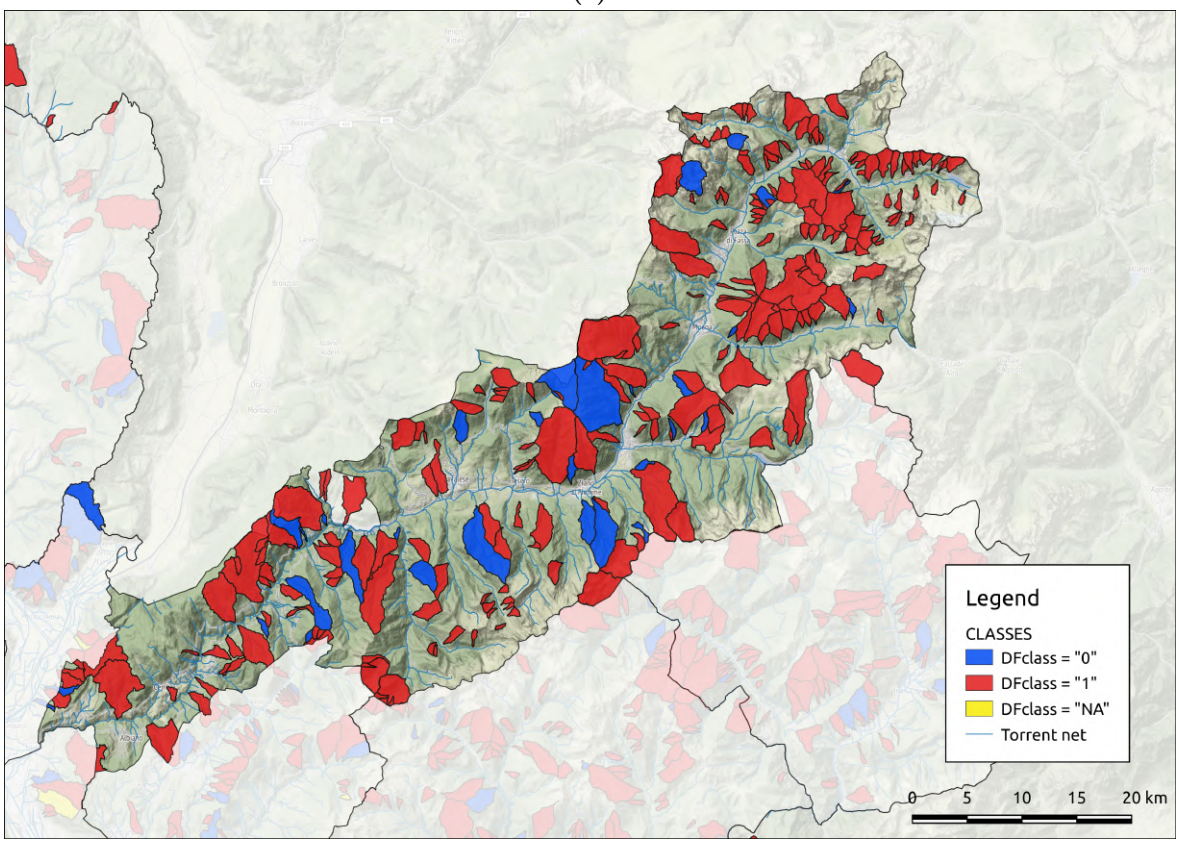


(b)

Figure 6.13: Chiese and Leno torrents, lower reaches Sarca and Adige: (a) predicted probability, (b) predicted classes.

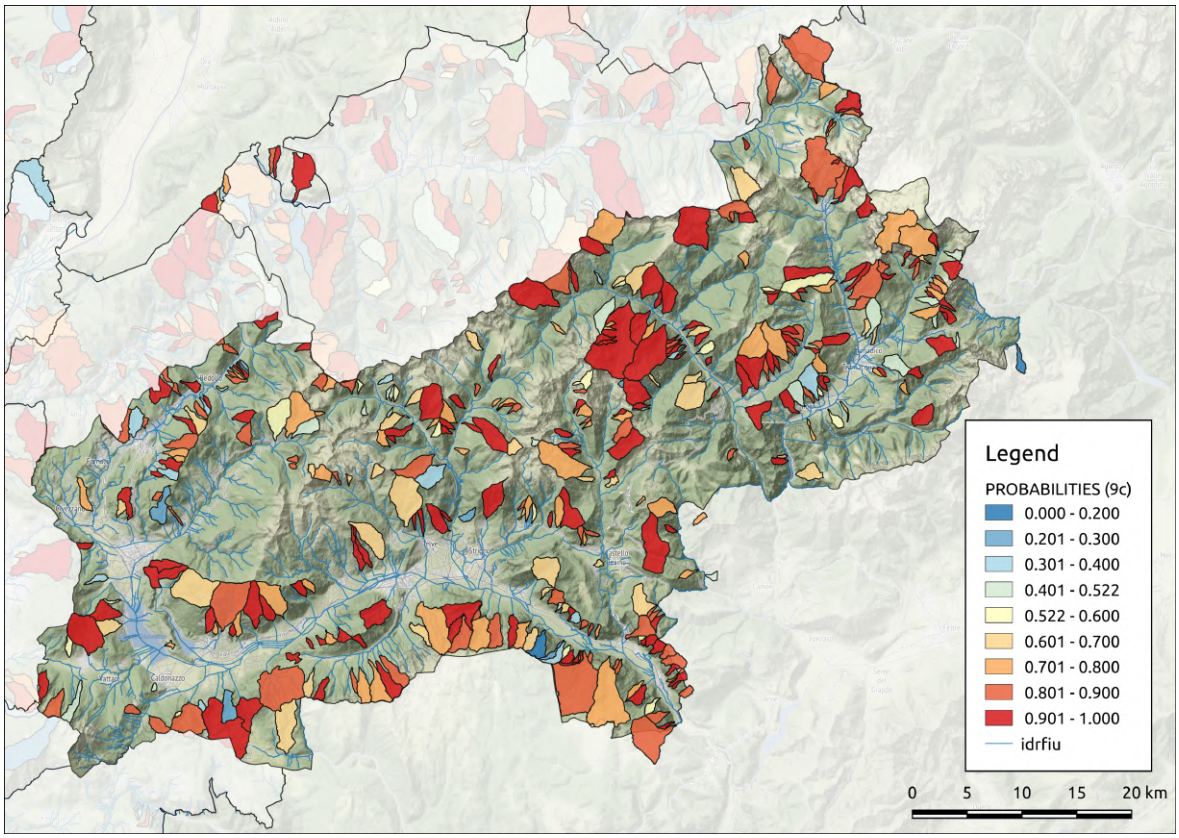


(a)

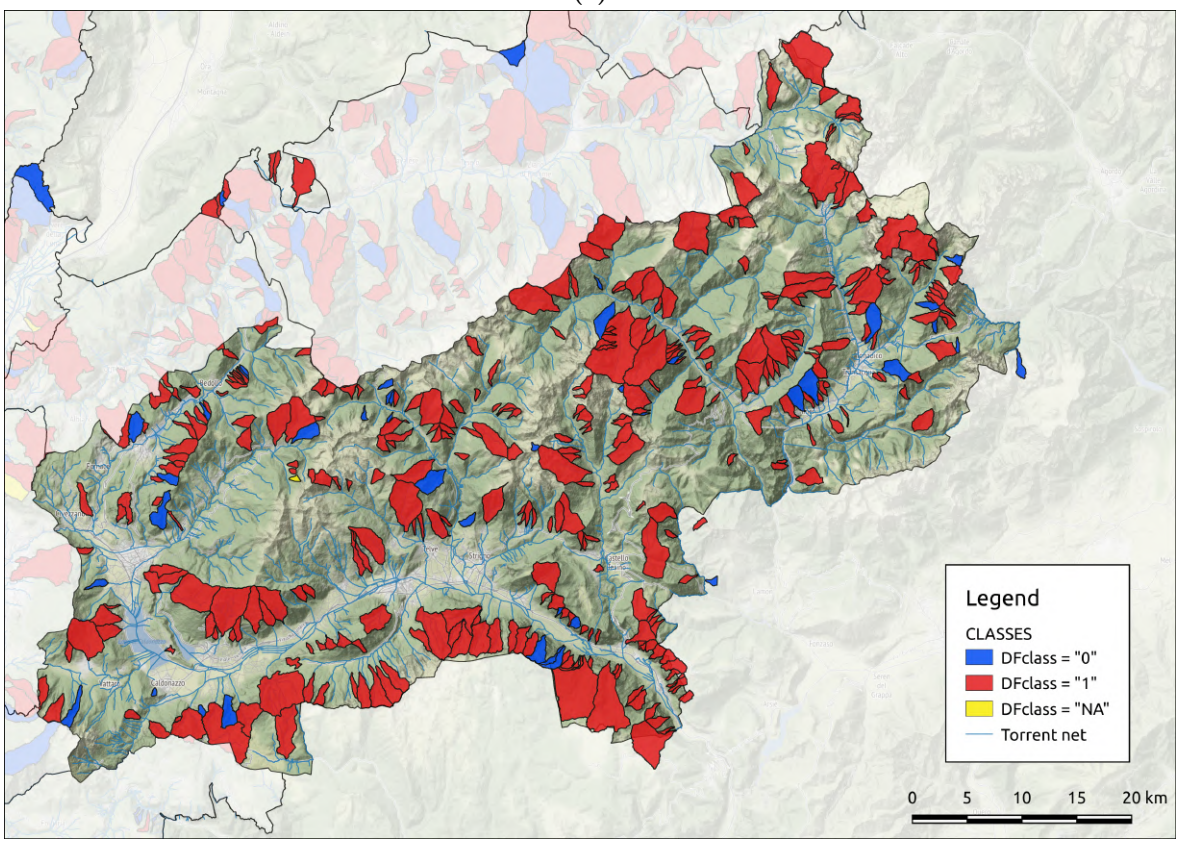


(b)

Figure 6.14: Avisio torrent: (a) predicted probability, (b) predicted classes.



(a)



(b)

Figure 6.15: Fersina torrent, Brenta river and Vanoi torrent: (a) predicted probability, (b) predicted classes.

Chapter 7

Discussion

Every step of the proposed model is briefly discussed considering available data, the method and the results.

7.1 Event selection

A precise selection of the reported events on the provincial territory is of crucial importance. This is particularly true because of the various information sources used to build the event cadaster. The 3522 events in total are characterized by strong differences in information source, time of occurence and precision of the description. Obviously old parochial registers or newspaper articles report imprecise descriptions of the occured torrential process, focusing often only on damages and casualties.
In addition to this, as already mentioned in Section 5.2.1, Castiglioni (1971) reported that in the Italian part of the Alps the most intense torrential processes have been often classify as landslides instead of debris flows or debris flood. Also Marchi and Cavalli (2007) stressed this missclassification for the events related to the epochal 1966 storm.
Only events with certain process type and location are used for the model. This is possible thanks to the presence of the arguments "reliability" in the event cadaster. Given that the model will predict the probability that an event reach the fan area, also a spatial selection is performed in order to get only signals from the fan of the considered basins. Observations connected with landslides, avalanches or river floods are also excluded. 406 events meet these criteria.

Important is the visual control of the filtering steps, in order to check the correct and plausible spatial location of the signals. When, from the description of the event, appears that the signal point is wrongly positioned a manual relocation is advisable. This ensures to maintain as much as possible events to build the model. Working with a GIS environment for this purpose is the best solution.

The used time window allows to cover the long recurrence intervall of the investigated processes but will inevitably bring also uncertainties into the model. On the other hand, the precise and punctual description of process characteristics and damages of lastly occured events will permit more reliable prediction in the future.

7.2 Basin selection

A model have to be built and trained based on a pool of selected basins. As widely explained in Section 5.2.2, only basins with a well defined behaviour are selected. The proposed procedure identifies 103 basins with only debris flows and debris floods reported in fan area. Other 36 basins show bedload transport and clear water discharge events. From the initial 1697 basins, only 139 are considered for the creation of the model.
The low amount of available basins does not allow to split the database into the train and test subsets. For this reason a resampling technique has to be implemented. Nevertheless, the low number of informative catchments remains one of the most unfavourable aspect of the suggested model.

It is to highlight that few fans are not properly drawn. This is mainly due to the close position of two neighbouring catchments with a simultaneous growth. The contribution of a single torrent on a certain event is not always straightforward. In order to maintain a good match with the provincial database, watershed's and fan's boundaries have not been modified.

The spatial distribution of the selected basins is visible in Figure 6.2. All the provincial territory is covered by catchments although not with the same density. Some areas, like the Val di Sole and the Valsugana are more extensively represented. The reasons of this fact are various. Beside the different climatic and geologic conditions also the types of torrent management and mitigation measures play an important role. In addition to the already mentioned parameters, the use of fan areas can differ in the valleys, passing from meadows, to vineyards or orchard, with comprehensible various monetary values and thus precision of damage description. The presence on the fans of infrastructures and settlements raises the likelihood of a torrential process to be signaled. Another possible source of diversity is the subdivision of the province into torrent management areas, each of them supervised by an officer. Expecially before the introduction of a common report protocol for the event description, the officer's sensibility about minor events could be strongly different.

7.3 Feature selection

The first feature selection approach, as explained in Sections 5.2.3 and 6.2, aims to choose only variables characterized by strong separation ability and low internal correlation. The method itself produce valuable results removing not informative variables and maintaining only 9 features.
Given the ability of the PCA to deal with highly correlated predictors, the consequent decision is to skip the manual feature selection and maintain only the PCA. This is mainly done in order preserve as much data information as possible.

7.4 Model building and fitting

7.4.1 Principal component analysis

This method delivers a simplified view of the entire dataset using linearly uncorrelated variables called principal components (PCs). The main part of the database total variance is condensed in the first few PCs. Selecting only the first 6 PCs allows to simplify the analysis maintaining a good explained variance ($81,5\%$ of the total) and thus information level (see Section 1.7.3 for the explanation).
Although the method calculates new predictors rearranging the input database, the contribution of every single variable on a principal component is expressed by the calculated loading values (see Section 6.3.1). For the first PC, the loadings are easy to interpret. The contribution on the following PCs, instead, are difficult to understand.
Representing more than 80% of the total variance using only 6 variables can be considered as a good compromise between information completeness and model simplification. As easy to understand, if the minimum amount of explained variance is reached with a minor number of PCs, the result will be a more clear and readable model. But this is not the case for the given database.

7.4.2 Logistic regression

Before the Logistic regression is performed, some preparatory steps have to be done. One of the first problem is to get a train and a test subset. This is quite troubling when dealing with small datasets.
A resampling technique is needed to overcome the problem of the low number of available basins. The method of bootstrapping with replacement is used. It allows to assign measures of accuracy to sample estimates. The useful idea of bootstrapping is that, if the analysis of a sample can lead to inferences about the population, in the same way the analysis of a resample can form inferences about the sample (the selected catchments). Furthermore, since the sample is known, inferences about it based on the resampled data are measurable and assessable.

R allows to easily launch a Logistic regression simply indicating the dependent variable, the predictors (the first 6 PCs) and the type of regression which has to be done.
The diagnostic ability of the Logistic regression is visually illustrated and investigated using the ROC curve. It depicts the tradeoff between sensitivity and specificity, varing the discrimination threshold. The ROC curve (Receiver Operating Characteristic curve) is an useful tool to determine the best cutoff and to get other numerical performance measurements.

The AUC (Area Under the ROC Curve) tells how much the model is capable of distinguishing between classes. For the proposed model it is equal to $69,1\%$, with a confidence intervall between $46,5\%$ and $91,7\%$. The AUC measures the quality of the model's predictions, regardless the chosen classification threshold. This is an appropriate approach when the cost of false negatives and false positives is the same, like in the present work. The value of the calculated AUC is moderate low or intermediate, considering that an AUC of $50,0\%$ represents a model without

discrimination capacity to distinguish between positive class and negative class. An AUC equal to $100,0\%$ accounts for perfect separation ability.

In order to maximize the Youden's J index, an alternative threshold value equal to $0,5222838$ is chosen. Passing from the default threshold of $0,50$ to the new cutoff value, the Youden's J index changes from $0,3636$ to $0,4182$. If sensitivity and specificity are of equal diagnostic importance, an index value of 1 represent a perfect test. Instead, an index equal to 0 means that the test has no diagnostic value.

The performances of the Logistic regression (estimated from the test subset and applying the alternative cutoff value) can be summarized and discussed using the below listed indeces.

- Overall accuracy rate vs. No-information rate: these two indices have to be considered together. The first index measures the amount of correctly predicted basins in both of the classes. The result is then divided by the total number of observations. The test of the model gets an overall accuracy rate equal to $0,7778$ meaning that the $77,78\%$ of the test subset was correctly predicted. The second index represents the best guess given no information beyond the overall distribution of the classes we are trying to predict. In this case the $81,48\%$ of the test subset belong to the class 1 (debris flow prone). The no-information rate becomes the value $0,8148$. It means that, if we predict all the basins to belong to the class 1, we will correctly assess the $81,48\%$ of the set. According to Kuhn and Johnson (2013), models with accuracy rate greater than the no-information rate might be considered reasonable. Considering these two indeces, the model does not achieve properly predicting basin's behaviour of the test subset. The reasons of this poor performance are unclear. Also other literature methods, discussed in the following section, show limited separation ability.

- Kappa statistic: this index accounts the accuracy that would be generated simply by chance (see Equation (5.1)). The proposed model gets a Kappa statistic value equal to $0,3622$. According to Kuhn and Johnson (2013), Kappa values within $0,30$ to $0,50$ indicate reasonable agreement, while values of 1 and 0 express perfect concordance and no agreement respectively.

- Sensitivity: the calculated sensitivity is equal to $0,8182$ which can be defined as a good level. This is an important and positive feature of the model because it ensures that the majority of the debris flow prone basins are detected and predicted into the class 1. The proposed approach does not make differences between false predicted class 1 and class 0 basins. However the ability to detect all the probable debris flow prone catchments is still of crucial interest.

- Specificity: this index measures the proportion of correctly predicted fluvial behaviour watersheds (no debris flow prone watersheds classified into the class 0). The calculated specificity is equal to $0,6000$ which is an intermediate value. This feature is not very high indicating that some basins characterized by fluvial behaviour are predicted as producing high intense events. This is problematic if the aim of the analysis is to detect basins with low debris flow

likelihood in order to exclude theyr planning and mitigation measures construction. For the same reason the misclassification of many basins leads to the lack of monetary founds for the operations on the most hazardous torrents.
A possible reason about the overestimation of the debris flow probability of many basins is the probable incomplete or erroneous knowledge of the event series. Working at a regional scale it is higly possible to not detect geomorphological features in fan areas as depositional lobes or levees. Coupling punctual information (e.g. silent witness, stratigraphical survey) with the event cadaster datas can sensibly improve the model prediction consistency.

- Youden's J: as already mentioned above, the Youden's J index is maximized adopting an alternative separation threshold and gets a value of $0,4182$. Using the alternative cutoff leads to a sensible increase of the Youden's J index, which however remains in a intermediate value range. The used cutoff aims to maximize this index choosing the more appropriate balance between sensitivity and specificity. If the interest is posed to have, alternatively, sensitivity or specificity as high as possible also the threshold has to be properly changed.

7.4.3 Comparison with other method

The proposed model is compared with a number of methods in the literature to assess its ability to dis-criminate between dominant behaviours. The applied methods of D'Agostino (1996) and DeScally and Owens (2004) use only already available predictors, as the Melton ruggedness number and the mean fan slope.
The specific performances have been described in the Section 6.3.4. It is interesting to discuss the results of the various methods on the basis of the descriptive indices.

Considering that the model is trained in order to maximize the Youden's J this is discussed first. The four methods from the literature obtain Youden's J values considerably lower than the proposed model, approaching 0, a value at which a model shows no prediction capability. This index mirrors values of sensitivity and specificity. On the test subset, the *Dago1* model achieves an extremely high sensitivity ($0,9773$) thanks to the low MRN threshold. The same model scores really bad specificity because the majority of the basins with fluvial behaviour are predicted as debris flow prone. Almost the opposite occurs for the *Dago2* model, where the high multiplication coefficient (set to 14) of the MRN leads to the classification of many torrents into the fluvial processes dominated catchments. Therefore, the specificity is quite high but the sensitivity remains low.
More balanced specificity and sensitivity values are shown by the two models from DeScally and Owens (2004) but the calculated Youden's J indexes remain lower than the one characterizing the proposed model. Each of the methods does not use weights or penalties for the two different types of errors and this makes possible to select the best model looking at the Youden's J index.

Looking at the Kappa statistic, the proposed model with $0,3622$ gets an higher evaluation in comparison with all the applied literature methods, which are rated worst than "moderate to good agreement" (Kuhn and Johnson, 2013).

The comparison between the overall accuracy rate and the no-information rate is suboptimal for both the proposed model and the adopted literature methods. This does not allow to discuss and compare the performance of the different procedures based on these index.

Although all the four literature methods perform in an insufficient way on the test subset, the best between the other seems to be the *DeSc1* model. The high MRN threshold, equal to $0,75$, reflects better than the other the local regional behaviour of small alpine torrents, It achieves good sesitivity and average specificity, but lacks for Kappa statistic, Youden's J index and overall accuracy rate vs. no-information rate.

Chapter 8

Conclusions

The available databases allow the creation of the required classification model for the small alpine catchments of the Autonomous Province of Trento.
The morphological features of the 1697 watershed-fan systems are created with an automated approach and are generally robust. More important is the filtering and inspection of the event database. This is done in order to build and train the model using only well reliable data. Based on these constituent database considerations, the proposed model is based on all available basin polygons and 269 reliable events that occurred in the fan areas.

Model building is based on 139 basins, selecting only those watersheds with well-defined behaviour. 103 basins showed only debris flows and debris floods, while 36 reported bedload transport or clear water discharge.

Some considerations can be made and a few questions are still waitimg for a certain answer:

- Did not the river-dominated fans undergo more intense processes? It is highly possible that, during the post-glacial period, large debris flows also occurred in basins now considered to be dominated only by fluvial processes. This ancient behavioural change did not affect the usefulness of a forecast because we are interested in the actual behaviour of torrents. However, geomorphological features created in the past can be responsible for forecast errors when used as predictors of models.

- Are all events reported? In basins with very intense processes, clear water discharge and bottom load transport events may not be recorded. This can also be said for basins located in remote and unused areas.

- The procedure only selects a basin if at least one transport process occurs in the fan area. It is possible that some truly informative basins did not produce any events in the time window considered or that deposition occurred in the transport area.

- Future climate changes, and thus possible alterations in the behaviour of torrents, cannot be taken into account (annual precipitation distribution, thawing of permafrost in high alpine regions, etc.).

Considering the not excellent quality of the available databases, the proposed model is of high interest despite the strong limitations described. This is especially true in comparison with other approaches that use only a few predicting parameters.
The proposed method can certainly be improved by changing the selection of events and thus the basins considered. For example, selecting all basins with events reported in the fan and or transport zone will result in a larger pool of usable basins. This will probably positively influence the reliability of the model.
Another possible improvement can be achieved by revising the geometry of the fans. Some problems are created by poorly delineated fans (e.g. eroded by the downstream river or fed by neighbouring basins).

The application of a resampling method such as bootstrapping efficiently solved the limited pool of selected basins. Then, the choice of using principal component analysis allows highly correlated variables to be treated without the need for severe feature removal. This is certainly a positive aspect of the analysis but brings with it a certain amount of uncertainty. For this specific element the proposed model did not meet the required transparency in data manipulation.
On the other hand, the application of logistic regression results in an intelligent model output that is easy to read and interpret, as requested at the beginning of this paper.

In conclusion, the proposed model does not perfectly fulfil the aim of the thesis work, but it does indicate the weaknesses of the provincial databases available so far. The improvement of the homogeneity of event registration and the cleaning of less informative signals will probably lead to a better evaluation of the cadastre of important events as well as to a more powerful and valid prediction, keeping the procedure automated.

REFERENCES

Bibliography

Abdi, H. and Williams, L.J. (2010). *Principal component analysis.* WIREs Comp Stat, 2: 433-459.

Ayalew, L., & Yamagishi, H. (2005). *The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan.* Geomorphology, 65(1-2), 15-31.

Aulitzky, H. (1984): *Vorläufige, zweigeteilte Wildbachklassifikation.* In: Wildbach- und Lawinenverbau (Hrsg.), 48. Jg., Sonderheft, 7- 60 (in German)

Aulitzky, H. (1992): *Die Sprache der „Stummen Zeugen“*, Tagungspublikation Interpraevent, Band 6. 139-174 (in German)

Bagnold, R. A. (1954). *Experiments on a gravity-free dispersion of large solid spheres in a Newtonian fluid under shear.* Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 225(1160), 49-63.

Bardou, E. (2002). *Méthodologie de diagnostic des laves torrentielles sur un bassin versant alpin* (Doctoral dissertation, Ecole polytechnique fédérale). (in French)

Bardou, E., Fournier, F. & Sartori, M. (2003). *Paleofloods reconstruction on Illgraben torrent (Switzerland): a need for today frequency estimation*, International Workshop on Paleofloods, Historical Data & Climatic Variability: Applications in Flood Risk Assessment, Barcelona.

Bergmeister, K. (2009). *Schutzbauwerke gegen wildbachgefahren: Grundlagen, entwurf und bemessung, beispiele.* John Wiley & Sons. (in German)

Bertrand, M., Liébault, F., & Piégay, H. (2017). *Regional Scale Mapping of Debris-Flow Susceptibility in the Southern French Alps.* Journal of Alpine Research— Revue de géographie alpine, (105-4).

Blais Stevens, A., & Behnia, P. (2016). *Debris flow susceptibility mapping using a qualitative heuristic method and Flow-R along the Yukon Alaska Highway Corridor, Canada.* Natural Hazards and Earth System Sciences, 16(2), 449.

Borga, M. (2003). *Analisi del regime delle piogge intense per la Provincia Autonoma di Trento.* Ufficio Pianificazione e Rilevazioni Idriche PAT. Associazione Italiana di Idronomia. (in Italian)

Breiman L., Friedman J. H., Olshen R. A., & Stone C. J. (1984). *Classification Regression Trees*, Wadsworth International Group, Belmont, California.

Bull, W. B. (1964). *Geomorphology of segmented alluvial fans in western Fresno County, California.* US Government Printing Office.

Bulmer, M. H., Barnouin Jha, O. S., Peitersen, M. N., & Bourke, M. (2002). *An empirical approach to studying debris flows: implications for planetary modeling studies.* Journal of Geophysical Research: Planets, 107(E5).

Bundes Ministerium für Land- und Forstwirtschaft (1984), *100 Jahre Wildbachverbauung in Österreich.* In self-publishing, Wien (in German)

Caldonazzi, M., & Avanzini, M. (2011). *Storia geologica del Trentino.* Provincia Autonoma di Trento (in Italian)

Carrara, A., Crosta, G., & Frattini, P. (2008). *Comparing models of debris-flow susceptibility in the alpine environment.* Geomorphology, 94(3-4), 353-378.

Castiglioni, G.B. (1971). *Le calamità naturali nelle Alpi.* Pages 7-37 in Atti del XXI Congresso Geografico Italiano, Verbania, Italy (in Italian).

Cavalli, M., & Grisotto, S. (2006). *Individuazione con metodi GIS delle aste torrentizie soggette a colate detritiche: applicazione al bacino dell'alto Avisio (Trento).* Nuova Bios, Idronomia Montana, 26, 1-11. (in Italian)

Cavalli, M., Goldin, B., Comiti, F., Brardinoni, F., & Marchi, L. (2017). *Assessment of erosion and deposition in steep mountain basins by differencing sequential digital terrain models.* Geomorphology, 291, 4-16.

Ceriani, M., Crosta, G., Frattini, P., & Quattrini, S. (2000, June). *Evaluation of hydrogeological hazard on alluvial fans.* In Proc. Int. Symp INTERPRAEVENT 2000, Villach, Austria (Vol. 2, pp. 213-225).

Coe, J. A., Godt, J. W., Parise, M., & Moscariello, A. (2003, September). *Estimating debris-flow probability using fan stratigraphy, historic records, and drainage-basin morphology, Interstate 70 highway corridor, central Colorado, USA.* In USA Proceedings of the 3rd International Conference on Debris-Flow Hazards Mitigation: Mechanics, Prediction, and Assessment, Davos (Switzerland) (Vol. 2, pp. 1085-1096).

Coe, J. A., Kinner, D. A., & Godt, J. W. (2008). *Initiation conditions for debris flows generated by runoff at Chalk Cliffs, central Colorado.* Geomorphology, 96(3-4), 270-297.

Costa, J. E. (1984). *Physical geomorphology of debris flows.* In Developments and applications of geomorphology (pp. 268-317). Springer, Berlin, Heidelberg.

Davies, T. R. H. (1986). *Large debris flows: a macro-viscous phenomenon.* Acta Mechanica, 63(1-4), 161-178.

D'Agostino, V. (1996). *Analisi quantitativa e qualitativa del trasporto solido torrentizio nei bacini montani del Trentino Orientale.* Associazione Italiana di Ingegneria Agraria, 1. (in Italian)

D'Agostino, V. (2006). *Le opere di idraulica torrentizia per il controllo dei sedimenti.* Quaderni di idronomia Montana, 26, 231-250. (in Italian)

de Scally, F. A., & Owens, I. F. (2004). *Morphometric controls and geomorphic responses on fans in the Southern Alps, New Zealand.* Earth Surface Processes and Landforms, 29(3), 311-322.

Dong, J. J., Lee, C. T., Tung, Y. H., Liu, C. N., Lin, K. P., & Lee, J. F. (2009). *The role of the sediment budget in understanding debris flow susceptibility.* Earth Surface Processes and Landforms, 34(12), 1612-1624.

Efe, R., Ozturk, M., & Atalay, I. (Eds.). (2011). *Natural environment and culture in the Mediterranean region II.* Cambridge Scholars Publishing.

Eisbacher, G. H. (1984). *Destructive mass movements in high mountains.* Geological Survey of Canada.

Ferro V., 2002, *La sistemazione dei bacini idrografici*, MCGraw-Hill (in Italian)

Heiser, M., Scheidl, C., Eisl, J., Spangl, B., & Hübl, J. (2015). *Process type identification in torrential catchments in the eastern Alps.* Geomorphology, 232, 239-247.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). *Unbiased recursive partitioning: A conditional inference framework.* Journal of Computational and Graphical statistics, 15(3), 651-674.

Hübl, J., Bunza, G., Hafner, K., & Klaus, W. (2003). *Stummer Zeugen Katalog.* Projektteam ETAlp (Bundesministerium fuer Land-und Forstwirtschaft, Umwelt und Wasserwirtschaft), Kompendium zu ETAlp-Erosion, Transport in alpinen Systemen, 48.(in German)

Hübl, J., Strauss, A., Holub, M., & Suda, J. (2005). *Structural mitigation measures.* na.

Hübl, J. (2006). *Einwirkungen von gravitativen Naturgefahren: Muren*, In: Inst. Für Ingenieurgeologie, TU-Wien (Hrsg.), Die Bemessungssicherheit von Schutzbauwerken und gefährdeten Objekten im Zusammenhang mit gravitativen Naturgefahren, 4.12.2006, Wien (in German)

Hübl, J. (2007). *Skriptum Wildbach- und Lawinenverbauung*, Institut für Alpine Naturgefahren, Universität für Bodenkultur Wien (unveröffentlicht, in German)

Hübl, J., Suda, J., Proske, D., Kaitna, R., & Scheidl, C. (2009, September). *Debris flow impact estimation.* In Proceedings of the 11th international symposium on water management and hydraulic engineering, Ohrid, Macedonia (pp. 1-5).

Hungr, O., Morgan, G. C., & Kellerhals, R. (1984). *Quantitative analysis of debris torrent hazards for design of remedial measures.* Canadian Geotechnical Journal, 21(4), 663-677.

Hungr, O., Leroueil, S., & Picarelli, L. (2014). *The Varnes classification of landslide types, an update.* Landslides, 11(2), 167-194.

Innes, J. L. (1983). *Lichenometric dating of debris flow deposits in the Scottish Highlands.* Earth Surface Processes and Landforms, 8(6), 579-588.

Iverson, R. M. (1997). *The physics of debris flows.* Reviews of geophysics, 35(3), 245-296.

Iverson, R. M. (2014). *Debris flows: behaviour and hazard assessment.* Geology today, 30(1), 15-20.

Kaiser, H. F. (1961). *A Note on Guttman's Lower Bound for the Number of Common Factors.* British Journal of Statistical Psychology 14: 1–2.

Kaitna, R., (2006). *Debris-flow experiments in a rotating drum.* PhD thesis, Universität für Bodenkultur, Vienna, 170 pp.

Kaitna, R., & Rickenmann, D. (2007, September). *Flow of different material mixtures in a rotating drum.* In Debris-flow hazards mitigation, Fourth International DFHM Conference: mechanics, prediction and assessment (pp. 10-13).

Kaitna, R., & Huebl, J. (2013). *Silent witnesses for torrential processes.* In Dating torrential processes on fans and

cones (pp. 111-130). Springer, Dordrecht.

Knighton, D. (1998). *Fluvial forms and processes*: London. Edward Arnold.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). New York: Springer.

Lappi, E. (2008). *L'epopea dei grandi lavori idroelettrici in Giudicarie nell'archivio fotografico di Dante Ongari*. Società degli Alpinisti Tridentini, Trento. (in Italian)

Legge Provinciale 27 Maggio 2008, N. 5. *Approvazione del nuovo piano urbanistico provinciale*. PAT (in Italian)

Legge Provinciale 1 Luglio 2011, N. 9. *Disciplina delle attività di protezione civile in provincia di Trento*. PAT (in Italian)

Major, J. J. (1997). *Depositional processes in large-scale debris-flow experiments*. The Journal of Geology, 105(3), 345-366.

Marchi L., Pasuto A., & Tecca P.R. (1993). *Flow processes on alluvial fans in the Eastern Italian Alps*. Z. Geomorph. N. F., 37, 447-458.

Marchi, L., & D'Agostino, V. (2004). *Estimation of debris flow magnitude in the Eastern Italian Alps*. Earth Surface Processes and Landforms, 29(2), 207-220.

Marchi, L., & Cavalli, M. (2007). *Procedures for the documentation of historical debris flows: application to the Chieppena Torrent (Italian Alps)*. Environmental management, 40(3), 493-503.

May, C. L., & Gresswell, R. E. (2004). *Spatial and temporal patterns of debris-flow deposition in the Oregon Coast Range, USA*. Geomorphology, 57(3-4), 135-149.

Melton, M. A. (1965). *The geomorphic and paleoclimatic significance of alluvial deposits in southern Arizona*. The Journal of Geology, 73(1), 1-38.

Mizuyama, T., and Uehara, S. 1981. *Debris flow in steep slope channel curves*. Japanese Journal of Civil Engineering, 23: 243–248.

ON (2007): ONR 24800 – *Schutzbauwerke der Wildbachverbauung – Begriffsbestimmungen und Klassifizierungen*, (in German)

Pascal, H., Michel, J., & Eric, B. (2008). *Debris flow susceptibility mapping at a regional scale*. In Proceedings of the 4th Canadian Conference on Geohazards: From Causes to Management. Presse de l'Université Laval, Québec.

Phillips, C. J., & Davies, T. R. (1991). *Determining rheological parameters of debris flow material*. Geomorphology, 4(2), 101-110.

Pierson, T. C. (1980). *Erosion and deposition by debris flows at Mt Thomas, north Canterbury, New Zealand*. Earth Surface Processes and Landforms, 5(3), 227-247.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.. ISBN: 1-55860-238-0

Revellino, P., Guadagno, F. M., & Hungr, O. (2008). *Morphological methods and dynamic modelling in landslide*

hazard assessment of the Campania Apennine carbonate slope. Landslides, 5(1), 59-70.

Rimböck A., Barben M., Gruber H., Hübl, J., Moser M., Rickenmann D., Schober S., Schwaller G. (2013): *OptiMeth – Beitrag zur optimalen Anwendung von Methoden zur Beschreibung von Wildbachprozessen.* Internationale Forschungsgesellschaft INTERPRAEVENT, Schriftenreihe 1, Handbuch 3, Klagenfurt (in German)

Salzer (1886): *Über den Stand der Wildbachverbauung in Österreich* (in German)

Santangelo, N., Daunis i Estadella, J., Di Crescenzo, G., Di Donato, V., Faillace, P. I., Martín Fernández, J. A., ... & Scorpio, V. (2012). *Topographic predictors of susceptibility to alluvial fan flooding, Southern Apennines.* Earth Surface Processes and Landforms, 37(8), 803-817.

Santi, P. M., Pyles, D. R., & Pederson, C. A. (2017). *Debris Flow Avulsion.* International Journal of Erosion Control Engineering, 10(1), 67-73.

Scheidl, C., McArdell, B. W., & Rickenmann, D. (2014). *Debris-flow velocities and superelevation in a curved laboratory channel.* Canadian Geotechnical Journal, 52(3), 305-317.

Schneuwly-Bollschweiler, M., Stoffel, M., & Rudolf-Miklau, F. (Eds.). (2012). *Dating torrential processes on fans and cones: methods and their application for hazard and risk assessment* (Vol. 47). Springer Science & Business Media.

Schneuwly-Bollschweiler, M., & Stoffel, M. (2012). *Hydrometeorological triggers of periglacial debris flows in the Zermatt valley (Switzerland) since 1864.* Journal of Geophysical Research: Earth Surface, 117(F2).

Scorpio, V., Santangelo, N., & Santo, A. (2016). *Multiscale map analysis in alluvial fan flood-prone areas.* Journal of Maps, 12(2), 382-393.

Selby, M. J. (1993). *Hillslope Materials and Processes.* Oxford University Press.

Stiny, J. (1910): *Die Muren*, Wagner, Innsbruck (in German)

Stiny, J. (1931): *Geologische Grundlagen der Verbauung der Geschiebeherde in Gewässern*, Springer Verlag, Wien (in German)

Strasser, H., & Weber, C. (1999). *On the asymptotic theory of permutation statistics.* 27. SFB Adaptive Information Systems and Modelling in Economics and Management Science, WU Vienna University of Economics and Business, Vienna.

Suda, J., & Rudolf-Miklau, F. (Eds.). (2011). *Bauen und Naturgefahren: Handbuch für konstruktiven Gebäudeschutz.* Springer-Verlag. (in German)

Takahashi, T. (1991). *Debris flow* (IAHR Monograph Series), Balkema, Rotterdam, The Netherlands, 165.

Tomasi, G. (1962). *Origine, distribuzione, catasto e bibliografia dei laghi del Trentino.* Museo di storia naturale. (in Italian)

Tomasi, G. (1997). *Il territorio trentino-tirolese nell'antica cartografia/Trentiner und Südtirol Landschaft auf alten Landkarten.* Priuli & Verlucca, Ivrea. (in Italian and German)

VanDine, D. F. (1985). *Debris flows and debris torrents in the southern Canadian Cordillera.* Canadian Geotechnical Journal, 22(1), 44-68.

VanDine, D. F. (1996). *Debris Flow Control Structures for Forest Engineering*. Working paper 22/1996, Ministry of Forest Research Program, British Columbia

Wan, Z., Wang, Z., & Julien, P. Y. (1994). *Hyperconcentrated flow*. Journal of Hydraulic Engineering, 120(10), 1234-1234.

Wang, Q., Kong, Y., Zhang, W., Chen, J., Xu, P., Li, H., ... & Zhu, Y. (2016). *Regional debris flow susceptibility analysis based on principal component analysis and self-organizing map: a case study in Southwest China*. Arabian Journal of Geosciences, 9(18), 718.

Weber, A. (1964): *Wildbachverbauung*, Kapitel XIII, in: "Taschenbuch landwirtschaftlicher Wasserbau" Uhden, Otto (Hrsg.), Franckh'scher Verlagsbuchhandlung, Stuttgart, 483 – 528 (in German)

Welsh, A., & Davies, T. (2011). *Identification of alluvial fans susceptible to debris-flow hazards*. Landslides, 8(2), 183-194.

Wilkinson, L. (2006). *The grammar of graphics*. Springer Science & Business Media.

Willgoose, G., & Hancock, G. (1998). *Revisiting the hypsometric curve as an indicator of form and process in transport-limited catchment*. Earth Surface Processes and Landforms: The Journal of the British Geomorphological Group, 23(7), 611-623.

Zimmermann, M., Mani, P., Gamma, P., Gsteiger, P. Heiniger, O. & Hunziker, G. (1997). *Murgefahr und Klimaänderung – in GIS-basierter Ansatz*. Zürich. (in German)

Zollinger, F. (1985, September). *Debris detention basins in the European Alps*. In Proc. Int. Symp. Erosion, debris flow and disaster prevention, Tsukuba, Japan (Vol. 1, pp. 433-438).

Zollinger, D. (1986). *Wildbachsperren: Klassifikationen und Definitionen*. Wildbach-und Lawinenverbau, 50(103), 70-93 (in German)

Sitography

Accordo De Gasperi - Gruber. https://it.wikipedia.org/wiki/Accordo_De_Gasperi-Gruber

Agricultural service Trentino. http://www.trentinoagricoltura.it/

Autonomous Province of Trento. http://www.provincia.tn.it/

Clima Trentino. http://www.climatrentino.it/

EGTC Tyrol-South Tyrol-Trentino. http://www.europaregion.info/

Forest and wildlife service PAT. https://forestefauna.provincia.tn.it/

GNU. http://www.gnu.org/

Language minorities PAT. http://www.minoranzelinguistiche.provincia.tn.it/

Logistic regression (Sayad S.). https://www.saedsayad.com/logistic_regression.htm

Partykit::ctree. https://www.rdocumentation.org/packages/partykit/versions/1.2-0/topics/ctree

PCA, explained variance (Cheplyaka R.). https://ro-che.info/articles/2017-12-11-pca-explained-variance

QGIS. https://www.qgis.org/en/site/

Reservoir database, Progetto Dighe. https://www.progettodighe.it/

Reservoir office PAT. http://www.floods.it/public/homepage.php

RStudio. https://www.rstudio.com/

Seismic classification PAT. http://www.protezionecivile.tn.it/territorio/Sismologia/

Servizio Bacini Montani Trento. http://www.bacinimontani.provincia.tn.it/

Statistic service PAT. http://www.statistica.provincia.tn.it/

The Comprehensive R Archive Network. https://cran.r-project.org/

The R Project for Statistical Computing. https://www.r-project.org/

Trentino. https://en.wikipedia.org/wiki/Trentino

Trentino OPENdata. http://dati.trentino.it/

Weather forecast service, Meteotrentino. http://www.meteotrentino.it/

ANNEX: R-code

```
#@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
# CODE PCA AND LOGISTIC REGRESSION - Torrential behaviour, Province of Trento
# Author: Gottardi Valentino
# Date: January 2020
#@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@


# ---------------------------------------------------------------------------------- #
# STEP 1: R PACKAGES AND RAW DATA IMPORT --------------------------------------------- #
# ---------------------------------------------------------------------------------- #
# uploading the utilized packages
# inporting the used shapefiles with informations about drainage basins and alluvial fans
# reprojecting the shapefiles
# ---------------------------------------------------------------------------------- #

# needed packages
library(proj4)
library(rgdal)
library(corrplot)
library(caret)

# raw data from Shapefiles
basins <- readOGR(dsn = "/home/valentino/Documenti/MASTERARBEIT/Shape",
              layer = "watershedMOD",
              p4s = NULL)
fans <- readOGR(dsn = "/home/valentino/Documenti/MASTERARBEIT/Shape",
              layer = "debris_fluvial_events",
              p4s = NULL)

# set projection
basins <- spTransform(basins, CRS("+init=epsg:25832"))
fans <- spTransform(fans, CRS("+init=epsg:25832"))



# ---------------------------------------------------------------------------------- #
# STEP 2: DATA PREPROCESSING -------------------------------------------------------- #
# ---------------------------------------------------------------------------------- #
# Unnecessary features are deleted and the correct data types are set
# log-transformation of areas and volumes
```

```r
# data is reduced to those catchments with at least one event
# ------------------------------------------------------------------------------ #

# use only dataframes and not shapefiles [1697 basins]
dBase <- merge(basins@data, fans@data, by.x = "classid", by.y = "classid")

# reduce features and set correct types
dBase <- dBase[, -c(44:39, 37:29, 27:25)]
dBase$nFluvial <- as.numeric(as.character(dBase$nFluvial))
dBase$nDebris <- as.numeric(as.character(dBase$nDebris))
dBase$danno_pote <- as.numeric(as.character(dBase$danno_pote))

# log (base = 10) transformations
dBase$barea <- log10(dBase$barea)
dBase$carea <- log10(dBase$carea)
dBase$cvol <- log10(dBase$cvol)
dBase$blungasta <- log10(dBase$blungasta)
dBase$bffactor <- log10(dBase$bffactor)

# reduce to those with at least one event [159 basins]
idActive <- which((dBase$nDebris + dBase$nFluvial) > 0)
dBaseActive <- dBase[idActive, ]



# ------------------------------------------------------------------------------ #
# STEP 3: NUMBER OF EVENTS TABLE ----------------------------------------------- #
# ------------------------------------------------------------------------------ #
# graphical display of events distribution, basins with at least one event
# rows: number of non-debris flow events (bedload transport and clear water discharge)
# columns: number of debris flow events (debris flow and debris flood)

# E.G. All catchments that are in the first row had 0 non-debris flow events
# but 1, 2, 3, 4, 5, or 6 debris flow events.
# All catchments in the first column had 0 debris flows
# but 1, 2, 3, 4, or 5 non-debris flow events.

# Removing all mixed type basins we will loose 20 observations
# 139 single process catchments remains.
# ------------------------------------------------------------------------------ #

table(dBaseActive$nFluvial, dBaseActive$nDebris)
table(dBase$nFluvial, dBase$nDebris)

sum(dBaseActive$nFluvial >= 1)
sum(dBaseActive$nDebris >= 1)

sum(dBaseActive$nFluvial >= 1 & dBaseActive$nDebris == 0)
sum(dBaseActive$nFluvial == 0 & dBaseActive$nDebris >= 0)


```

```
92
93  # ------------------------------------------------------------------------------ #
94  # STEP 4: DEBRIS FLOW VS. NON DEBRIS FLOW BASINS ------------------------------- #
95  # ------------------------------------------------------------------------------ #
96  # the feature DF1 is added, getting the following values for:
97  # [ 1 ] at least one DEBRIS FLOW but NONE non-debris flow event
98  # OR
99  # [ 2 ] at least one non-debris flow but NONE DEBRIS FLOW event
100 # OR
101 # [ NA ] otherwise.
102 # ------------------------------------------------------------------------------ #
103
104 n <- nrow(dBaseActive)
105 dBaseActive$DF1 <- rep(NA, n)
106
107 for(i in 1:n){
108 if(dBaseActive$nDebris[i] >= 1 & dBaseActive$nFluvial[i] ==0){dBaseActive$DF1[i] <- 1}
109 else if(dBaseActive$nDebris[i] == 0 & dBaseActive$nFluvial[i] >=1){dBaseActive$DF1[i] <- 0}
110 else{dBaseActive$DF1[i] <- NA}}
111
112
113
114 # ------------------------------------------------------------------------------ #
115 # STEP 5: IMPUTE "NA" VALUES --------------------------------------------------- #
116 # ------------------------------------------------------------------------------ #
117 # One of the basins shows ‘NA‘ values in the features blungasta, bslo_asta and bffactor.
118 # The imputation of those values is performed by selection of a basin
119 # with the closest euclidian distance regarding all other features.
120 # ------------------------------------------------------------------------------ #
121
122 # choose only numeric features
123 colidx <- seq(3, (ncol(dBaseActive)-4), 1)
124 colidx <- colidx[colidx !=15 & colidx != 24 & colidx != 23 & colidx != 22]
125
126 # reduce data set and standardize values
127 dBaseDistance <- dBaseActive[, colidx]
128 dBaseDistance <- scale(dBaseDistance)
129
130 # get distance matrix and that basin which has the lowest distance to the impute basin
131 rowID <- rownames(dBaseActive[is.na(dBaseActive$bslo_asta), ]) # ID of the basin to impute
132 dmat <- as.matrix(dist(dBaseDistance, upper = TRUE)) # distance between basins
133 bidx <- which(rownames(dmat) == rowID) # get distance matrix row of the impute basin
134 nbidx <- names(sort(dmat[bidx, ])[2]) # get row ID of nearest basin
135
136 # get nearest basin data and impute
137 idxnb <- which(rownames(dBaseActive) == nbidx) # nearest basin
138 idxib <- which(rownames(dBaseActive) == rowID) # impute basin
139
140 dBaseActive[idxib, c(22:24)] <- dBaseActive[idxnb, c(22:24)] # impute
141
142
```

```
143
144  # ------------------------------------------------------------------------------ #
145  # STEP 6: FEATURE SELECTION ---------------------------------------------------- #
146  # ------------------------------------------------------------------------------ #
147  # Reduction of the dataset to the most meaningful features.
148  # These were chossen by their seperation ability in the one-dimensional case:
149        # manteining only the features with different boxplots between classes.
150  # The reduced dataset is used to check for co-correlation.
151  # Features with high co-correlation are then excluded.
152  # ------------------------------------------------------------------------------ #
153
154  # boxplots used to plot the univariate distributions all features are excluded which
155  # show a high overlap of their feature values
156  colidx <- seq(3, (ncol(dBaseActive)-4), 1)
157  # delete factor because not plotable as a boxplot ("bstadio")
158  colidx <- colidx[colidx !=15]
159
160  # build boxplots
161  par(mfrow=c(1,4))
162  for(i in 1:length(colidx)){
163  y <- (dBaseActive[,colidx[i]]-min(dBaseActive[,colidx[i]]))/
164       (diff(range(dBaseActive[,colidx[i]])))
165  y <- scale(dBaseActive[,colidx[i]])
166  boxplot(y~dBaseActive$DF1,
167      main = paste(i),
168      notch = TRUE,
169      col = c("blue", "red"),
170      names = c("non-DF", "DF"))
171  }
172  idx <- c(22, 21, 16, 13, 12, 10, 9, 7, 8) # excluding features
173  dBaseSel <- dBaseActive[, -colidx[idx]]
174
175
176  # Based on high correlation (start from the 13 choosen variables)
177  par(mfrow=c(1,1))
178  corrplot(cor(dBaseSel[, c(3:9, 11:16)]), method = "number")
179  corrplot(cor(dBaseSel[, c(5:9, 11:16)]), method = "number") # sequential feature reduction
180  corrplot(cor(dBaseSel[, c(7:9, 11:16)]), method = "number") # sequential feature reduction
181  dBaseSel <- dBaseSel[, -c(3:6)]
182
183
184
185  # ------------------------------------------------------------------------------ #
186  # STEP 7: PREPARING THE MODEL -------------------------------------------------- #
187  # PRINCIPAL COMPONENT ANALYSIS ------------------------------------------------- #
188  # ------------------------------------------------------------------------------ #
189  # The PCA is used to overcome the problem of high collinearity between the predictors.
190  # ------------------------------------------------------------------------------ #
191
192  ## PCA using Definition 1
193
```

```r
# Whole predictor set
dBaseActive <- dBaseActive[!is.na(dBaseActive$DF1), ] # only basins with DF1 = 1 or DF1 = 0
dBaseMod <- dBaseActive[, c( -2, -15, -26, -27, -28)] # modelling dataset without ID's etc.
dBaseMod$DF1 <- as.factor(dBaseMod$DF1) # add observed behaviour vector


# Conducting the Principal Component Analysis
PCAmodel <- prcomp(dBaseMod[,-c(1,23,24)], center = TRUE, scale = TRUE) # Estimate PCs
summary(PCAmodel)
pcaData <- as.data.frame(predict(PCAmodel)) # predict the pc's
pcaData$DF1 <- dBaseMod$DF1 # add class vectors
pcaData$classid <- dBaseMod$classid
dBaseModPCA <- pcaData




# ------------------------------------------------------------------------------ #
# STEP 8: APPLICATION OF OTHER METHODS ---------------------------------------- #
# ------------------------------------------------------------------------------ #

## Using only basin parameters
# From D'Agostino (1996) - Melton Ruggedness Number
dBaseMod$Dago1 <- as.factor(ifelse(dBaseMod$bmelton > 0.5, 1, 0))
pcaData$Dago1 <- dBaseMod$Dago1
# From DeScally and Owens (2004) - Melton Ruggedness Number
dBaseMod$DeSc1 <- as.factor(ifelse(dBaseMod$bmelton > 0.75, 1, 0))
pcaData$DeSc1 <- dBaseMod$DeSc1

## Using both basin and fan parameters
# From D'Agostino (1996) - Fan's medium slope against Melton Ruggedness Number
dBaseMod$Dago2 <- as.factor(ifelse(dBaseMod$cslo_med > (dBaseMod$bmelton * 14), 1, 0))
pcaData$Dago2 <- dBaseMod$Dago2
# From DeScally and Owens (2004) - Melton Ruggrdness Number and Fan's mean slope
dBaseMod$DeSc2 <- as.factor(ifelse(dBaseMod$bmelton > 0.75 &
                           dBaseMod$cslo_med > 7.5, 1, 0))
pcaData$DeSc2 <- dBaseMod$DeSc2

dBaseModPCA <- pcaData

# ------------------------------------------------------------------------------ #
# STEP 9: FITTING THE MODEL --------------------------------------------------- #
# ------------------------------------------------------------------------------ #
# The model will be fitted using bootstraping in order to asses their performance
# based on Youdens J (Specificity + Sensitivity - 1).
# ------------------------------------------------------------------------------ #

library(e1071)
## Bootstrap using 6 Most Important PCA's (82% of Variance explained)

dBaseFit <- dBaseModPCA # generic name of data to build the model

```

```r
# Bootstrap Settings
n <- nrow(dBaseFit)
m <- 1000 # number of bootstrap samples
Specificity <- rep(NA, m)
Sensitivity <- rep(NA, m)
youdensJ <- rep(NA, m)

# Resampling
set.seed(800)
for(i in 1:m){
idx <- sample(1:n, n, replace = TRUE) # random selection of index (row number)
datTrain <- dBaseFit[idx, ]
datTest <- dBaseFit[-idx, ]

# Fit the model (6 most important PCs, resampled watersheds)
ModelTRAIN <- glm(DF1 ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6,
                  data = datTrain,
                  family = "binomial")

# Predict test set and estimate
testProb <- predict(ModelTRAIN, newdata = datTest, type = "response")
testPred <- ifelse(testProb > 0.5, 1, 0)

# Evaluate Prediction
confMat <- confusionMatrix(data = factor(testPred, levels = 0:1),
                     reference = factor(datTest$DF1, levels = 0:1),
                     positive = "1")
Specificity[i] <- confMat$byClass[2]
Sensitivity[i] <- confMat$byClass[1]
youdensJ[i] <- Specificity[i] + Sensitivity[i] - 1
}

datTest$Prob <- testProb
datTest$Pred <- testPred


# ------------------------------------------------------------------------------ #
# STEP 10: RECEIVER OPERATING CHARACTERISTIC ---------------------------------- #
# ------------------------------------------------------------------------------ #

library(pROC)

ModelROC <- roc(response = datTest$DF1, predictor = datTest$Prob) # create the ROC

par(pty = "s") # parameter for square plotting
roc(response = datTest$DF1, predictor = datTest$Prob, # plotting the ROC
       plot = TRUE,
       legacy.axes = TRUE,
       percent = TRUE,
       ci = TRUE,
       print.auc = TRUE, auc.polygon = TRUE, auc.polygon.col ="#377eb822", print.auc.x =60,
```

```
296           col = "#377eb8", lwd = 3,
297           main = "ROC: PCA and Logit Model")
298
299
300   # alternative cutoff - optimal threshold
301   cutoff <- coords(roc = ModelROC,
302             x = "best",
303             best.method = "youden",
304             transpose = TRUE)
305   cutoff
306
307
308   # use new cutoff
309   # Predict test set and estimate
310   testPredNew <- ifelse(testProb > 0.5222838, 1, 0)
311
312   # Evaluate Prediction
313   confusionMatrix(data = factor(testPredNew, levels = 0:1),
314           reference = factor(datTest$DF1, levels = 0:1),
315           positive = "1")
316
317
318   # ------------------------------------------------------------------------------ #
319   # STEP 11: CONFUSION MATRIX ---------------------------------------------------- #
320   # ------------------------------------------------------------------------------ #
321
322   # MODEL: confusion matrix for PCA and LR model - (only on test subset)
323   confusionMatrix(data = factor(datTest$Pred, levels = 0:1),
324            reference = factor(datTest$DF1, levels = 0:1),
325            positive = "1")
326
327   # LITERATURE: confusion matrix - (only on test subset)
328   # D'Agostino (1996) - Dago1
329   confusionMatrix(data = factor(datTest$Dago1, levels = 0:1),
330            reference = factor(datTest$DF1, levels = 0:1),
331            positive = "1")
332   # DeScally and Owens (2004) - DeSc1
333   confusionMatrix(data = factor(datTest$DeSc1, levels = 0:1),
334            reference = factor(datTest$DF1, levels = 0:1),
335            positive = "1")
336   # D'Agostino (1996) - Dago2
337   confusionMatrix(data = factor(datTest$Dago2, levels = 0:1),
338            reference = factor(datTest$DF1, levels = 0:1),
339            positive = "1")
340   # DeScally and Owens (2004) - DeSc2
341   confusionMatrix(data = factor(datTest$DeSc2, levels = 0:1),
342            reference = factor(datTest$DF1, levels = 0:1),
343            positive = "1")
344
345
346
```

```r
# @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@ #
# ---------------------------- APPLICATION OF THE MODEL ----------------------------- #
# @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@ #
# ----------------------------------------------------------------------------------- #
# STEP 12: FIT THE MODEL --- [WHOLE PREDICTOR SET] ---------------------------------- #
# ----------------------------------------------------------------------------------- #

# use already performed PCA (6 most important PCs, original selected watersheds)
ModelALL <- glm(DF1 ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6,
         data = dBaseModPCA,
         family = "binomial")

# ----------------------------------------------------------------------------------- #
# STEP 13: INPUT "NA" VALUES -------------------------------------------------------- #
# ----------------------------------------------------------------------------------- #

## add "bslo_asta" NA values
# choose only numeric features
colidx <- seq(3, (ncol(dBase)-4), 1)
colidx <- colidx[colidx !=15 & colidx != 24 & colidx != 23 & colidx != 22]

# reduce data set and standartize values
dBaseDist <- dBase[, colidx]
dBaseDist <- scale(dBaseDist)

# get distance matrix and that basin which has the lowest distance to the impute basin
rowID1 <- rownames(dBase[is.na(dBase$bslo_asta), ]) # ID of the basin to impute
dmat1 <- as.matrix(dist(dBaseDist, upper = TRUE)) # distance between basins
bidx1 <- which(rownames(dmat1) %in% rowID1) # get distance matrix row of the impute basin

nbidx1 <- rep(NA, length(rowID1))
for(i in 1:length(rowID1)){
nbidx1[i] <- names(sort(dmat1[bidx1[i], ])[2]) # get row ID of nearest basin
}

# get nearest basin data and impute
for(i in 1:length(rowID1)){
idxnb1 <- which(rownames(dBase) == nbidx1[i]) # nearest basin
idxib1 <- which(rownames(dBase) == rowID1[i]) # impute basin

dBase[idxib1, c(22:24)] <- dBase[idxnb1, c(22:24)] # impute
}

## add "cslo_med" and "cvol" NA values
# replace NaN with NA
dBase[dBase == "NaN"] <- NA

# get distance matrix and that basin which has the lowest distance to the impute basin
rowID2 <- rownames(dBase[is.na(dBase$cslo_med), ]) # ID of the basin to impute
dmat2 <- as.matrix(dist(dBaseDist, upper = TRUE)) # distance between basins
bidx2 <- which(rownames(dmat2) %in% rowID2) # get distance matrix row of the impute basin
```

```
398
399   nbidx2 <- rep(NA, length(rowID2))
400   for(i in 1:length(rowID2)){
401   nbidx2[i] <- names(sort(dmat2[bidx2[i], ])[2]) # get row ID of nearest basin
402   }
403
404   # get nearest basin data and impute
405   for(i in 1:length(rowID2)){
406   idxnb2 <- which(rownames(dBase) == nbidx2[i]) # nearest basin
407   idxib2 <- which(rownames(dBase) == rowID2[i]) # impute basin
408
409   dBase[idxib2, c(18:19)] <- dBase[idxnb2, c(18:19)] # impute
410   }
411
412
413
414   # ------------------------------------------------------------------------------ #
415   # STEP 14: PREDICT THE PROBABILITY OF DEBRIS FLOW ---------------------------------- #
416   # ------------------------------------------------------------------------------ #
417
418   # predict the debris flow probability for the entire drainage basin set
419   # Modelling data set without ID's etc.
420   dBasePred <- dBase[, c(-1, -2, -15, -26, -27, -28)]
421
422   # Conducting the PCA (use the model fitted on selected basins as reference)
423   dBasePred <- as.data.frame(predict(PCAmodel,
424                              newdata = dBasePred))
425
426   # Predict with the model
427   dfProb <- predict(object = ModelALL,
428                 newdata = dBasePred,
429                 type = "response")
430
431   # Add prediction vector of debris flow probability to the database
432   basins@data$DFprob <- dfProb
433
434   # Add class prediction vector of debris flow probability to the database
435   basins@data$DFclass <- ifelse(basins@data$DFprob > 0.5222838, 1, 0)
436
437
438
439   # ------------------------------------------------------------------------------ #
440   # STEP 15: PLOT THE MAP ------------------------------------------------------- #
441   # ------------------------------------------------------------------------------ #
442   # plotting the basin's polygons with color and class options
443   # ------------------------------------------------------------------------------ #
444
445   rbPal <- colorRampPalette(c('red','blue'))
446   # Probability map
447   basins@data$Col <- rbPal(10)[as.numeric(cut(basins@data$DFprob, seq(0, 1, by = 0.1)))]
448   plot(basins, col = basins@data$Col)
```

```
449
450
451
452   # ------------------------------------------------------------------------------ #
453   # STEP 16: SAVE AND EXPORT SHAPEFILE ------------------------------------------ #
454   # ------------------------------------------------------------------------------ #
455   # export of the shapefile for the use in GIS environment
456   # ------------------------------------------------------------------------------ #
457
458   # folder creation, dataset to shapefile
459   dir.create("ShapeExport")
460
461   # export shapefile
462   writeOGR(obj = basins,
463          dsn = "ShapeExport",
464          layer = "ModelOutput",
465          driver = "ESRI␣Shapefile")
466
467
468
469   # ------------------------------------------------------------------------------ #
470   # END OF THE CODE ------------------------------------------------------------- #
471   # ------------------------------------------------------------------------------ #
```