Vrije Universiteit Amsterdam

Faculty of Science

Risk map development for a snow avalanche hazard

Master project Business Analytics

GRADUATION SUPERVISOR:
prof. Dr. Sandjai Bhulai
s.bhulai@vu.nl

EXTERNAL SUPERVISOR:
Dr. Daphne van Leeuwen
d.vanleeuwen@brightcape.nl

SECOND READER:
Dr. Wan Fokkink
w.j.fokkink@vu.nl

AUTHOR:
Ieva Brantevica
i.brantevica@brightcape.nl

July 2019

Preface

This paper is written as a compulsory part of the Master's degree in Business Analytics at Vrije Universiteit Amsterdam and serves as a graduation project. The research has been developed during a 6 months long internship in a data consultancy company called Bright Cape. Bright Cape has an expert knowledge in four different areas, but the internship focused more only on one - an innovation project. International partners were working together on this project to create a drone platform that can support Search and Rescue teams during their operations in order to rescue the avalanche victim faster. As a part of this project - a predictive model was created to indicate the likelihood of finding an avalanche victim in the French Alps.

First of all, I want to thank Daphne van Leeuwen, my external supervisor, for the support and guidance during this internship period. Especially, for the opportunity to travel to Rennes (France) and join the face-to-face meeting with partners from Italy and France. I also want to thank my graduation supervisor Sandjai Bhulai for the regular bi-weekly meetings, the moral support and the strong theoretical support in the field of geospatial modelling. I would like to thank Wan Fokkink for accepting to be the second reader of this report in such a short notice. In addition, I want to thank all my colleagues from Bright Cape for the help and the good time spent together. I am truly happy to have been able to experience and be part of a development of such an innovative product. Finally, a special thanks to my family for their loving support during my studies abroad.

Ieva Brantevica
July 2019

Executive summary

Every year, more than 250 people worldwide die from snow avalanches and the number of incidents are growing every year since 1990. When a person is buried underneath the snow, the survival rate of a victim drops quickly after 15 minutes since burial. The time is crucial when saving an avalanche victim. The mountain Search and Rescue teams are localising the victim by their own experience and gut feeling, sometimes using trained search dogs and probe lines. This requests a lot of manpower and time, while the latter is crucial. In addition, helicopters are also deployed when necessary, but it is very costly.

There is a need for tools that can help to reduce this time. The host organization Bright Cape collaborates with University of Rennes 1 and Research institute FBK on an EIT Digital partly subsidised innovation project to make a flexible autonomous drone platform for Search and Rescue (SAR) operations. This report summarizes a part of Bright Cape's contribution to the project - a development of an avalanche risk map.

Therefore, the research question of this report is as follows:
How accurately can we predict the location of an avalanche/victim caught in an avalanche, based on the historical and available open source data?

A literature review of avalanche contributing factors have been performed to understand and acquire the necessary data sets for modelling from open sources. Preliminary analysis have been performed and data from different sources have been processed. A predictive model has been developed that can construct a risk map of the area of interest. The map represents the likelihood of an avalanche victim to be located in the area - a grid cell.

Three predictive models - Generalized linear model, Generalized additive model, Autoencoder, and 3 evaluation methods - accuracy, Matthews correlation coefficient and Fraction skill score were considered to train the risk model on historical avalanche victim data for time till 01-01-2017, and tested on the victim data from the past 2 years. GAM model with geospatial factors elevation, distance to a downhill slope, road, path and slope steepness were indicated as significant. The Fraction skill score of value 0.028 for a moving window of 7 x 7 cells was gained.

This is greater than GLM and Autoencoder performance. The value close to 0 indicates that the model still struggles to recognize the avalanches for the test set.

The model can be improved by including the dynamic avalanche factors like properties of snow pack and meteorological conditions. Using the distances of geospatial factors as numerical variables instead of ordinal could leave to a different model considerations and results. The model performance regarding a smaller area of interest and different sizes of grid cells could be investigated in the future.

Contents

Chapter 1

Introduction

1.1 Host organization and project

Bright Cape is a small & big data consultancy firm with offices in Eindhoven and Amsterdam active throughout Europe. Small & big data might sound somewhat confusing but what the company means by this slogan is that there is not always a need for big data to extract insights and to make decisions - small data can be enough for business improvements.

Bright Cape has expert knowledge in four different areas that help customers achieve their data driven goals. These areas are:

- Analytics & Applied Data Science,
- Data Driven Experience Design,
- Process Mining,
- Innovative Products.

Bright Cape helps customers by extracting value out of their data, embedding solutions into customer's processes & governance and educating them in data analytics and their created solutions. Through Data Analytics & Science, Data Driven Experiences (DDEX), and Process Mining, Bright Cape helps companies increase revenue, diminish costs and increase process efficiency.

Additionally, the company focuses on various Europe-wide innovation projects. This 6 months long graduate internship was focused on one of these innovation projects - UAV Retina. The project is partly subsidised by EIT Digital, an accelerator for innovative projects, and various international parties are collaborating to build one product - a flexible autonomous drone

platform for Search and Rescue (SAR) operations. The parties are Université de Rennes 1 from France, a research institute FBK from Italy, and Bright Cape from the Netherlands. The university in France provides the hardware for the product - architecture of the platform and the drone itself with the necessary sensors. FBK works on the mission planning (route optimization) and image recognition module. Bright Cape is responsible for the user experience research and development of the risk map.

The internship was supervised by Daphne van Leeuwen. Daphne is an alumna of the VU Amsterdam with a Master's degree in Business Mathematics and Computer Science. After graduation she continued studying as a PhD student at Centrum Wiskunde en Informatica (CWI) where she immersed into the topic of queueing theory to model road traffic congestion.

1.2 Problem statement

Every year more than 250 people worldwide die from snow avalanches and the number of incidents is growing every year since 1990. The majority of global avalanche incidents is initiated by a human [5]. This means that people often get carried away in avalanches. When a person is buried underneath the snow, time is crucial as most of the victims die due to lack of oxygen [27]. Within the first 15 minutes the chance of survival is 92%, after 30 minutes the chance of survival falls below 50%.

Search and Rescue (SaR) team typically comes into action when a phone call is received from people who have noticed an avalanche. Sometimes it is not known whether there are any victims in the incident. During the call, a description of the area of interest is given. Sometimes the area is specified accurately, but quite often the SaR team has to decide on the area they will investigate first. It is done by their own historical experience and gut feeling. They also use dogs that are trained to recognize human odor rising from the snow and also they often stick probe lines in the snow in order to find out where the victim is. However, this requires manpower and time while time is crucial. In addition, helicopters are also deployed when necessary, but it is very costly.

1.3 Objective

An introduction of unmanned aerial vehicles (UAV) for search and rescue operations can assist in finding the victims quicker. Also, drones are more agile and cheaper than helicopters, making it attractive to introduce them in search and rescue operations. As soon as an avalanche report arrives, an autonomous drone can fly from a central point or a ground station to the location with the greatest likelihood of finding the victim there. This location is determined on the basis

of a risk map. The risk map is divided into grid cells, which is an area of 500 by 500 meters. When the description of the location from the call is more specific than the size of the grid cell, the risk map is not necessary and the drone can fly straight to the given location. Hence, the risk map will be used only when a wide area has to be searched, and this varies depending on the incidents.

The drone flies to the highest risk area and looks for a victim. Sensors like infrared camera and RECCO transceiver can be used with the drone to localise the victim without SaR members being in the area. If the victim is not found in the area of the grid cell, the drone will move to the next location, continuing until the victim gets localized or the SaR team arrives and takes over the operation.

1.4 Research description

The internship goal is to develop quantitative models and methods for predicting and visualizing the likelihood of the avalanche victims' location. This application will contain a model that is based on historical data and characteristics of the landscape acquired via open source. Through mathematical analysis of these two types, the risk map (which is based on a grid) that indicates the probability of each grid cell to contain a victim will be developed.

The research does not include the route optimization plan for the drone and the processing of the on-site data acquired by the sensors. It only focuses on the development of the search map using static geospatial factors and historical avalanches.

This whole paper is completely focused on the predictability of avalanches. As snow avalanches are natural disasters, some realistic expectations have to be set. Overall, the predicted values from the model outcome will indicate the probability of natural disasters, but it is not possible to develop a predictive model that gives an exact value for the riskiness of an avalanche. While it is possible to indicate where the likelihood of an avalanche occurrence is greater, it is impossible to be completely certain that if an avalanche does happen, its actual location is the one with the highest probability of risk.

To reach the goal of the research, the following research question is formulated:

How accurately can we predict the location of an avalanche/victim caught in an avalanche, based on the historical and available open source data?

This research question can be answered with the help of the following sub-questions:

- *What are the main variables that increase this likelihood and accuracy?*

- *How these environmental factors should be used?*
- *What are the recommended methods in this field of subject?*
- *What validation criteria apply best for this specific problem?*
- *Which risk model fits the current problem the best?*
- *How applicable are the results for the usage in the field?*

1.5 Structure of the report

The remaining report is organised in six chapters. In Chapter 2, the literature review is performed. It discusses the different types of avalanches, the contributing factors of avalanches, a background of geographic data and the methods behind processing the data used for the modeling, a background information of Search and Rescue operations and unmanned aerial vehicles. In Chapter 3, the historical avalanche and geospatial data is described with the findings from preliminary analysis. The reasoning behind the data preparation is explained. Chapter 4 contains the explanation of theoretical methods and evaluation measures used for the research. The results of the models and their performance metrics are summarized in Chapter 5, followed by the conclusion chapter in Chapter 6. Finally, Chapter 7 discusses and reflects on the research followed by giving recommendations for further research.

Chapter 2

Literature review

Avalanches have been studied for decades and there are many researches done in this domain. The factors that cause the avalanches hardly change over time, so the research findings do not become outdated fast. This chapter summarizes the available literature to explain what are avalanches, what causes different types of avalanches, what are the factors that contribute to the snow avalanche danger in mountainous areas, and what are search and rescue operations. An explanation will be provided on geographical data files and how these files are used in the project.

2.1 Avalanches and types of avalanches

Snow avalanches are snow masses that rapidly descend steep slopes. They can contain rocks, soil, vegetation, or ice [34]. The majority of the world-wide avalanche incidents are human-initiated. In the European Alps, an average of 103 people per year died in avalanches from 1970 to 2015 [36]. To better understand the domain it is important to look at different types of avalanches.

In many literature sources, avalanches are divided in two sub-types - dry and wet avalanches. As described in [5], 90% of all dry avalanches are triggered by the victims or someone in the victim's party by putting too much additional stress on the snow-pack. Usually wind is drifting the snow or a new fresh snow is loading the unstable snow-pack. The avalanche flow moves very fast (130 km/hr) and a dust cloud accompanies the avalanche.

Wet avalanches are caused by decreasing the strength of the snow-pack and they are difficult for people to trigger. Most of the wet avalanche accidents are from natural causes. The weather conditions like rain, prolonged melting by sun or very warm temperatures contribute to these natural hazards. The flow of wet avalanches is slower (10-65 km/hr) and moves concretely without a dust cloud.

Wet snow avalanches can be released as slab avalanches, loose snow or glide avalanches. Dry snow avalanches usually emerge from slab avalanches [5]. Let's look deeper on how these types differ from each other.

2.1.1 Slab avalanche

A "slab" is a compact snow layer that slides as a unit on the weaker layer of snow underneath. It means that the top layer of the snow is stronger than the layer it overlies. A slab can occur anywhere in the snow-pack but avalanche professionals usually think of a slab as the layer that slides off the slope to create the avalanche [4].

The bonds holding a slab in place typically fracture in a single section at 350 km/hr and it appears to shatter like a pane of glass. Usually it is about the size of half a football field, about 30-80 centimeters deep and reaches speeds of 30 km/hr within the first 3 seconds and quickly accelerates to around 130 km/hr after the first 6 seconds. Dry slab avalanches can lie patiently, teetering on the verge of catastrophe, sometimes for days to even months. The weak-layers beneath slabs are also extremely sensitive to the rate at which they are stressed. In other words, the rapid addition of the weight of a person can easily initiate the fracture on a slope that would not have avalanched otherwise. The crack often forms well above the victim, leaving little room for escape [3].

The slab avalanches are the most dangerous from all 3 types. They can be released in different ways - naturally, remotely by using explosives or triggered by a person. As stated in the introductory paragraph, 85 - 90% of worldwide avalanche incidents are triggered by a human and the majority of slab avalanches causes fatalities.

2.1.2 Loose snow or sluff avalanche

Loose snow sliding down a mountainside is called a loose snow avalanche. Small loose snow avalanches are called "sluffs" [3]. These avalanches usually occur either during or right after the snowfall and are released naturally. It may also occur if the temperature changes and a significant warming starts. The reason these type of avalanches are called loose snow avalanches is because they consist of cold, dry and powdery snow that has not been able to stick to the layer beneath it. The sluff avalanches require a steep slope - an angle of 40° or steeper. When the snow is wet and the slope is continuously steep, these avalanches can reach a considerable size as they spread out from a point of triggering and sweep along more and more snow.

A relatively small number (less than 10%) of avalanche fatalities can be tied with the loose snow avalanches. Most of the incidents and deaths occur in summer, when mountaineers in steep terrain are swept along and then fall. When a loose slow avalanche is triggered by a snow

sport participant, he is rarely buried because the snow slides down the slope away from him and usually releases only small snow masses [33]. Sometimes a sluff is considered as a sign of stability within the deeper snow because the new snow slides down without triggering deeper slabs.

2.1.3 Glide avalanche

Gliding avalanches are considered when not only part, but the entire snow-pack as one unit slides down the slope. This usually is a slow process that can occur over several days. Both slab and glide avalanches have a wide, well-defined fracture line, the difference is that glide avalanches have the whole snow-pack released. They occur because there is water underneath the snow-pack which lubricates the ground and "glides" down on a smooth substratum, typically consisting of flattened grass or slabs of rock. The steeper the slope, the more likely the snow is to slide [3]. Water can penetrate the basal layer in three different ways:

- from the ground - the ground is either still warm or wet during the first snowfall and the base layer of the snow-pack gets moist or absorbs the moisture from the ground;
- from solar radiation - the sun melted snow can seep through the entire snow-pack and concentrate between the ground and the snow-pack;
- from rain - the rain water penetrates through the snow-pack and concentrates between the ground and snow-pack [1].

As stated in [33], in very snowy winters the gliding avalanches are a major problem for roads. As these avalanches release naturally, they are difficult to trigger by winter-sport enthusiasts or with the use of explosives. Glide avalanches are hard to predict and pose a difficult challenge to the avalanche control teams at ski resort areas.

2.2 Avalanche contributing factors

An avalanche occurrence is a result of complex interactions between multiple contributing factors. According to the author of [34], there is still uncertainty over what timing and which factors influence the avalanche occurrence, but almost all research mentions terrain, snow-pack, and meteorological conditions. Meteorological conditions like temperature, wind, precipitation and snow-pack are dynamic, constantly changing and filled with uncertainty. However, terrain is static and constant over a long period of time.

On one hand, it is simple to recognize a terrain which could be a potential place for an avalanche. On the other hand, recognizing when this same terrain presents an actual avalanche danger can be very uncertain and complicated. When looking at the terrain of the avalanche slopes, the characteristics can be divided in sub-factors - elevation, aspect, steepness of the slope, land cover

and forest cover. Another important contributing factor is the proximity of people. In the next chapters the importance of these factors will be explained in more detail.

2.2.1 Elevation

Elevation is an important factor that has a direct influence on the meteorological conditions within a certain altitude and land cover. The higher the altitude, the more suitable meteorological conditions of temperature, wind and precipitation are for avalanche formation. That is, when the air temperature is lower, the precipitation amount increases and the wind speeds are higher. According to [15], the relationship between the starting zones of the avalanches and elevation are highly correlated. Depending on the altitude, the properties of the snow in the upper and lower parts of the mountain changes. This introduces a challenge for to properly assess the stability of the snow pack. [29]

2.2.2 Aspect

Another static terrain factor is the direction of the mountain slope, also known as the aspect. The snow pack is directly influenced by wind speed and exposure to the sun. The properties of this factor change depending on the latitude. The overview on how aspect and season influences the snow pack stability in the north hemisphere is depicted in Figure 2.1.

Table 2.1: The snow pack stability depending on the slope aspect and time of the year in Northern hemisphere.

	Winter	$\longrightarrow$	Summer
North	unstable		stable
South	stable		unstable

During the winter months, the sun exposure to the north aspect slopes is small, therefore, the snow pack is not developing stability and has a tendency to create weak layers. The south facing slopes have the opposite tendency - the snow pack is more stable since it is more exposed to the sun and the snow is melting during the daytime and freezing during the night. In the time between winter and summer, the overall air temperature increases, therefore, the snow pack on the north side of the mountain becomes more stable and it is still protected from the sun. The south aspect has more direct sun radiation and the air temperature is increasing, therefore the snow pack might get more unstable due to rapid melting. This can lead to wet slab avalanches described in Section 2.1.1. Authors in [25] mention that statistics have proven that most avalanche accidents occur in the north facing slopes. However, this is not based on slope usage which has to be considered.

2.2.3 Slope

The steepness of the slope is another important avalanche contributing factor. Sources like [17] mention that it might be the most influencing parameter in avalanche occurrences. As stated in [25], the slope inclination has to be between 20 to 50 degrees to enable the initiation of an avalanche. When the slopes inclination is smaller than 20 degrees, the shear deformation is too small to initiate fractions in the snow pack. The largest avalanches occur between 30 and 45 degrees. The areas with the most frequent avalanche occurrences are between 35 and 40 degrees. It can be explained as the least stable angle at which most of the snow pack can collect. When the inclination is bigger than 55 degrees, nature naturally releases the snow pack in sluffs and prevents slab formation in the area.

2.2.4 Land cover type and forest cover

Land cover or vegetation map has been mentioned in multiple resources. It is self-evident that some of the land cover types are safer than others. For example, avalanches are hardly ever released in urban areas or lakes. The land cover and forest cover property that influences avalanche occurrences the most, is the anchoring effect. The presence of trees and forests works as anchors holding the snow pack in one place. When an avalanche is already occurring, the trees can slow down the snow flow and minimize the damage. The denser the forest is, the better the anchor works [25]. It is estimated that to prevent avalanches, the density of the trees in the forest has to be 1000 per hectare (100 x 100 meters).

Another interesting property of the land cover that was already mentioned in the Section 2.2.2 is that other contributing factors can be linked to land cover. For example, trees just simply can not grow at high altitudes due to moisture and air temperature factors. In flat areas, where the steepness of the slope is smaller, more built-up areas are present.

2.2.5 Proximity of people

The number of winter sports activists has grown in the past decades. The number of avalanches has increased as well. As mentioned in Section 2.1, around 90% of the dry avalanche cases have been triggered by humans. Therefore, one of the avalanche contributing factors is proximity of people. The majority of avalanche incidents has been triggered during human leisure activities like skiing, snowboarding, climbing and hiking. To include this factor, more information about human locations is needed, such as ski resorts, actual skiing pistes and ski lifts. The proximity of smaller and bigger cities, roads and even railways can indicate this as well.

2.3 Geographic data

The research problem is a spatial problem and one of the most challenging aspects of geospatial analysis is the data. Geospatial data can be stored in dozens of file formats and database structures. Additionally, almost any file format can technically contain geospatial information simply by adding a location. Another problem with the Geographic information system (or GIS) is the different coordinate reference systems. The following subsections explain the geographic data file types used in the research, and describe coordinate reference systems and the theory behind extracting slope steepness and aspect from elevation data, as well as the rasterization of polygons, lines, points and transformation algorithms of raster files when changing the resolution. The spatial data sets used in this research is described in the Section 3.3.

2.3.1 Shapefiles and raster files

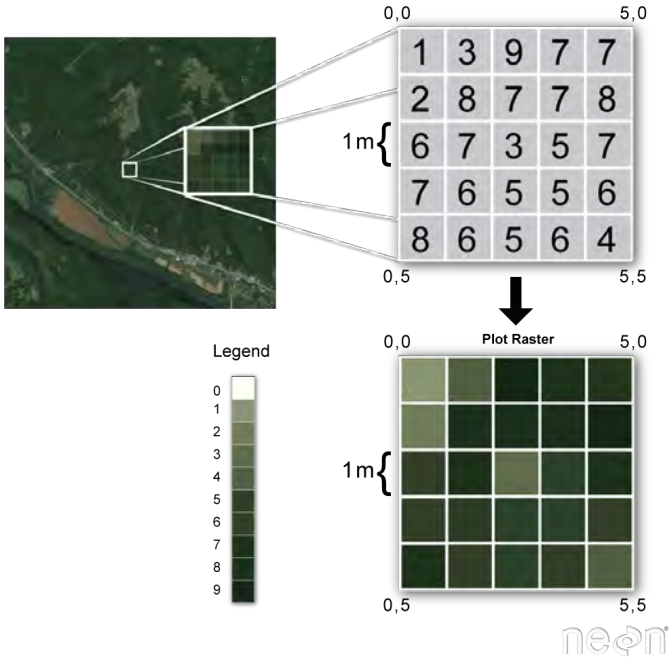
The geographical data types used in this project are mainly shapefiles (files with extension .shp) and raster files (.tif). An introduction of these file types is given in the following section.

Shapefile. Shapefiles were introduced by Environmental Systems Research Institute (ESRI) with ArcView GIS version 2 in the early 1990s. The following information has been taken from the ESRI Shapefile Technical Description [10] published in 1998. A shapefile format stores geometric location and additional information about the attribute in a vector. The shapefile does not save topological information. This file format is relatively simple - it stores the data as a shape of points, lines or area features. The latter are called polygons and they represent a vector of closed loop coordinates.

Although the term "shapefile" is commonly used, it might be misleading as this format does not consist only of one .shp file. Shapefile is a collection of at least 3 mandatory files with suffixes .shp, .shx and .dbf. All three of these files have to be named the same in order to be able to use them. The .shp file alone is incomplete.

Raster file. The second geographical data type that has been used in this research is the raster file, which has a very simple structure. It consists of a matrix of regular, same sized cells, known as pixels, that are organized into a grid - rows and columns. Every pixel represents an area of land on the ground and contains a value that represents some information about this area. The raster concept is represented in Figure 2.1.

The information saved in rasters can be either discrete (positive or negative integer) or continuous (positive or negative floating point). Discrete data can represent categorical information. These



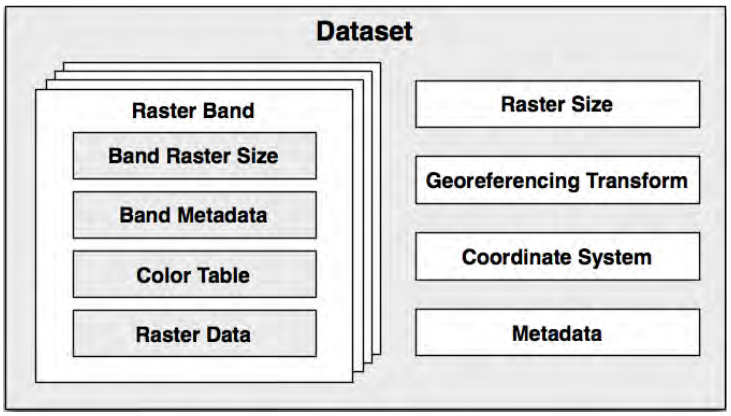
Source: Colin Williams, NEON.
Figure 2.1: A concept of a raster file with resolution of 1 x 1 meter, values in range $[0, 9]$ and extent of $[0, 5]$ X $[0, 5]$ meters.

categories could be different types of the same variable such as land use, or soil, aspect of the mountain slope or even indicate a presence of a factor like forested areas and non-forested areas. Continuous data can represent features like elevation, temperature or precipitation. These rasters are used to display data layers along with other geographic data and to perform spatial analysis.

In order to work with raster files, a Geospatial Data Abstraction Library (GDAL) has to be introduced. Originally, GDAL was just a package for working with raster geospatial data. Nowadays, there is an extra library called OGR integrated in GDAL, which was intended to work with vector data. The research focuses more on raster files, so only GDAL will be explained in more detail. The Figure 2.2 represents the structure of a data model in order to describe a raster geospatial dataset.

A raster data model consists of various parts. The following properties are taken from the source [20]. Let's look at them separately:

- raster band:
  - band raster size - size (number of pixels both horizontally and vertically) for the data within the band. This may be the same as the raster size for the overall dataset, in which case the dataset is at full resolution, or the band's data may need to be scaled to match the dataset;
  - band metadata providing extra information specific to this band;
  - color table - saves information on how the pixel values are translated to colors;



Source: [20]

Figure 2.2: Raster over the same extent, at 4 different resolutions

  - raster data itself.
- raster size - width and height of the image, in pixels;
- georeferencing transform - converts the raster coordinates (x, y) into georeferenced coordinates, that is, coordinates on the surface of the earth, like latitude and longitude;
- the coordinate system - describes the georeferenced coordinates produced by the georeferencing transform. The coordinate system includes the projection and datum, as well as the units and scale used by the raster data;
- metadata - additional information of the dataset as a whole.

2.3.2 Geographic coordinate systems (GCS)

Spatial data is very similar to regular datasets as they represent an array of numbers. The only difference is that spatial data also carry numerical information that can locate the features on the surface of the earth. This is necessary in order to combine different variables from other data sets to one specific location and to perform accurate spatial analysis. Every spatial dataset (rasters, polygons, lines, points) is saved in some coordinate system. The representation of coordinates may differ depending on the coordinate system but they might be specified as decimal degrees for latitude and longitude, meters, feet or kilometers.

In total there are a couple of hundred geographic coordinate systems and a few thousand projected coordinate systems. All these systems are defined as follows [9]:

- Its measurement framework, which is either geographic (in which spherical coordinates are measured from the earth's center) or planimetric (in which the earth's coordinates are projected onto a two-dimensional planar surface);
- Units of measurement (typically feet or meters for projected coordinate systems or decimal degrees for latitude-longitude);
- The definition of the map projection for projected coordinate systems;

- Other measurement system properties such as a spheroid of reference, a datum, one or more standard parallels, a central meridian, and possible shifts in the x- and y-directions.

Knowing these properties, two datasets from different coordinate systems can be aligned using geographic (datum) transformation. This is a well-defined mathematical method to convert coordinates between two geographic coordinate systems. As with the coordinate systems, there are several hundreds of predefined geographic transformations that can be accessed. It is very important to correctly use a geographic transformation if it is required. When neglected, coordinates can be in the wrong location by up to a few hundred meters. Sometimes no transformation exists, and a third GCS like the World Geodetic System 1984 (WGS84) has to be used to combine two transformations.

2.3.3 Extracting aspect and slope steepness from DEM

It was mentioned in Sections 2.2.2 and 2.2.3, that aspect and steepness of the slope are important avalanche contributing factors. These features can be extracted from the digital elevation model (DEM) - a 3D computer graphics representation of a terrain's surface. It is created from the terrain's elevation data. The following section describes the algorithms behind the calculations of aspect and slope steepness. These algorithms are based on the sources [7] and [8].

The input raster for both algorithms is elevation data. Every cell contains a value indicating the average altitude in the pixel area. A moving 3 x 3 window visits each cell in the input raster, and for every cell in the center of the window, an aspect or slope value is calculated. These algorithms use all of the 8 neighbors of the central cell to calculate the new features. The surface 3 x 3 window is illustrated in the Figure 2.3. The cells are notated with letters a to i, where the cell e represents the center for which the aspect and slope is calculated.

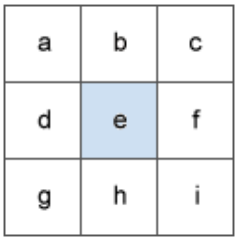


Figure 2.3: Moving 3 x 3 window for aspect and slope calculation.

The Aspect algorithm. Let's indicate the elevation above the ground surface with a plane z. Conceptually, the aspect algorithm fits a plane to the z-values (altitude) of a 3 x 3 cell neighborhood around the center cell. The direction the plane faces is the aspect for the center cell.

The rate of change in the horizontal (x) direction for cell e is calculated with the following formula:

$$\left[\frac{dz}{dx}\right] = \frac{(c + 2f + i) - (a + 2d + g)}{8}. \tag{2.1}$$

The rate of change in the vertical (y) direction for cell e is calculated with the following formula:

$$\left[\frac{dz}{dy}\right] = \frac{(g + 2h + i) - (a + 2b + c)}{8}. \tag{2.2}$$

Taking the rate of change in both the x and y direction for cell e, aspect is calculated using:

$$aspect = 57.29578 * arctan2\left(\frac{dz}{dy}, -\frac{dz}{dx}\right), \tag{2.3}$$

where $arctan2(x, y)$ is a 2-argument arctangent and the number 57.29578 is a rounded version of the result from $\frac{180}{\pi}$.

The aspect value is then converted to compass direction values (0-360 degrees), according to the following rule:

$$cell_{aspect} = \begin{cases} 90 - aspect, & \text{if } aspect < 0, \\ 360 - aspect + 90, & \text{if } aspect > 90, \\ 90 - aspect, & \text{otherwise.} \end{cases} \tag{2.4}$$

The compass direction values can be used to categorize the aspect using the following rule:

$$aspect_{cat} = \begin{cases} N, & \text{if } cell_{aspect} \text{ (337.5;360] or [0;22.5]}, \\ NE, & \text{if } cell_{aspect} \text{ (22.5;67.5]}, \\ E, & \text{if } cell_{aspect} \text{ (67.5;112.5]}, \\ SE, & \text{if } cell_{aspect} \text{ (112.5;157.5]}, \\ S, & \text{if } cell_{aspect} \text{ (157.5;202.5]}, \\ SW, & \text{if } cell_{aspect} \text{ (202.5;247.5]}, \\ W, & \text{if } cell_{aspect} \text{ (47.5;292.5]}, \\ NW, & \text{if } cell_{aspect} \text{ (292.5;337.5].} \end{cases} \tag{2.5}$$

Slope steepness The output raster from the slope calculation can be expressed by two types of units - degrees or percent (percent rise). The percent rise is easier to interpret, therefore, it is used in this research and explained in the following paragraph.

The change rate of the surface in the horizontal $\left[\frac{dz}{dx}\right]$ and vertical $\left[\frac{dz}{dy}\right]$ directions from the center cell determines the slope. The same notation as indicated in the Figure 2.3 is used in order to

calculate the rate of change for cell e in the x direction:

$$\left[\frac{dz}{dx}\right] = \frac{(c + 2f + i) - (a + 2d + g)}{8 * x_{cell\_size}}, \tag{2.6}$$

and y direction:

$$\left[\frac{dz}{dy}\right] = \frac{(g + 2h + i) - (a + 2b + c)}{8 * y_{cell\_size}}. \tag{2.7}$$

The slope steepness value in percentage for cell e can be computed by the following formula:

$$cell_{slope} = \sqrt{\left[\frac{dz}{dx}\right]^2 + \left[\frac{dz}{dy}\right]^2} * 100. \tag{2.8}$$

The percent rise is the rise divided by the run, multiplied by 100. The percent rise may observe values bigger than 100. For example, consider a triangle B in the Figure 2.4. When the angle θ is 45 degrees, the rise is equal to the run, so the percent rise is 100%. As the slope angle gets steeper and approaches 90 degrees, like the triangle C in Figure 2.4, the percent rise begins to approach infinity.

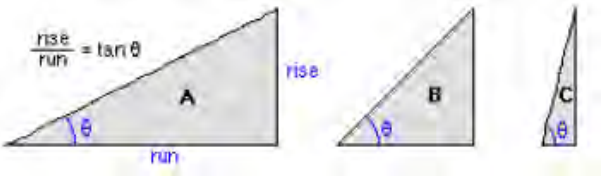


Figure 2.4: Examples for different slope steepnesses depending on the angle.

2.3.4 Spatial data preparation

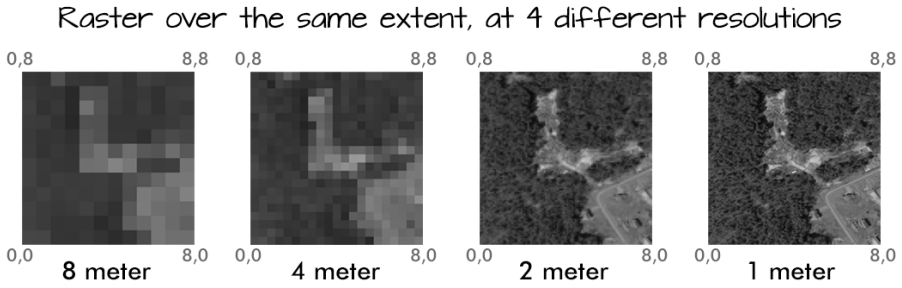
The main goal of this research is to predict the riskiness of a location. Therefore, it is defined as a classification problem. Some adjustments to the data files had to be done to fit this problem. In a classical classification problem, the data is saved in a table (data frame) with columns indicating the independent variables $x_j(i)$, and a separate column indicating the class $y(i)$, whereas the rows are the observations $i = 1, ..., n$. In this spatial problem, so far there have been separate files with terrain characteristics saved in rasters, point, line and polygon shapefiles. This information has to be combined in one structured file. When the rasters have the same pixel size and the same size of the territory, it is not that hard. But that is not always a case. Therefore, this section explains the concept behind changing the resolution for numerical and categorical type rasters, and how to save the information from shapefiles to rasters.

The area of interest is divided in the same sized h x h cells by the coordinates. As the spatial data always contain information about the location, the values can be merged by the unique values of location. Therefore, the observations indicate different locations - cells in a grid (pixels).

2.3.4.1 Change resolution of a raster

The different raster files can be stored in different resolutions and different spatial extents (bounding boxes). A bounding box represents the (X,Y) coordinates of the raster corners in geographic space. It tells the GIS how to distribute every pixel in a 2D space. To correctly combine information about the same territories, the cell location and resolution has to be set the same for all variables. When the extent with x and y coordinates is set, the resolution can be changed.

A raster can be re-sampled in order to adjust the resolution. When a raster with a higher resolution is needed, a grid with more pixels can be applied to the same extent. If a lower resolution raster is needed, a grid with less pixels within the same extent can be applied. An example of different resolutions of the same extent are illustrated in the Figure 2.5.



Source: Colin Williams, NEON.
Figure 2.5: Raster over the same extent, at 4 different resolutions.

The best way to do this, is to create a raster template of the resolution and spatial extent that is needed, and resample the original variable raster files. There are multiple methods to perform the resampling, depending on the structure of raster information, but nearest neighbor assignment is used for categorical data and bilinear interpolation is used for numerical data.

The nearest neighbour method is the fastest of the interpolation methods. As it does not change the values of the cells, it is best used for discrete data. First, in the output raster, the location of the cell's center has been determined. Then, the algorithm looks for the closest cell center from the input raster to this location. When it is found, this value is assigned to the output raster cell. The maximum spatial error in this method is one-half of the cell size.

The bilinear interpolation looks at the weighted distance average of the four nearest input cell centers. The value of an output cell is calculated by averaging (weighted for distance) the values of the surrounding 4 pixels. This method is used for continuous data as it smooths the surface more than the nearest neighbor. The values can take decimal values and can still be interpreted.

2.3.4.2 Rasterize shapefiles

It is clear that spatial data saves a geographic location together with features. The shapefiles does not say anything about the topology around the shapes. The shape values are needed to transfer and save in a raster cell. This can be done via rasterization. There are some differences between rasterizing points, lines and polygons so let's look at them separately.

If the shape represents points, the output cell saves this information in the whole cell. The value for the output grid cell can be determined by a function. In this research, when a point or multiple points fall into the same cell, only one value is assigned.

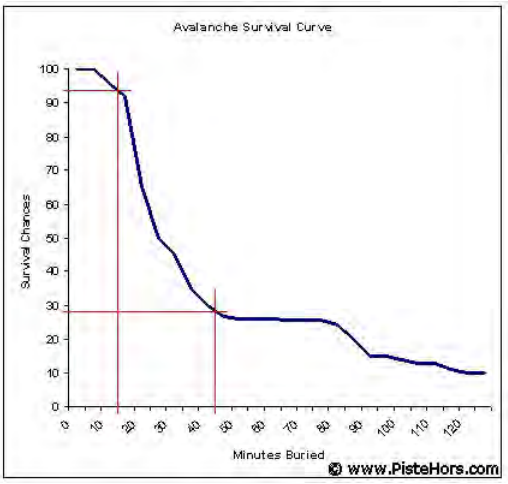
If the shape represents lines, then for every cell, through whom the line goes, a value is assigned even if the line barely touches the cell. All the cells that do not touch the shape have an empty value.

By the standard procedure the polygon shape gets the values saved in the cell only if the polygon covers the center of a raster cell. The procedure in this research was similar to the line approach. First, the fraction of each grid cell that is covered by the polygons is calculated. If a value in a cell is bigger than 0, a value is assigned to the cell. If the fraction is 0, the cell value stays empty.

2.4 Search and Rescue operations

As described by [27], the majority of victims die from asphyxia which is the inability to acquire sufficient oxygen through breathing or, in simple words, death due to the lack of oxygen. The reason why victims hardly can survive for more than 90 minutes is hypothermia - the body can not keep the temperature, therefore, the person freezes to death. Another cause of death is physical trauma during the fall in an avalanche. Hence, surviving an avalanche is a race against time. For the Search and Rescue (SaR) teams to save an avalanche buried victim, the first 30 minutes are crucial. According to the Figure 2.6, at 15 minutes about 9 in 10 people buried in an avalanche can survive, but by 30 minutes, only 50% survive. This information is extracted from accidents between 1981 and 1991 by Swiss Avalanche Research Center at Davos.

The SaR team comes into action when they receive an alarm message from skiers who have noticed an avalanche. Quite often the people are from the same group as the victim. The avalanche SaR teams use different ways to locate the victims. When victims are not equipped with electronic transceivers like RECCO, the trained avalanche-rescue dogs can be used to find human odor that rises from the snow. If the victim is equipped with an electronic transceiver, fellow skiers can immediately start searching for the missing person, even when the SaR team is only on its way to the incident place. However, these transceivers have a finite transmission time. As soon as the area of interest has been determined, collapsible probe sticks are extended



Source: http://pistehors.com/backcountry/wiki/Avalanches/Avalanche-Survival-Curve
Figure 2.6: Avalanche survival curve depending on the time a person has been buried.

and used to locate the victim by penetrating the snow. The problem with probe lines is that it costs a lot of energy and time, as the area of interest has to be covered by small, spiral-shaped areas. And time is crucial when saving the avalanche victims.

2.5 Unmanned aerial vehicles

The technology has been advancing and humanity has reached a point where there are countless possibilities for utilizing unmanned aerial vehicles (UAV) in different industries. One of the applications is for SAR operations. As the time factor is so crucial in the SAR operations, drones can be used to gather evidence about the location of a missing person while the SAR team is still preparing for its operation. UAVs are agile, fast and can exhibit autonomous behavior. This helps perform operations that are difficult for humans at relatively low costs.

The benefits of drone development are clear, but there are some limitations and disadvantages for drone usage:

- regulations for flying drones are still under development and differ from country to country;
- restricted areas such as close to military zones, airports and densely populated places;
- drones can be used for privacy invasion or to target specific population groups;
- the battery management could be improved as this currently limits flight times.

The internship focuses on the mountain Search and Rescue operations, but the project UAV Retina considers also other applications such as fire fighter operations and detecting improvised explosive devices (IED). The main focus has been put on the fire fighters operations. During these, a drone can be used to capture footage about the situation inside a building that is on fire from above, and angles that fire fighters can not access themselves. An image recognition on

information extracted by infrared cameras can be used to support firemen by indicating locations of people, spot the fire source and give a confirmation to fire fighters that it is safe to execute their plan. The same footage can be used after the operation in order to perform training for new firefighters.

Chapter 3

Data and preliminary data analysis

3.1 Data

In this chapter a description of the data used for the risk model as an input is explained. Additionally, an initial analysis and performed data processing is described. When thinking about the application of victims within avalanches, you have to think of the contributing factors and where to find the data. First, the datasets of historical avalanches will be described. Second, the contributing factors that are mentioned in the Section 2.2 will be explained in more detail in the following sections. All the data sets used in this project are open source and the websites of the sources are summarized in Table 3.1.

Table 3.1: Data sources used in the project.

Data	Source
Historical avalanches	[6]
Elevation	[11]
Landcover	[13]
Forest cover	[35]
Coordinates of ski resorts	[23]
Ski slopes	[22]
Shape files	[21]

Working with natural disasters like avalanches, you have to understand that the interpretation of data is complicated as the set contains only the reported occurrences and often rely on a relatively small number of events. It means that this data set does not contain all the events that have happened. As discussed in the Chapter 2.1, 90% of the avalanches are initiated by a human but there are still a lot of naturally released avalanches in areas untouched by humans. Further, the severity of incidents influences the reporting rate: accidents resulting in casualties

are generally well reported and documented, and less-severe incidents are often under-reported [36].

3.2 Historical avalanches

Historical avalanche datasets over different territories are available on the Internet. The initial plan in this research was to use the WSL Institute for Snow and Avalanche Research SLF dataset of fatal avalanche accidents in Switzerland over the period from 1995 to 2016. It is publicly available via their website. Although a lot of information can be extracted from this dataset, it contains only 401 accidents which is a relatively small set to use for spatial modelling.

The dataset used in this research was successfully obtained via getting in touch with the source [6]. The received file contains 1671 avalanches from all over the world but mostly recorded in France. The decision was made to focus only on one area and explore it in more detail. Therefore, the territory of the French Alps is the target location for modelling. The steps taken to clean historical avalanche data set are presented in Table 3.2. The 1181 avalanche observations in the French Alps are used in this research to build a grid-based spatial risk model.

Table 3.2: Cleaning procedure of the historical avalanche dataset.

Step	Action	# of observations
0	initial dataset	1671
1	longitude/latitude value empty	256 (1415 left)
2	take France territory	142 (1273 left)
3	take French Alps territory	92 (1181 left)

The language throughout these observations is French. A small preview of data is captured in Table 3.3. The initial analysis is done using Latitude/Longitude coordinates in EPSG projection 4326 - World Geodetic System (WGS) 84. The geospatial factors are also converted to the same coordinate projection to make the analysis convenient and concise.

Table 3.3: Avalanche data set preview

observer	massive	Date	Description	cause_1	danger_lev	latitude	longitude
Duclos Alain	Vanoise	14/02/2005	Après un tir d'explosif négatif...	Skieur hors piste	4	45.252708	6.74296
Xuereb François	Vanoise	03/03/2018	Vaste avalanche partie...	Neige	3	45.270099	6.746063
Lacheré François	Lauzière	03/05/2012	Vrai plaque partie soudainement...	Skieur rando. descente		45.540305	6.419608
Pautas Roland	Haute Maurienne	06/11/2011	Nombreuses avalanches spontanées...	Neige	4	45.306025	7.03645
Mondon Gaël	Aiguilles Rouges	05/04/2010	Ce sont trois skieurs de randonnée...	Skieur rando. descente	3	46.000921	6.888653

Later, in the data preparation part, all the data files are converted to EPSG projection 3395. This is done in order to use coordinates and distances in meters to be more precise, as the distances in different latitude and longitude between coordinates differ.

3.2.1 Analysis

The number of avalanches within this dataset over a period of time is depicted in Figure 3.1. The file covers the past 30 years. Overall, the number of avalanches is going up. One of the reasons is the average temperature over the years is increasing. Therefore, the rise in temperature increases liquid water in the snowpack which in turn increases the shear deformation rate, causing stress, which is released when the snowpack collapses in an avalanche [2]. Another reason - the number of snow activists is increasing. Finally, avalanche reporting might not have been that common in the past as technology has been used more in only the past 20 years.

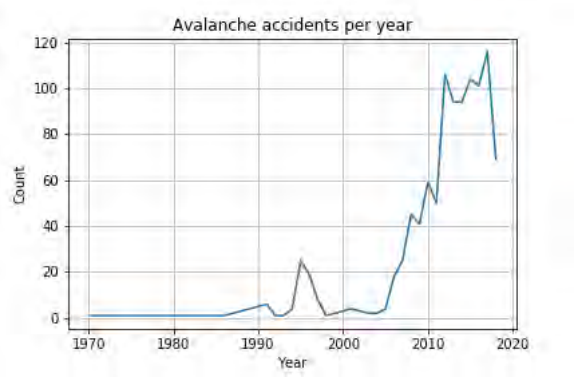
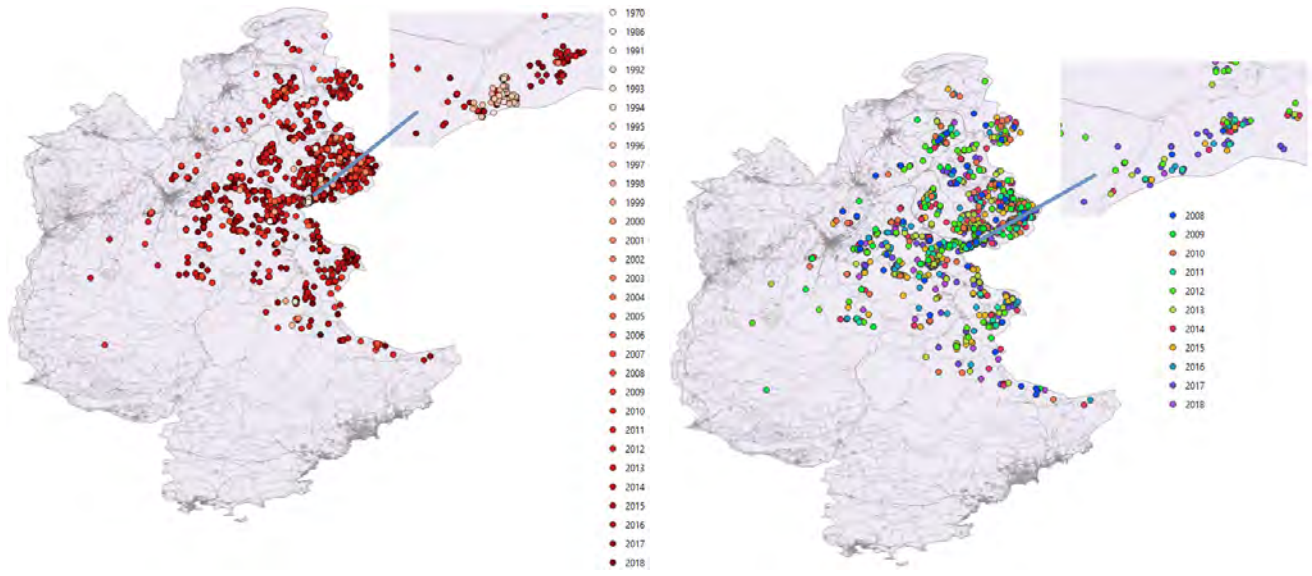


Figure 3.1: Number of avalanches over the years

Figure 3.1 shows a spike in the years 1995-1996. Looking closer on the avalanche locations by years in Figure 3.2a, the spike can be explained with occurrences around a specific ski resort within these years. After checking the observers of these records, 42 out of 44 records within these two years are made by the same person.

The Figure 3.2b represents all the avalanches from year 2008 (1051 avalanches $\approx$ 89%). The same region is zoomed in and shows that there are more avalanches over the years in this region, not only the ones in 1995-1996.

The data is very dense around certain regions, while in other areas the density becomes almost zero. This can partly be explained by geospatial factors and characteristics of the landscape. This will be explained in more detail in the following sections. A visual representation of the number of avalanche occurrences in every 100 by 85 km grids in French Alps can be found in Figure 3.3. This gives an overview of the differences in findings for each area.



(a) (b)

Figure 3.2: Visual representation of historical avalanches in period of (A) 1970 to 2018 and (B) 2008 to 2018.

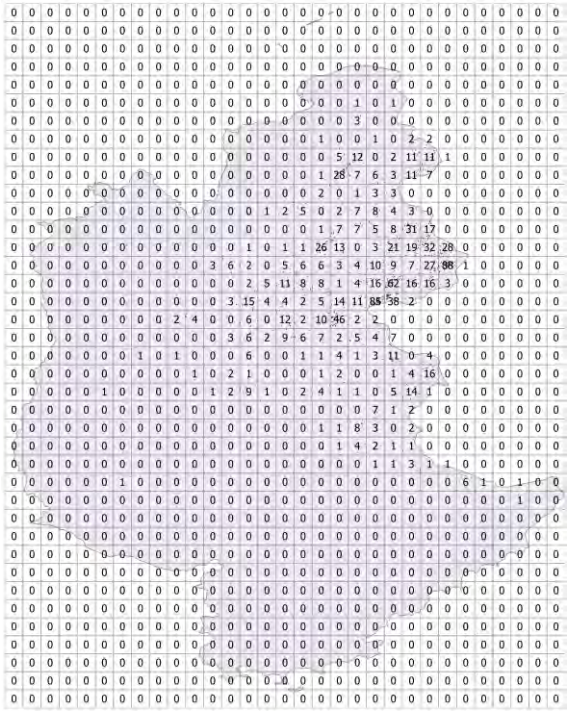


Figure 3.3: French Alps divided in 100 x 85 km quadrates each giving the number of historical avalanches in the quadrate.

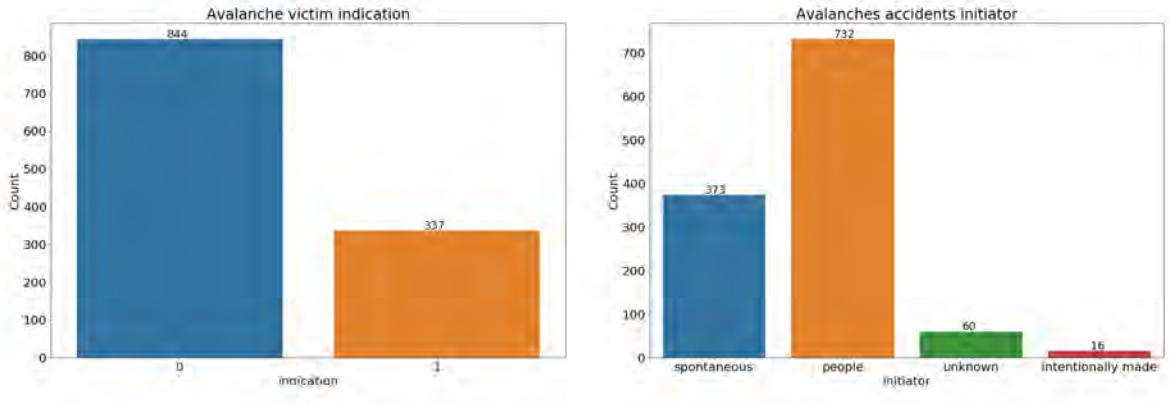
The further research focuses more on the mountainous area in the French Alps. This is done in order to avoid training the predictive model on areas where no avalanches have happened and where they can not even occur.

3.2.2 New variable creation

There is no clear indication in the historical avalanche dataset of whether an observation involves any victims. As the project plan is to predict the locations of potential victims buried in avalanches, the following text mining steps are used to extract the most common words from the *Description* field to find out more about the presence of the victims:

1. all words are changed to lower case;
2. remove punctuation;
3. remove French stopwords;
4. remove digits;
5. correct spelling with Levenshtein distance [14] equals 1;
6. lemmatization - get the root word of every verb.

A new indicator variable is introduced by looking at the cleaned most common stems in the *Description*. For the new variable *victim_indication*, the observations with words like 'victim', 'injured', 'died', 'buried', 'carried' are set to value 1 and 0, otherwise. The distribution of this variable is depicted in Figure 3.4a. We can conclude that the historical avalanche dataset does not include as many victims as would be needed to make predictions. Therefore, an assumption has been made that every avalanche in the dataset could potentially involve victims as these observations were reported by people. Now, all the observations can be used to predict the likelihood of avalanches and, therefore, victims.



(a) (b)

Figure 3.4: Visual representation of new variables (a) *victim_indication* and (b) *initiator*.

The same text mining procedure is used to extract information of the avalanche cause from *Description* and *cause_1* fields. The example of key words can be found in the Table 3.4. The first column indicates the word in French, second column translates the same word in English,

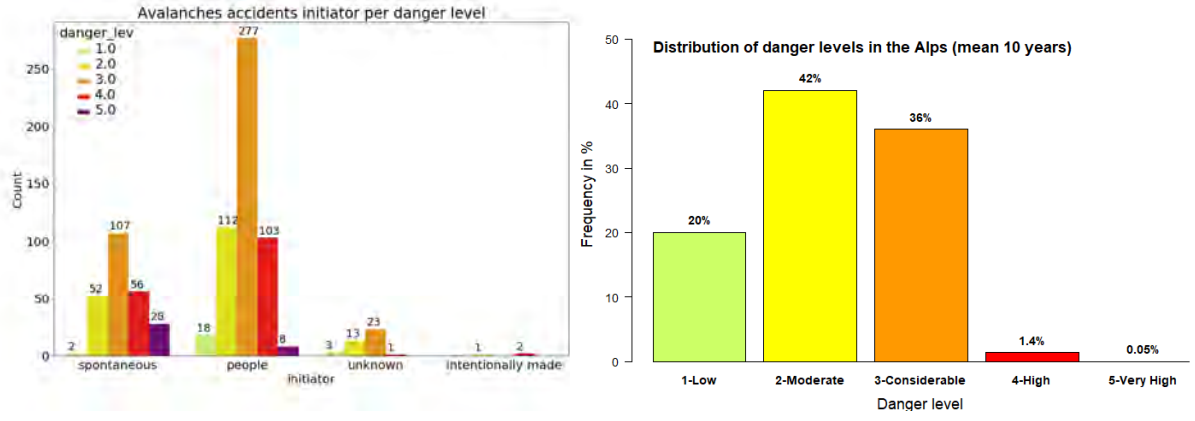
third column indicates the values for the *initiator* variable. The distribution of this new variable is pictured in Figure 3.4b.

Table 3.4: Key words used to make variable *initiator*

French	English	value
spontané, naturelle	spontaneous, natural	spontaneous
skieur, groupe, randonnée, snowboarder, pisteur, alpinistes, morts, déclenchée, raquetteurs, personne, tué, accidentellement	skier, group, hiking, snowboarder, tracker, mountaineers, dead, unleashed, snowshoers, person, killed, accidentally	people
mauvais, temps, vent, neige, pluie, réchauffement	bad weather, wind, snow, rain, warming	weather
grenadage hélico, grenadage à main, gazex, catex ,minage de corniche	grenadage helicopter, hand grenading, gazex, catex, cornice mining	intentionally made
inconnue, avalancheur	unknown, avalancheur	unknown

The intentionally made avalanches stand for the artificially initiated falls of avalanches. This is a standard method of reducing avalanche hazard in order to protect highways, railways and ski slopes, all in circumstances where the traffic can be restricted or diverted while the avalanche falls [19].

Intuitively one would guess that spontaneous avalanches happen when the weather conditions are bad. In Figure 3.5a, the distribution shows this pattern - relatively more spontaneous avalanches historically happened when there has been a higher danger level. In the meantime, there are still a lot of spontaneous avalanches when the danger level is 2 and 3. Another usual pattern that has been mentioned in related literature is people initiated avalanches occur when the danger level is 4 or less, mostly 3. This can be explained by human nature to not go in dangerous places when a risk is said to be very high. Danger level 3 means that on many steep slopes the snow is only moderately or weakly stable [32]. The 10-year average distribution of danger levels in Alps from Figure 3.5b indicates that 36% of time there is a danger level 3. Although it is still unsafe to go off-piste, many people are taking the risk in order to feel the adrenaline.



(a) (b)

Figure 3.5: Visual representation of (a) number of avalanches per danger level by initiator and (b) danger level distribution in Alps over 10 years.

3.3 Geospatial factors

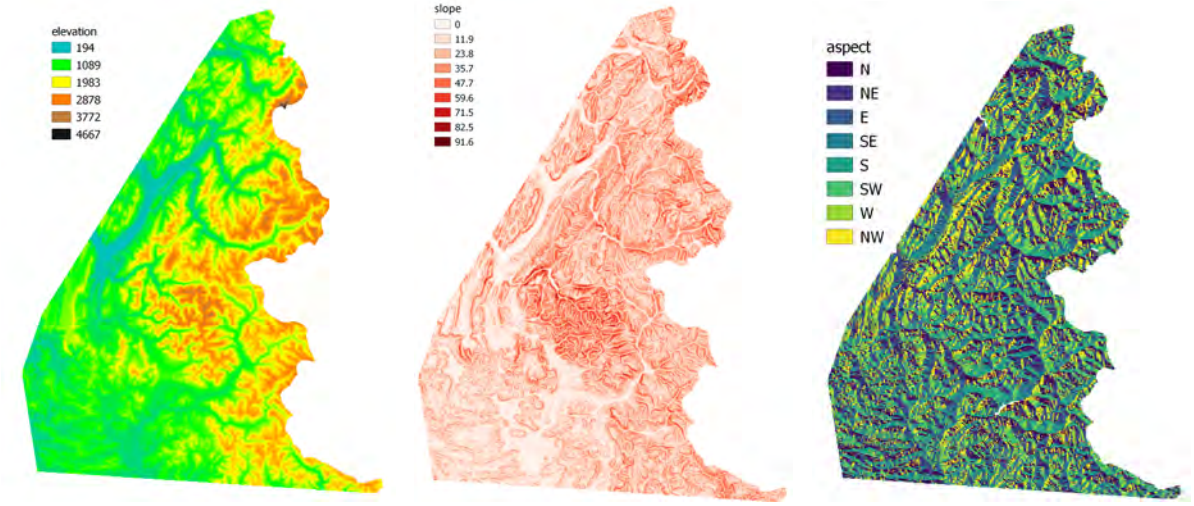
This section contains the exploratory data analysis of geospatial factors. It consists of a descriptive analysis in order to get useful insights and patterns about the data. First, some general information retrieved from the data is presented. Then, the distributions of the terrain characteristics and distances between avalanches and other factors are extracted. This information defends the decisions made in the data processing part.

3.3.1 Description

Geospatial factors are used as a landscape characteristics in order to estimate a risk of an area. Geospatial factors include elevation, roads, railways, settlements, waterways, skiing pistes, skiing lifts, ski resorts, land cover and forested ares. These files were open source and obtained online. From the elevation (DEM - Digital Elevation model) file, the slope aspect and steepness were calculated. The algorithms behind this procedure were described in Section 2.3.3.

These 12 factors were combined and used to calculate risk, based on the historical avalanches. As has been detailed in the introduction of this section, the data is quite rough and incomplete. For example, not all the roads are displayed properly and many small roads are missing entirely. The areas for soil type can be the same for areas of several square kilometers, which is not always the case. These facts have to be kept in mind when evaluating the built predictive models.

A visual representation of first three geospatial factors - elevation, slope steepness and aspect is illustrated in Figure 3.6.

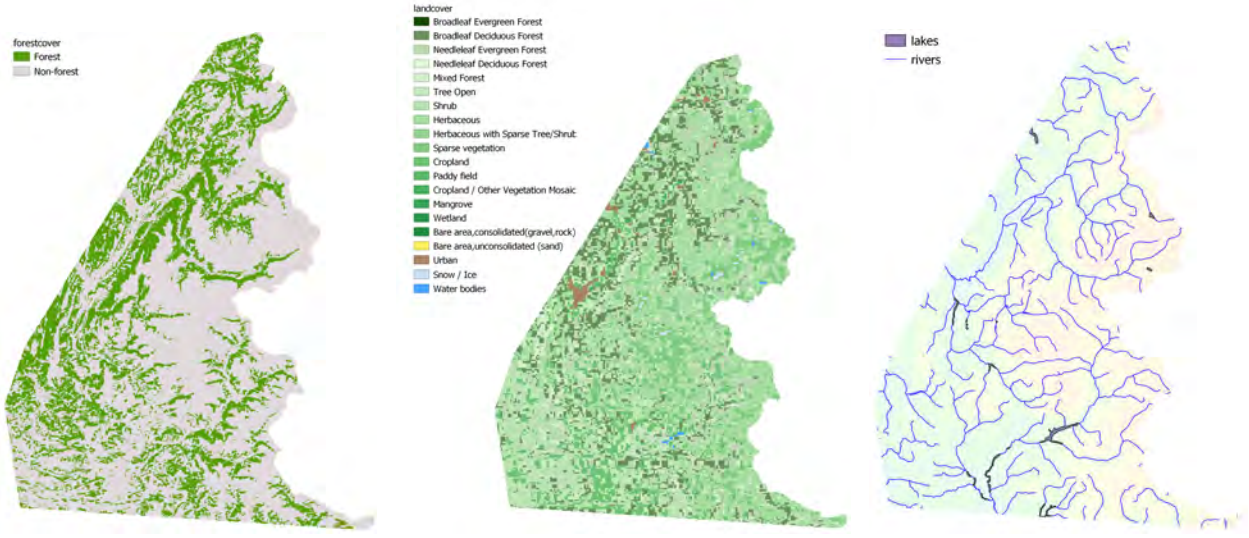


(a) (b) (c)

Figure 3.6: Geospatial factors for French Alps: (a) elevation, (b) slope steepness and (c) slope aspect.

In illustration 3.6a, the darker orange color represents a higher altitude and darker green/blue - lower altitude. The figure in the middle represents slope steepness. The steeper it is, the darker the red color. These values for every cell are represented as a percentage of the average steepness of an area. The light values give an indication of a more flat area. These values can be very well linked with the elevation data. Figure 3.6c illustrates the direction of the mountain slope that was computed from the elevation. The 8 values for the aspect indicate the compass directions - North, North-East, East, South-East, South, South-West, West, North-West and are coloured from dark blue to yellow. We are interested if the avalanches in the historical incident dataset indicates aspect values that are more risky than others.

In Figure 3.7, the representation of forested areas, land cover types and rivers are shown. Even by a quick look to the illustrations 3.7a and 3.7b, the resolution differences can be noticed. The forest cover raster is more detailed, meaning the raster resolution is smaller. This has to be corrected for every raster data set in order to align the data values for the grid cells. Figure 3.7c gives an indication of the river locations in French Alps.

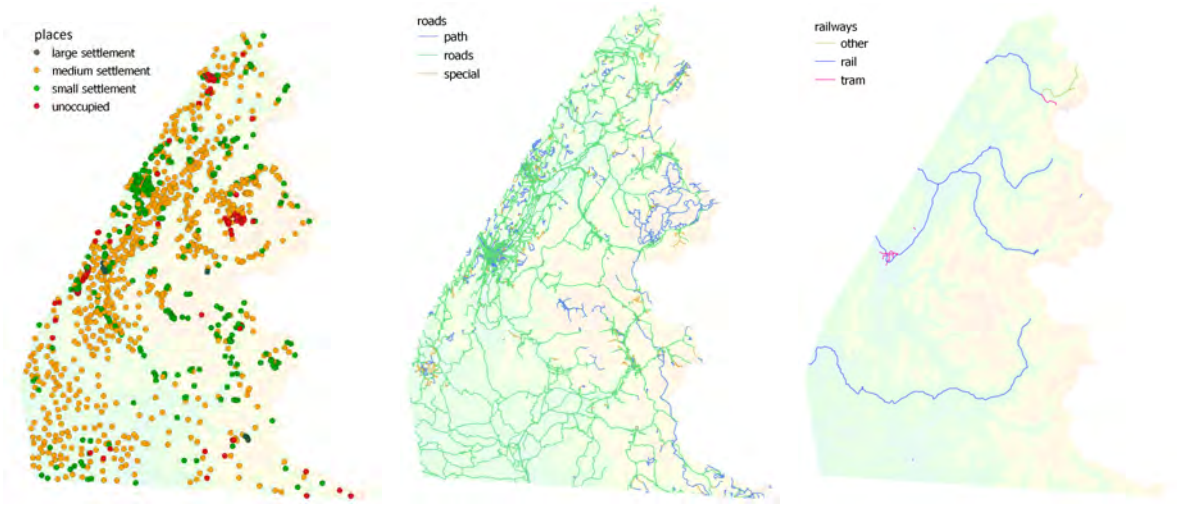


(a) (b) (c)

Figure 3.7: Geospatial factors for French Alps: (A) forest cover, (B) land cover and (C) lakes and rivers.

The illustration 3.8a gives an indication of locations for different types of populated areas - small, medium, large and unoccupied settlements. Figure (b) in 3.8 represents 3 different priority roads.

Path are only for pedestrians and not for any motorized vehicles. Type roads include highways and smaller priority roads that can be used by cars. Special roads are bicycle paths, cycling streets, raceways. The colors in the legend specify the type. As can be seen, paths are mostly in the mountainous areas and the highways are spread all over the area. The railways are not widely spread in the French Alps area as Figure 3.8c shows. The pink color presents a tram,

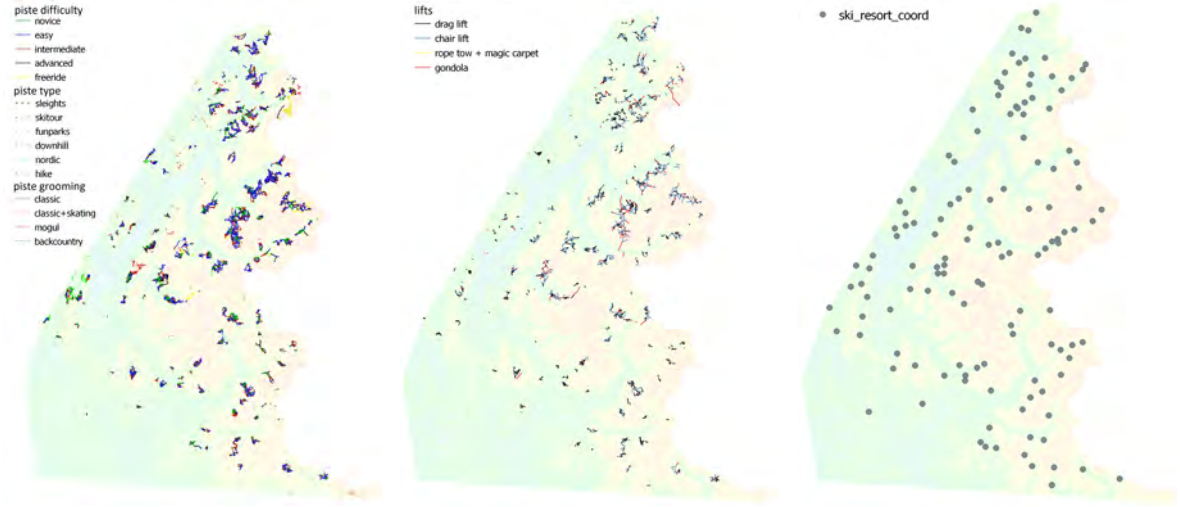


(a) (b) (c)

Figure 3.8: Geospatial factors for French Alps: (a) settlements, (b) roads and (c) railways.

meaning that it is also a city area. The train rails are in blue and the type "other" indicates a tourism train route.

Further, in order to acknowledge the potential places of people in mountains, the data of the skiing pistes and lifts were acquired. This information is collected by enthusiasts of the winter sports and is still growing. The benefit of this is that people are logging the paths they are taking so the information about off-pistes are also represented.



(a) (b) (c)

Figure 3.9: Geospatial factors for French Alps: (a) pistes , (b) ski lifts and (c) ski resort coordinates.

Solid lines in the legend of Figure 3.9a represent the difficulty of the ski piste, the dashed line indicates the activity type of the piste, like hiking trail, a downhill skiing slope, nordic skiing

slope and so on. The piste grooming indicates the preparation of the pistes and are presented in doted line. The same grooming name for different piste types are applied. For example, classic grooming for downhill piste is a standard ski run that is mechanical groomed by a tracked vehicle. For the nordic ski piste, classic grooming means there are two rails for classic style nordic.

In order to be able to ski down the mountain slopes, the person first has to get higher up in the mountain. This is done by ski lifts. Different types of ski lifts are presented in plot (c) of Figure 3.9. The illustration (c) shows ski resort distribution in French Alps.

The data illustrated in this section has been already cleaned. Data retrieved from the OpenSnowMap was in a dictionary format and tagged by different properties. These tags were used in order to extract the geometries from the .osm file by specific values and combined to reduce the number of categories in a variable. The analysis of the factors and the further preparation are described in the next sections.

3.3.2 Analysis

The historical avalanche data locations are combined with the geospatial factors in order to get some insights of their importance. The insights are compared with the predetermined expectations per factor on the basis of literature, indicated in Section 2.2.

Elevation In the literature, elevation has been mentioned as an important avalanche contributing factor. Linking the locations of avalanches with the altitude of these places, the frequency is presented in Figure 3.10.

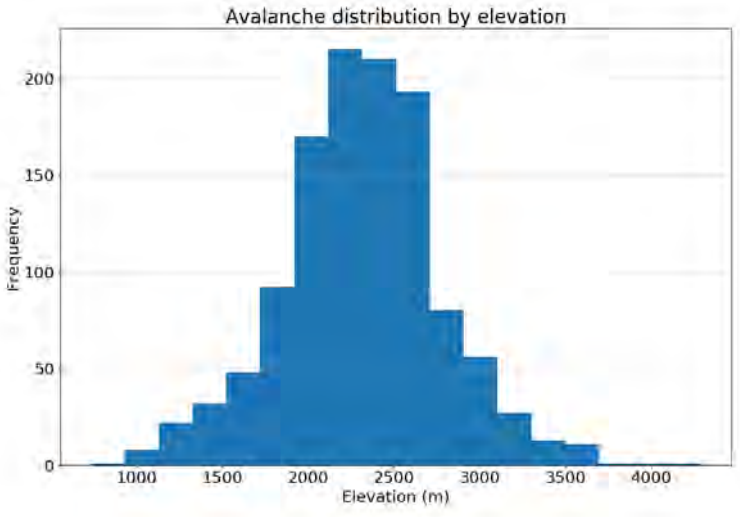
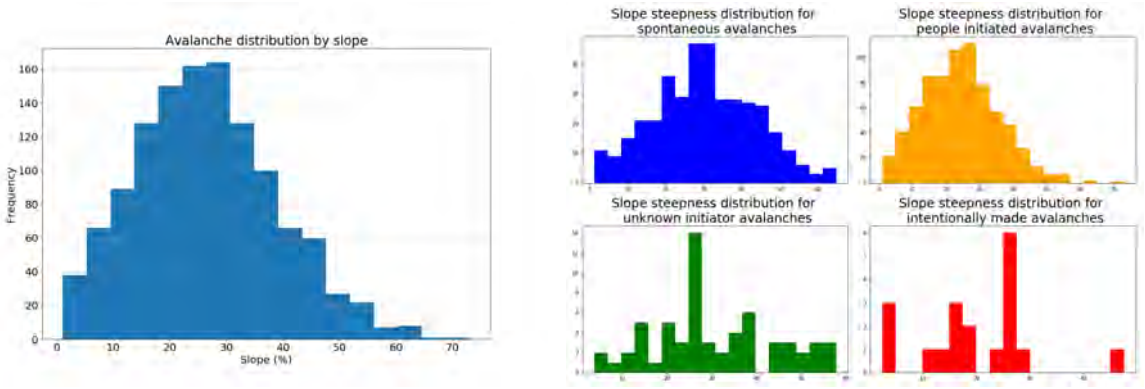


Figure 3.10: Distribution of elevation by avalanche occurrences.

It is clearly visible that the avalanche occurrence is dependent on the altitude. Remarkably, most avalanches occur around 2000 to 2700 meters. From 2700 meters, the number of avalanches decreases as the altitude increases. This means that most avalanches do not occur in the highest areas. This can be explained by the fact that $\approx 62\%$ of the avalanche initiators in the dataset are humans (see Figure 3.4b) and it is hard for people to perform activities at high altitude, therefore, there are also fewer ski areas. Since there is less snow and the slopes are less steep in lower altitude areas, it is logical to have fewer avalanches there. The histogram has a shape of a normal distribution and this pattern was also expected from the literature review.

Slope steepness The next important terrain factor is slope steepness. The distribution of the slope steepness (in percentage) by avalanche occurrence is presented in Figure 3.11a.



(a) (b)

Figure 3.11: Distribution of (a) slope steepness by avalanche occurrences and (b) slope steepness by avalanche initiator.

Most avalanches in the dataset have happened in the slope inclination of a range 15 to 40%. This aligns with the literature. Although slopes with less that 20% inclination should not be dangerous, the dataset reports avalanches even at the 5% inclination. This can be explained by the information of the avalanche dataset itself. We know that the avalanche observations were reported by people. This can lead to inaccurate coordinates. It may also be that the coordinates reported by observers are not the place where the avalanche was initiated, but only the end of an avalanche. The Figure 3.11a also shows occurrences in places steeper than 55%. According to the literature, these avalanches are happening spontaneously and released by nature. Looking closer to the avalanche initiators by slope inclination in Figure 3.11b, the spontaneous avalanches are released in a broad range. Again, one has to keep in mind that the locations can be misleading or the classified avalanche initiators may not be correctly identified during the text mining phase.

Aspect To gain an insight into the aspect influence on the avalanche occurrences, Figure 3.12 gives an overview of the number of avalanches per aspect per month.

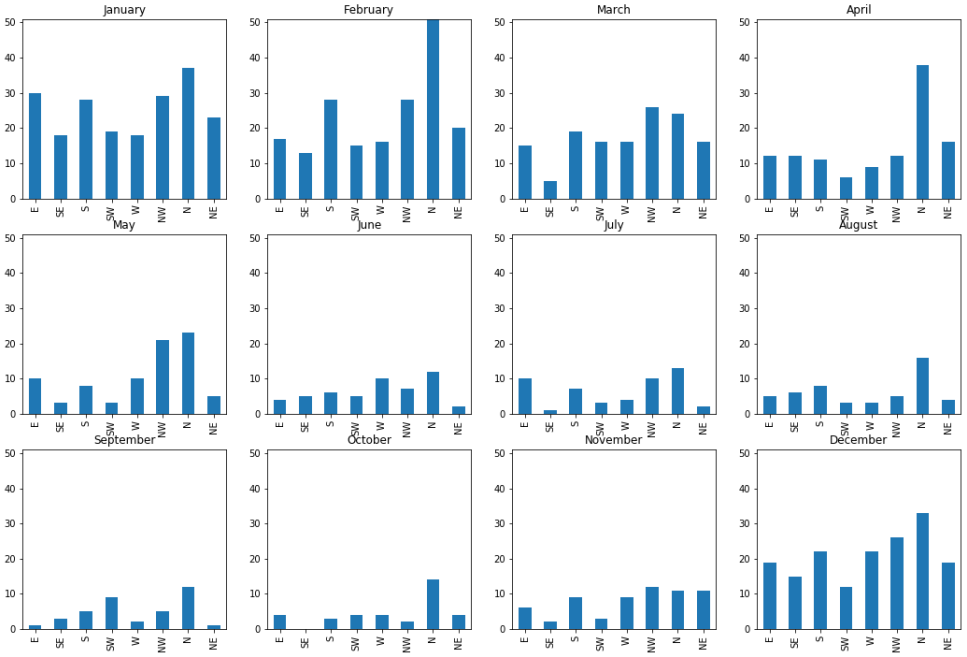
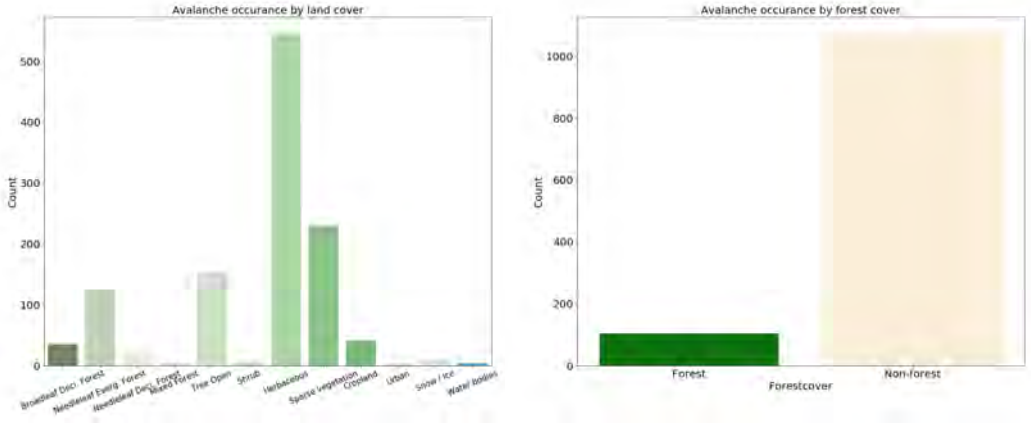


Figure 3.12: Distribution of avalanches per aspect and month.

The figure shows that most avalanches occur on the North slope no matter the time of the year. From the literature, the North side of the hill should be more unstable during the winter months and usually becomes more stable towards summer months (see Table 2.1). Overall, avalanches in summer months are not that frequent according to the historical avalanche data set, so it is hard to make stronger conclusions out of this.

Land cover type and forest cover When looking at the distribution of avalanche occurrences by land cover in the left graph of Figure 3.13, the urban areas and the water areas represent a small number of the total number of avalanches in the data set. In the literature, forested areas were mentioned as snow anchors. It is expected to have avalanches where less trees can be found. Both the land cover and the forest cover figures in 3.13 depict this behavior for the historical avalanches in the French Alps.

Proximity of people The descriptive analysis is straight-forward when the factors are saved in a raster file, as the whole area of French Alps is covered with values and it can be extracted for every location. The data sets used to indicate the proximity of people are shapefiles. The lifts, pistes, railways, roads and rivers are represented by lines. The coordinates of ski centers and settlements are represented by points, but lakes are represented by polygons. As described in 2.3.1, these files do not include information about the topography around the factor, but do



(a) Land cover (b) Forest cover

Figure 3.13: Distribution of avalanches per (a) land cover and (b) forest cover.

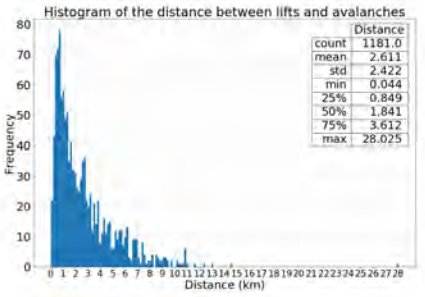
include information about the factor itself. We want to explore the influence of these factors to the avalanche locations. Therefore, to give an indication of the proximity of these factors to avalanches, the distances between avalanches and different factors were determined and analyzed.

In the first step, the importance of the factor itself is considered. In the next step, the patterns and the importance of the sub-types of these factors are considered. The histograms for the first step are illustrated in the Figure 3.14. From the histograms, it can be seen that some factors like lifts and pistes are close to the avalanche victims, but some factors are far away from the avalanches such as railways, rivers and lakes.

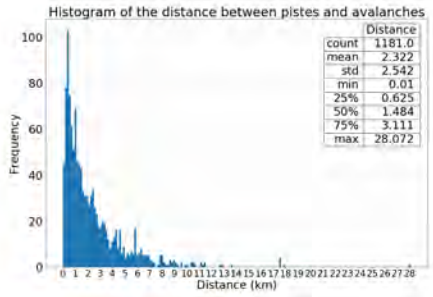
In the next step, the distances between the types of lifts, pistes, railways, roads, settlements and avalanches are analyzed. The histograms can be found in Appendix A. The information of the proximity of the factors have to be included in the grids. Therefore, different sizes of buffers around the factors can be created in order to indicate the significance of the distance to the incident. The next section describes the logic behind the buffers created.

3.3.3 Creating buffers

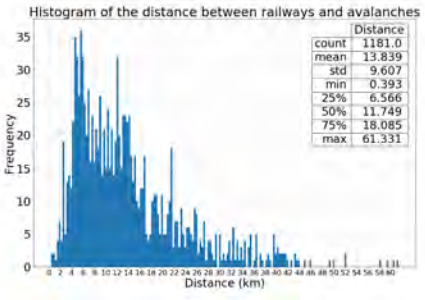
As already introduced in the previous section, buffers around variables had to be created in order to generalize the distances from the factor. With every buffer size, we want to make a separation between significant and not significant distances to avalanches. Therefore, histograms from Appendix A are used to make this separation. In order to indicate the location of the factor, a value of 1 has been assigned to these locations. There is only a small number of avalanche occurrences within the first couple of 100 meters from the factors, therefore a buffer of this size is made to indicate non-significance. The number of buffers per factor varies as they indicate different patterns. The border values for buffers have been chosen by visual inspection of the



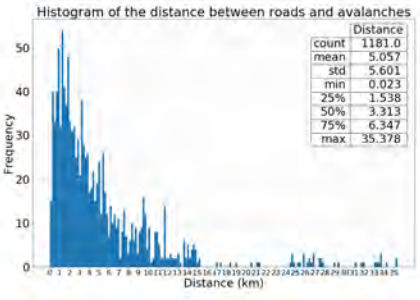
(a) Lifts



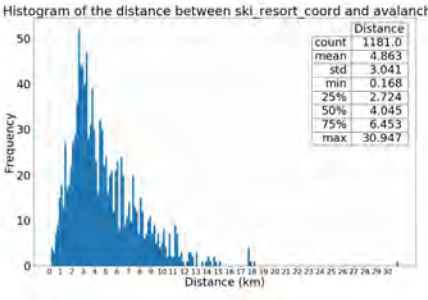
(b) Pistes



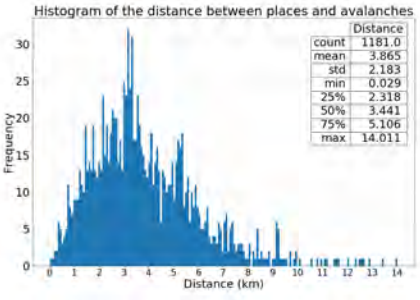
(c) Railways



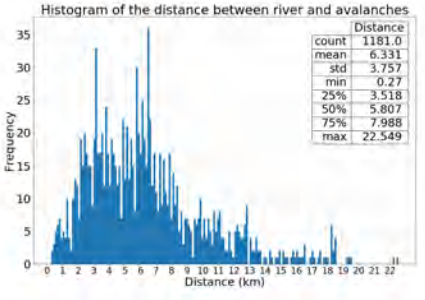
(d) Roads



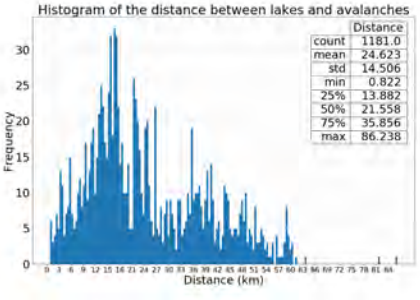
(e) Ski resort coordinates



(f) Settlements



(g) Rivers



(h) Lakes

Figure 3.14: Histograms of the distances between factors and avalanches.

distance histograms. The areas of buffers with non-zero values have been created so that it approximates around 75% of avalanche data in these buffers.

The sizes of buffers for every sub-type of factors - lifts, ski piste difficulty, ski piste type, ski piste grooming, roads, settlements and railways - individually are summarized in the Table 3.5. These buffers are used as categorical variables with values indicated in the "Value" column. The sub-types of factors that do not show clear patterns have not been used in the predictive modelling. Also, lakes and rivers indicate too big of a distance from avalanches so these factors will not be used for modelling.

Table 3.5: Buffer sizes (in meters) for sub-types of factors

Factor	Type	Value						
		1	2	3	4	5	6	0
pistes difficulty	novice	0	0-800	800-3500				>3500
	easy	0	0-250	250-2900				>2900
	intermediate	0	0-250	250-2000	2000-3500			>3500
	advanced	0	0-250	250-3000	3000-5000			>5000
	freeride	0	0-650	650-4000	4000-6000	6000-8000	8000-12000	>12000
pistes activity	hike				no clear pattern			
	funparks	0	0-2400	2400-6000	6000-9600	9600-12200		>12200
	nordic	0	0-1300	1300-9800				>9800
	downhill	0	0-1700	1700-3600				>3600
pistes grooming	mogul	0	0-600	600-6000	6000-9600	9600-12200		>12200
	skitour	0	0-2500	2500-16000				>16000
	sleights	0	0-3000	3000-9000				>9000
	backcountry	0	0-380	380-9000				>9000
	classic	0	0-1500	1500-5000	5000-8600			>8600
	classic+skate				no clear pattern			
lifts	chair lift	0	0-3000	3000-4800				>4800
	drag lift	0	0-250	250-2100	2100-4900			>4900
	gondola	0	0-6000	6000-11000				>11000
	rope tow	0	0-500	500-5200	5200-10500			>10500
roads	roads	0	0-1300	1300-4300				>4300
	path	0	0-170	170-4000				>4000
	special				no clear pattern			
railways	rail	0	0-4000	4000-6500				>6500
	tram				no clear pattern			
	other				no clear pattern			
places	small	0	0-1000	1000-2000	2000-3000	3000-7500		>7500
	medium	0	0-1300	1300-5600				>5600
	large				no clear pattern			
	unoccupied				no clear pattern			
ski resorts		0	0-2000	2000-3500				>3500

These buffers are still represented as shapefiles, only now they are polygons. To save the values of polygons in the grid cells, rasterization of shapefiles have to be performed. The algorithms behind rasterization have been described in Section 2.3.4.2. The next section summarizes the rasters of the factors that are used in modelling.

3.3.4 Preparation

As already mentioned in Section 2.3.1, spatial data can be saved in different file types. The factors - roads, highways, rivers, places, pistes, lifts and ski resorts - were saved in shapefile format as vectors of lines or points. The created buffers indicated in Table 3.5 are polygons. The factors - elevation, slope steepness, aspect, forest cover and landcover - were saved in raster files, but with different extents and resolutions. In order to combine all factors to a usable dataset for modelling, these have to be rasterized in a way that every pixel of a raster indicates the same area on every data set. Therefore, a raster template of the area has been defined. The summary of the parameters for the raster template can be found in Table 3.6.

Table 3.6: Summary of raster template for rasterization.

	parameter	value
extent	x_{min}	578210.8
	x_{max}	833710.8
	y_{min}	5448009
	y_{max}	5792009
dimensions	n_{row}	688
	n_{col}	511
	n_{cell}	351568
resolution	resol	500
coordinate reference system	CRS	EPSG:3395

The territory of x coordinates from 578210.8 to 833710.8 (in EPSG:3395 coordinate system) have been divided in 511 cells with every cell of a length 500 meters. The area of y coordinates from 5448009 to 5792009 were divided in 688 cells with every cell of height 500 meters. This way every cell in the grid represents a territory of French Alps. As the factors and their created buffers have been saved in a shapefile, these contain information about the location of an area. The buffer sizes of value 2 for factors are often small. It happens that the buffers often overlap with one another as the pixel sizes are wide. In this case, during the rasterization process, the smallest buffer value has a priority over the bigger one so that every pixel in the raster contain only one value that indicates the smallest distance to the factor.

Using the rasterization algorithms described in Section 2.3.4.2, the factors and their sub-types were transformed from shapefiles to separate raster files. The numerical rasters, such as elevation and slope steepness, were transformed to 500 x 500 meter resolution using the bi-linear interpolation. The categorical raster files like forest cover, landcover and aspect were transformed to the template resolution using the nearest neighbor method. These methods were described in Section 2.3.4.1. The avalanche dataset has been split in two sets, depending on the date of occurrence - observations until 01-01-2017 are rasterized and used as a train data set. Avalanches from 01-01-2017 until 02-12-2018 are rasterized and used as a test set.

The visual representation of the created raster files are summarized in Appendix B. Now, as every pixel indicates the same area for every raster file, a .csv file is created. The rows indicate the location of a 500 x 500 meter quadratic grid cell, and every column represents the factors and factor sub-types with values for the buffers. The .csv file now can be used as a dataset to perform predictive models. But first, the relationship between these factors has to be discovered.

3.3.5 Correlation between factors

The relation between all variables is analyzed in the final topic of this chapter. The correlation methods usually cannot deal with missing values, since complete pairs of observations are needed. In this research, all data is cleaned in a way that, if the geospatial factor has no value in the grid cell, the value of 0 is assigned (Table 3.5). Since this study uses continuous and ordinal variables, the Spearman's rank correlation coefficient is used to explore correlations between the variables [18]. The Spearman's rank correlation coefficient evaluates the monotonous relationship between two continuous or ordinal variables. In a monotonous relationship, the variables tend to change together, but not necessarily with a constant speed. The Spearman's rank correlation coefficient is based on the ranked values for each variable instead of the raw data. The Spearman's rank correlation coefficient is represented by the following formula:

$$\rho_S = \frac{\sum_i^n (r_1 - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i^n (r_1 - \bar{r})^2} \sqrt{\sum_i^n (s_i - \bar{s})^2}}, \tag{3.1}$$

where r_i and s_i represents ranking numbers of n data pairs, and the value ρ_S takes a value between -1 and +1.

The difference between the Spearman's correlation and the widely used Pearson correlation is that when both variables increase with different speed (but ranking stays the same), the Pearson correlation coefficient is then positive and less than +1, while the Spearman's coefficient is +1. Figure 3.15 represents the correlations between variables in the train set.

The sub-types of ski lifts, skiing slopes and coordinates of ski resort centers have a high and positive correlation with each other. This has to be kept in mind during the modelling part as some models are sensitive to highly correlated variables. There are some negative relations between elevation and ski slope information as well as positive relation between elevation and settlements, roads, but the correlation is still small.

Further, since the data has been explored and prepared for modelling, the considerations of the models used in this research will be described in the next chapter.

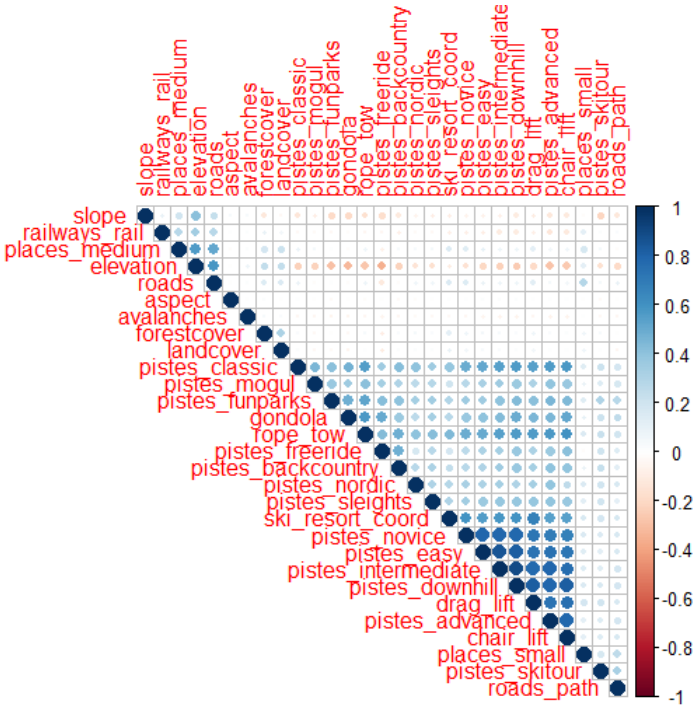


Figure 3.15: Spearman's rank correlation between variables in train set.

Chapter 4

Methodology

4.1 Predictive classification models

Predictive models are used to identify the variables that play a role in a certain problem and to be able to predict an unknown event. Various models can be used for specific problems. A possible model has to be considered. This research looks at a classification problem as the target variable in the dataset indicates two classes - an avalanche occurs or not. We are interested in the likelihood that this phenomena can happen in different locations, as well as in the variables that contribute and to what extent.

Standard linear regression does not work as the target variable is discrete. So the Generalized linear models (GLM) are considered. GLM are used when modeling discrete data, such as an avalanche occurrence. The outcome of the model gives a probability of the positive outcome, therefore, in this research, that is the likelihood of an avalanche to happen in a certain location. A similar model to GLM is the Generalized additive model (GAM). The difference between GAM and GLM is that GLM is based on linear relationships, while GAM can take non-linear relationships into account. Therefore, it is possible to recognize important patterns of a dependent variable as a function of one or more independent variables.

These two methods do not work that well, when the dataset is unbalanced, meaning that there are small number of cases for one class, and many observations from the other class. One of the suggestions in literature is to re-sample the dataset - down-sample the biggest class, up-sample the smallest class, or do both. But this approach does not really give good results. An autoencoder, which is a type of a neural network, is introduced. The next sections will describe all three of the mentioned classification models in more detail.

4.1.1 Generalized Linear Model

First, let's look at the GLM models. They were popularized by McCullagh and Nelde [26] in order to summarize and unify multiple statistical models such as linear regression, logistic regression and Poisson regression. The target variable Y_i in these models is assumed to follow a distribution from the exponential family with mean μ_i. This mean is assumed to be some function of $X_i^T\beta$.

GLM optimizes the weights of the factors by using the maximum likelihood method. GLM models are built from three components:

1. random component - the probability distribution of the target variable Y (also called the noise or error model),
2. systematic component - specifies the linear combination of explanatory variables $X = (X_1, X_2, ...X_k)$:
$$\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}, \tag{4.1}$$
3. link function η_i or $g(\mu_i)$ - specifies the link between random and systematic components:
$$g(\mu_i) = \eta_i = g^{-1}(X_i^T\beta). \tag{4.2}$$

The link function describes, how the mean of the target variable $\mu_i = E(Y_i)$ relates to the linear predictor of explanatory variables. Different link functions define different relationships between the linear predictor and the mean of the distribution function.

In this research the target variable takes binary values - 1, if an avalanche has occurred in the location, and 0, if the cell is a non-avalanche location. The model with a binary response variable is called the binary logistic regression model or Logistic regression models with a binary link function. So a set of k explanatory variables $X = (X_1, X_2, ...X_k)$ tries to predict the likelihood of an avalanche occurrence.

Distribution of Y_i is assumed to be Binomial(n, p), where p is the probability of an avalanche, therefore, the random component is $Y_i \sim$ Binomial(n, p). The explanatory variables are $X = (X_1, X_2, ...X_k)$ and they can be both continuous and categorical. They are linear in their parameters, but the transformation of the X's themselves is allowed.

The link function is
$$\eta_i = g(\mu_i) = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right), \tag{4.3}$$
where the mean value of Binomial distribution is p. The link function models the log odds of probability of an avalanche as a function of explanatory variables. Binary logistic regression models are also known as logit models, when the predictors are all categorical.

4.1.2 Generalized Additive Model

From the analysis of the data in Section 3.3.2, it can be seen that the variables like elevation and slope steepness are not linear but have Gaussian properties. GAM takes a function f_j for each individual independent variable X_j, which makes this model more complex, but explains the response variable better.

GAM is a class of statistical models, where the usual linear relationship between the target variable and covariates is replaced by several non-linear smooth functions to model and capture the non-linearities in the data. GAMs are just a generalized version of linear models in which the predictors X_j depend linearly or non-linearly on some smooth non-linear functions like splines, polynomials or step functions.

The formula for GAM is similar to the GLM formula with the difference that the linear terms $\beta_j X_{ji}$ are replaced by more flexible functions $f_j(X_j)$, for $j = 1, \ldots, k$. The core of a GAM is still a sum of feature effects, but it gives the option to allow nonlinear relationships between some features and the output. Linear effects are also covered by the framework, because for features to be handled linearly, you can limit their $f_j(X_j)$ only to take the form of $\beta_j X_{ji}$. So, the regression function for GAM is

$$F(X) = Y_i = \beta_0 + f_1(X_{1i}) + \cdots + f_k(X_{ki}) + \epsilon_i, \tag{4.4}$$

where the functions $f_1, f_2, \ldots, f_k$ are different non-linear functions on variables $X = (X_1, \ldots, X_k)$ and ϵ_i is the random component. The Regression Function $F(X)$ gets modified in Generalized Additive Models, and only due to this transformation the GAMs are better in terms of generalization to random unseen data, fits the data very smoothly and flexibly without adding complexities or much variance to the model most of the times.

GAMs are parameterized just like GLMs, except that some predictors can be modeled non-parametrically in addition to linear and polynomial terms for other predictors. The probability distribution of the response variable must still be specified, and in this respect, a GAM is parametric. In this sense they are more aptly named semi-parametric models. A crucial step in applying GAMs is to select the appropriate level of 'smoother' for a predictor. This is best achieved by specifying the level of smoothing using the concept of effective degrees of freedom. A reasonable balance must be maintained between the total number of observations and the total number of degrees of freedom used when fitting the model (sum of levels of smoothness used for each predictor).

4.1.3 Autoencoder

The dataset in this research contains a very unbalanced response variable. There are many cells with no avalanche occurrences and only around 1000 cells with avalanches. We can look at this problem as an anomaly detection. A very common approach for the anomaly detection is to use a type of Neural Networks called the Autoencoder in order to reduce the dimensions of a regular, non-avalanche cell. When the pattern for the non-avalanche cells is defined, it can be used to detect anomalies such as the occurrences of avalanches.

An autoencoder is described as a Neural Network that is trained to reconstruct the data given as input. This method is a special case of Neural Networks, therefore, an autoencoder is trained using the same techniques as NN, usually gradient descent following gradients computed by back-propagation [16]. There are 4 types of autoencoders:

1. Vanilla autoencoder - contains only 1 hidden layer;
2. Multilayer autoencoder - contains multiple (more than 1) hidden layers;
3. Convolutional autoencoder - regularized multi-layer autoencoder;
4. Regularized autoencoder - using a loss function, encourages the model to not only copy the input to the output, but to also have other properties. There are two types:
   - sparse - learn features for another task (for example, classification);
   - de-noising - adds some noise to the input, and makes the autoencoder learn how to remove it.

Popular applications for autoencoders - dimensionality reduction to learn compact representation of data, One-Class Classification to recognise the biggest class in an unbalanced data problem, data de-noising. This research only looks at the vanilla autoencoder - a three-layer Neural Network, which is widely used for the anomaly detection. The three layers are:

- input layer - the input is a training set with cells from only the "regular"(non-avalanche) class;
- hidden layer - encodes the input x with a function $h = f(x)$;
- output layer - decodes the encoded input with a reconstruction function $\hat{x} = g(h)$,

where $\hat{x}$ is the reconstructed input x. This autoencoder minimises the distance (loss) function $L(x, \hat{x})$, which in this case is taken as Mean Squared Error (MSE):

$$L(x, \hat{x}) = \frac{1}{n} \sum_{i=1}^{N} (x - \hat{x})^2. \tag{4.5}$$

Finding a function $g(f(x)) = x$ seems easy, but the design of autoencoders is made to be unable to copy their input perfectly [16]. They are built to copy only the input similar to the training

data. A way how to make this work is to choose a smaller number of nodes in the hidden layer than the number of nodes in the input layer. In this way, the encoder compresses the representation of input data and is only able to optimally reconstruct input, which was seen frequently in the training data.

In other words, useful properties of the data are learned because the model needs to minimize the reconstruction error while retaining less information than the original data.

If this reconstruction error goes beyond a threshold, it means that the input was not well represented in the training data and therefore constitutes an outlier. In this research - a grid cell with an avalanche.

4.2 Evaluation measures

Many evaluation measures for classification problems have been described in the literature. Before looking for an optimal evaluation method, one has to understand the characteristics of the dataset, and has to address the problem of the certain research first. When the problem statement is clear, an optimal evaluation method can be introduced. In this research, the riskiness of every cell in the risk grid needs to be predicted. The evaluation measures considered in this research are the classification accuracy, Matthews correlation coefficient and Fraction skill score.

Accuracy is used as a statistical measure of how good the binary classification test correctly identifies or excludes a condition. The accuracy looks at the proportion of the true results among the total number of cases examined and can give misleading results, when the classes are not in the same proportions. The Matthews correlation coefficient gives more insight in how well the model predicts each classification class.

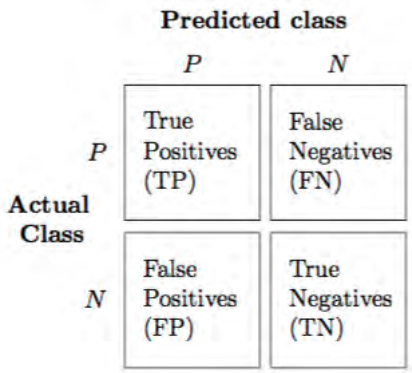
These two measures only look at the predictions for exact grids meaning that, if the model predicts a neighboring cell as risky, it is not taken into account for the evaluation. As this is a spatial problem, we want to look at the neighboring cells and reward the model, when the prediction has been made in a certain distance. For spatial prediction problems, the Fraction skill score will be introduced. The first two metrics are based on the confusion matrix. Therefore, this metric is briefly explained first.

4.2.1 Confusion matrix

In predictive analytics, the confusion matrix is used to represent the performance of a classification algorithm. The confusion matrix can be calculated for classification problems, where the target

variable consists of multiple classes, but the most common approach is to have the binary classification. That is also the case in this research, therefore the confusion matrix will be described for two classes.

The two class problem is called the binary problem. Each case of the target variable can be predicted correctly or incorrectly. If the latter is the case, the classification has predicted the other class. Every prediction of an observation can be classified in this manner and visualized in the confusion matrix, illustrated in Figure 4.1.



Source: `http://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/`
Figure 4.1: Confusion matrix for binary classification.

The confusion matrix is 2 x 2 table that report for numbers of the following metrics:

- True Positive (TP) - the number of data predicted as positive and the actual output is positive;
- False Positive (FP) - the number of data predicted as positive but the actual output is negative;
- False Negative (FN) - the number of data predicted as negative but the actual output is positive;
- True Negative (TN) - the number of data predicted as negative and the actual output is negative.

This distribution allows more insights and detailed analysis. When an unbalanced data set is used, the proportion of correct guesses (accuracy) can be misleading as the classifier might not correctly distinguish the difference between two classes and assign the most frequent class to all of observations. In this way, the confusion matrix is more reliable as it shows the real performance. This approach avoids misleading a high success rate of a classifier.

4.2.2 Accuracy

Classification accuracy is what people usually mean by accuracy. It is a proportion of correct predictions to the total number of input samples. When the confusion matrix has been introduced,

it can be used to explain accuracy. The accuracy formula for a binary classification problem is defined as follows:

$$Accuracy = \frac{\text{\# of correct predictions}}{\text{\# of total predictions made}} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4.6}$$

As mentioned before, the accuracy works well only if the data set is balanced, meaning that the target variable has to have a close to the equal number of observations belonging to each class.

4.2.3 Matthews correlation coefficient

To avoid making faulty conclusions out of accuracy results, the Matthews correlation coefficient (MCC) is introduced. This metric helps represent the results of the confusion matrix with a single value. MCC is a metric to represent the quality of a binary classifier and can even be used when the classes are of very different sizes. Basically, MCC shows a correlation between the actual outcomes of the target variables and predicted classes.

The Matthews correlation coefficient (MCC) can be directly calculated from the confusion matrix:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{4.7}$$

MCC can return a value in range [-1;1]. The value +1 represents a perfect prediction, MCC close to 0 means that the prediction is no better than random and -1 indicates total disagreement between prediction and observation [24]. By looking at the proportion of each class from the confusion matrix in formula 4.7, the coefficient will be high only when the classifier is doing well on both classes, that is - negative and positive elements.

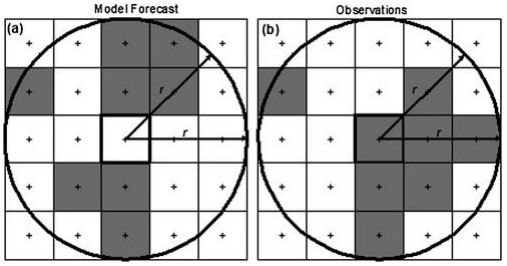
4.2.4 Fraction skill score

The previously mentioned evaluation methods use a double penalty. For example, one model predicts an avalanche in a grid cell close to the actual avalanche and another model does not predict the pattern at all. In the confusion matrix, first model would be punished twice as it does not have predicted avalanche where it supposed to be and, secondly, for predicting an avalanche where it is not supposed to be. This is called double penalty. The other model would be penalized only once - for not predicting the avalanche where it is supposed to be [31].

That is why the neighborhood verification methods have been developed. The fractions skill score (FSS) was introduced by Roberts and Lean in 2008. [30]. The binary events of a target variable within a small area in space are treated probabilistically rather than deterministically.

In this research - whether the probability of the cell exceeds a certain avalanche risk threshold or not [12].

For a given n x n window size, the FSS considers a perfect forecast to be one with the same frequency of events as observed within the window, regardless of their particular placement within the window. For example, consider a 5 x 5 window. The left square in Figure 4.2 illustrates a prediction/ forecast from the model and the right square represents the actual observations. The grey areas indicate the occurrence (1) and the white areas represent no occurrence (0). The model predicts 8 events in the window with radius $r = 2.5$. The actual observations are also 8. Although the events are not predicted in the same cells, the FSS would indicate a perfect prediction with value 1.



Source: 2012 Spring Forecast Experiment: Forecast Verification Metrics by NOAA HWT
Figure 4.2: Example to visualize fraction skill score.

The fractions skill score is computed as the fractions Brier score, divided by the sum of the mean squared forecast and observed fractions and can be written as:

$$FSS = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(p_{pred} - p_{obs})^2}{\frac{1}{N}\sum_{i=1}^{N} p_{pred}^2 + \frac{1}{N}\sum_{i=1}^{N} p_{obs}^2}, \tag{4.8}$$

where N is the number of cells in the domain (dependent on sliding window size), p_{pred} is the frequency of events in the window (probability) for prediction and p_{obs} is the observed fraction of the sliding window. The sum in the numerator in equation 4.8 is the fractions Brier score (FBS). This score indicates the squared difference between the predicted and the observed fractions. The FBS for a perfect forecast is 0 but the upper limit depends on the event frequency. Roberts and Lean ([30]) normalized the FSS with the FBS_{worst}, which is the largest possible FBS in the absence of an overlap of the nonzero observed and forecast fractions as defined by the sums in the denominator of equation 4.8. However, using FBS_{worst} endows the FSS with two very useful properties. First, it constrains the FSS to values between 0 and 1, and second it allows the FSS to be symmetric with respect to the fractional bias defined by p_{pred}/p_{obs}. [28]

The FSS can return a value from the range [0;1]. The value 1 represents a perfect prediction and 0 indicates no skill. As radius r of the window w expands and the number of cells in the neighborhood increases, the value of the FSS improves. The reason for this is that the observed and predicted probability fields are smoother and the overlap increases.

Chapter 5

Results and evaluation

This chapter presents the results obtained with the models and evaluation metrics specified in Chapter 4. First, Section 5.1 presents the results obtained from the GLM model. Then, Section 5.2 explains the results obtained with a GAM model. Finally, in Section 5.3 the results from the autoencoder are presented and explained. Every section contains considerations on the performance of the models using the evaluation metrics.

The only restrictions for the GAM and GLM models with binary response variable is that the observations has to be independent. Intuitively, only the neighboring areas (cells) have similar values for variables like elevation, forest cover, etc. But that is not always the case. Most of the ski resorts are located in similar altitudes, but still in different areas and can have similar characteristics of landscapes like slopes, lifts, etc. The order of the observations has been shuffled to meet this condition. Therefore, the assumption that the observations are independent has been met.

5.1 GLM

First, the generalized linear model has been used to simulate the risk areas of the French Alps. The model estimates the weights of the geographical factors and they are used to predict the risks in the grid cells. The variable selection has been done using the forward selection approach:

1. 28 models with every variable alone has been created and their Akaike information criterion (AIC) has been calculated;
2. The variable from the model with the smallest AIC is taken as the first variable. Then, the determined variable from step 1 is used to create 27 separate models with each variable.
3. The covariate from the model with the smallest AIC is used as the second variable.

4. Compare the new model with the previous one to assess whether the more complex model is significantly better at capturing the data than the simpler model.
5. Repeat steps until none of the additional variables improves the model to a statistically significant extent.

To assess the statistical significance, the Anova Chi-squared test has been used. If the resulting p-value is sufficiently low (less than 0.05), a conclusion is made that the more complex model is significantly better than the simpler model, and thus favor the more complex model. If the p-value is not sufficiently low (greater than 0.05), the simpler model is statistically significant and that is the final model.

5.1.1 Factor weights in the model

Following the steps explained in the previous section, the final model with 5 variables was created. When the GLM models are built, the factors from the categorical variables are extracted as indicator variables of every category. The obtained output from the GLM is represented in Table 5.1. The numbers in the variable names, for example x_2, x_3, x_4, x_5 represent the buffer category explained in Table 3.5. The $*$ next to the p-value indicates the 5% significance of the variable.

Table 5.1: GLM final model output

Variable	variable value x_j	weights β_j	odd-weights e^{β_j}	p-value
Intercept		-9.01299	0.000121817	<2e-16*
elevation	x_1	2.08247	8.024296389	<2e-16*
pistes_intermediate2	x_2	-1.43804	0.237391726	<2e-16*
pistes_intermediate3	x_3	-0.48727	0.614301285	2.20e-08*
pistes_intermediate4	x_4	-0.04608	0.954962127	0.6761
pistes_intermediate0	x_5	-0.06688	0.935311656	0.4658
roads2	x_6	-1.04992	0.349966959	1.43e-15*
roads3	x_7	-0.20064	0.818211068	0.0348*
roads0	x_8	-0.14915	0.861439016	0.0417*
roads_path2	x_9	-0.72609	0.483795694	1.81e-12*
roads_path3	x_{10}	-0.12754	0.880262300	0.2914
roads_path0	x_{11}	0.13620	1.145909636	0.3190
slope	x_{12}	0.02025	1.020460079	9.10e-12*

The weights β_j in Table 5.1 give the change in the log odds of the outcome for a one unit increase in the predictor variable. In order to explain and better interpret the coefficients, the odds-rations are computed and presented in the 4th column of Table 5.1. Now the interpretation of the influence of numerical variables can be done as follows: for every one unit (1 km) change in elevation, and keeping other variables at a fixed value, the odds of an avalanche occurrence (versus non-occurrence) increase by a factor of $\exp(2.08247) = 8.024$. Slope steepness is a significant variable, but the 1% change of the slope steepness increases the odds of avalanches

only by 1.02. The indicator variables for pistes_intermediate, roads and roads_path have a slightly different interpretation. For example, the cell in the distance of buffer 2 (0-250m) from the pistes_intermediate versus an actual piste with an intermediate difficulty (buffer 1), changes the odds of an avalanche occurrence by 0.237391726.

An interesting pattern can be concluded from the weights - the proximity of people (like pistes, roads and roads_path) reduces the probability of avalanches happening in the area of these locations as the weights are negative (and odd-weights are below 1). Another fact that the model shows - the buffer 4 of the distance 2000-3500m from the intermediate difficulty skiing slope has a very close coefficient with the cells that are further than 3500m from the intermediate piste. This indicates that the approach of making a buffer for such a distance is not significant when predicting avalanche occurrences.

Every grid cell has a value for every factor. Now, taking the weights β_j and factors x_j from the Table 5.1, the riskiness of every grid cell i can be calculated by using the following formula:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_1 2 x_{12i}. \tag{5.1}$$

The performance of the model will be described in the next section.

5.1.2 Predicted riskiness of the area

To calculate the risk of avalanches, the following comparison must be applied:

$$\eta_i = log\left(\frac{\mu_i}{1-\mu_i}\right). \tag{5.2}$$

The value η_i for every i is calculated using the equation 5.1, therefore the μ_i, which is the probability of an avalanche occurrence, can also be calculated. Figure 5.1 represents these values (y-axis) for every cell (x axis) both for train (left) and test (right) set.

As there is no clear indication of when probabilities for avalanche cases are higher than non-avalanche cases, the top 5% probabilities value has been chosen as a cutoff. This is a value of 0.019. The red line in the Figure 5.1 indicates the threshold of 0.019. We say that all the cells with the predicted probability higher than this value are indicated as an avalanche danger. This has to be done in order to use the evaluation metrics described in the Section 4.2. It can be seen that many cells have a small chance of an avalanche occurrence which is logical, as there has been only a relatively small number of cells with actual avalanches. In the meantime, there are also outliers like the one with value 0.27. When investigating this cell closer, there have not been avalanches. This indicates the complexity of the research problem as the included features can not perfectly model the avalanche danger.

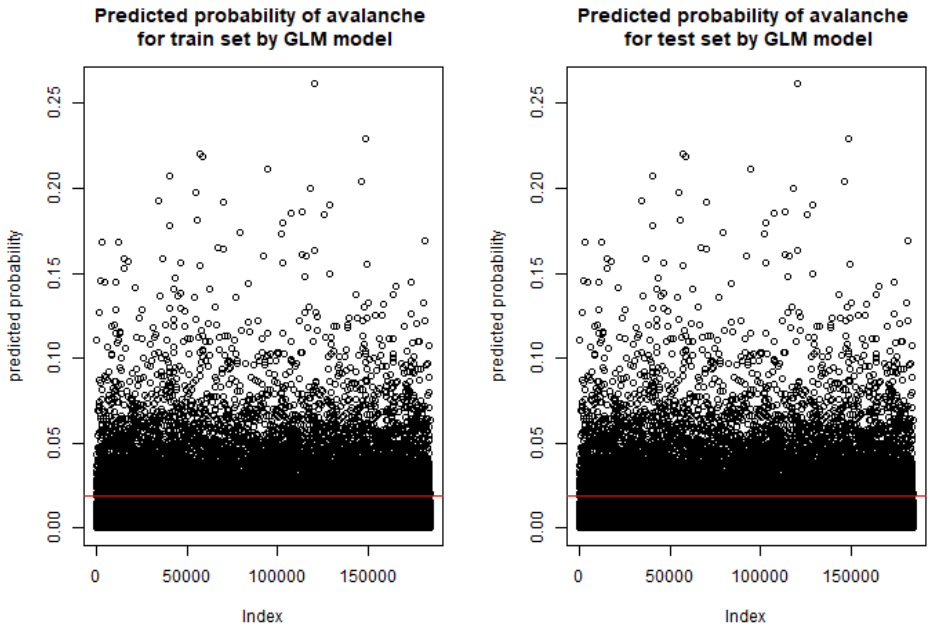


Figure 5.1: Predictions from the trained GLM model for train and test set with the threshold of 0.019 (red line).

An overview of the predicted locations and the actual avalanches are illustrated in Figure 5.2. The model reasonably recognizes the risky areas on a bigger scale. However, the focus in this research is on a smaller scale - areas of 500 x 500 meters and the applicability of these results for the usage in the field will be discussed in the Chapter 6.

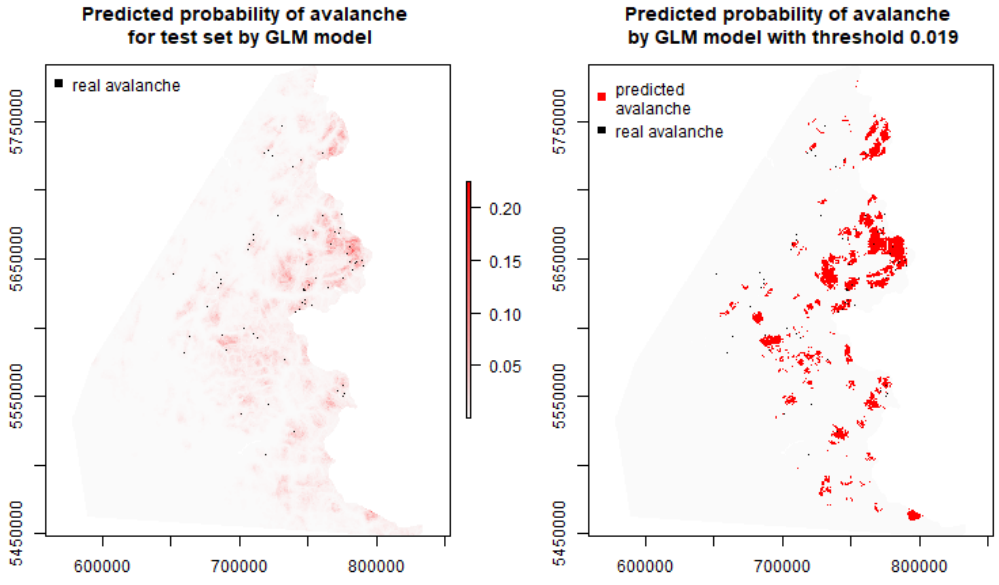


Figure 5.2: Predictions from the trained GLM model for the test set in the French Alps: the predicted values and actual avalanches (left), the areas of probabilities higher than the threshold 0.019 (right).

The illustration on the right side of Figure 5.2 shows the predicted avalanches with probability higher than the threshold. The indicated avalanches are often next to each other as the characteristics of the landscape for these cells are similar. The confusion matrix of these results are presented in the Table 5.2.

Table 5.2: Confusion matrix of the test set for the GLM model

		Predicted	
		0	1
True	0	175100	9124
	1	141	95

The confusion matrix shows that there are many True Negatives in the model. There are 9124 False Positives, which is partly due to the low threshold value. The number of correctly predicted avalanches is 95, but the model could not recognize 141 avalanches, therefore, these are classified as False Positives.

The evaluation metrics are summarized in Table 5.3.

Table 5.3: GLM model performance summary

Evaluation metric	Value
Accuracy	94.98%
MCC	-0.029
FSS	0.0211
FSS, w=1	0.0257
FSS, w=2	0.0269
FSS, w=3	0.0275
FSS, w=4	0.02817

The accuracy of the GLM model is high as a big proportion (94.79%) of the cells without avalanches have been classified correctly. The MCC gives a value below zero indicating that the model is bad in classifying both of the classes correctly. As explained previously, Accuracy and MCC are based on the confusion matrix and these metrics do not take into account the neighboring cells, therefore they penalize the outcome twice. That is why the FSS was introduced. The FSS for the cell itself, a 3 x 3 cell moving window (w=1), 5 x 5 cell moving window (w=2), 7 x 7 cell moving window (w=3) and 9 x 9 cell moving window (w=4) has been calculated. Despite the fact that FSS increases when the moving window increases, the difference is still small. The value of 0.028 is a low score for such a metric as it is close to 0. This means that the model struggles to identify the weights from the factors and to predict avalanche cases in the test set. The potential improvements for the application will be discussed in the next chapter.

5.2 GAM

As the GAM model is a generalized version of the GLM models, the same forward selection approach as described in Section 5.1 has been used for building a GAM model. Only this time, for the numerical variables elevation and slope, different splines have been fitted. And again, the variable from the model with the smallest AIC has been used in the forward selection until the more complex model does not significantly improve the results.

5.2.1 Factor weights in the model

The obtained model output is represented in Table 5.4. The $*$ indicates the 5% significance.

Table 5.4: GAM final model output

Variable	variable function f_j	weights β_j	odd-weights e^{β_j}	p-value
Intercept		-6.14333	0.002147768	<2e-16*
pistes_downhill2	$\beta_2 x_2$	-1.25755	0.284349387	<2e-16*
pistes_downhill3	$\beta_3 x_3$	-0.22477	0.798702843	0.00241*
pistes_downhill0	$\beta_4 x_4$	0.01089	1.010954181	0.88117
roads2	$\beta_5 x_5$	-1.13311	0.322029931	<2e-16*
roads3	$\beta_6 x_6$	-0.05811	0.943545851	0.54994
roads0	$\beta_7 x_7$	-0.10215	0.902889597	0.16710
roads_path2	$\beta_9 x_9$	-0.65298	0.520490453	3.02e-10*
roads_path3	$\beta_{10} x_{10}$	-0.17421	0.840121709	0.14967
roads_path0	$\beta_{11} x_{11}$	0.07665	1.079660565	0.57551
Smooth terms	variable function f_j	edf		p-value
s(elevation)	$f_1(x_1)$	3.593		<2e-16*
s(slope)	$f_8(x_8)$	5.103		1.98e-08*

The categorical variables are explained with linear functions $\beta_j x_j$, but the numerical variables are non-linear functions. First, lets look at the linear terms. These are explained in the same manner as the results from the GLM model. To better interpret the odds (instead of log-odds), the weights for odds are computed and presented in the 4th column of Table 5.4. The base case for categorical variables are the cells that contain the actual factor - downhill piste, road or a small road (path). Every category 2,3 and 0 indicate the distances with buffer values explained in Table 3.5. The model compares these categories to the reference case and the p-value shows, if the probability is different from the category 1 or not.

The GAM model shows that, if the area is in the distance of 0-1700m (buffer 2) from a downhill slope, the odds from the reference case changes by 0.284. If the area is in the distance of 1700-3600m (buffer 3) the odds from the reference case change by 0.799. As these values are smaller than 1, the odds are against the avalanche, therefore these factors are reducing the

probability of an avalanche occurrence. If the area is further than 3600m away from the downhill slope (buffer with a value 0), the odds of an avalanche occurrence in this location are not different from the odds, if there would be a downhill slope in this location. In the same way the roads and roads_path can be explained. An interesting finding is that the buffer 3 for roads and roads_path is not significant and the further the location from the factors, the odds of an avalanche occurrence reduce. Another interesting finding is that the roads_path buffer 1 and buffer 2 could have been combined, because these two categories explain the same partial, as the Figure 5.3 presents.

Now, let's look at the non-linear terms. The edf in Table 5.1 for numerical variables stands for effective degrees of freedom. This value represents the complexity of the smooth. Edf of 1 is equivalent to a straight line, 2 is equivalent to a quadratic curve, and so on, with higher edfs describing more wiggly curves. The visual representation for all 5 variables is given in the Figure 5.3. Here, the elevation is approximated with the edf just above 3.5 indicating that the fit is similar to a GLM with a 4th order polynomial function

$$f_1(x_1) = a_0 + a_1 x_1 + a_2 x_1^2 + a_3 x_1^3 + a_4 x_1^4, \tag{5.3}$$

and the slope steepness is approximated with a spline similar to a 5th order polynomial

$$f_8(x_8) = b_0 + b_1 x_8 + b_2 x_8^2 + b3 x_8^3 + b3 x_8^4 + b4 x_8^4 + b_5 x_8^5. \tag{5.4}$$

Both of these non-linear functions are statistically significant when predicting the log-odds of avalanches. Unlike the 'traditional' regression, the coefficients can not be interpreted or expressed as a formula, but the fit can be visualised.

5.2.2 Predicted riskiness of the area

Figure 5.4 illustrates the probabilities of avalanches for every cell in the train (left) and test (right) sets. The red line indicates the threshold of 95% of the predicted values. The assumption has been made - the cells with the prediction higher than the value 0.021 are indicated as risky, but otherwise as non-risky. Again, there are some outliers that show a high likelihood of avalanches, while it is not true.

Figure 5.5 illustrates the risky locations in the right plot, but the predicted values on the left figure. The black dots are the actual avalanches in the test set.

The performance of the GAM model is very similar to the previously described GLM model. The slope and elevation approximation with a function instead of a straight line, has improved the results. Also, the pistes_downhill instead of the pistes_intermediate as a second variable was used for the GAM model.

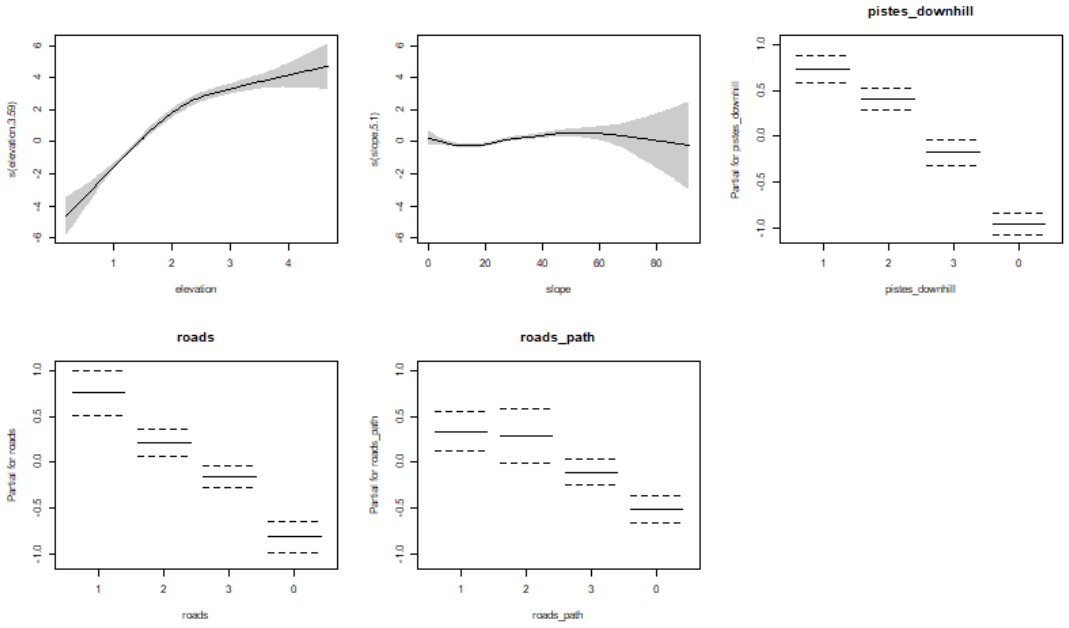


Figure 5.3: Variable function representations of the GAM model.

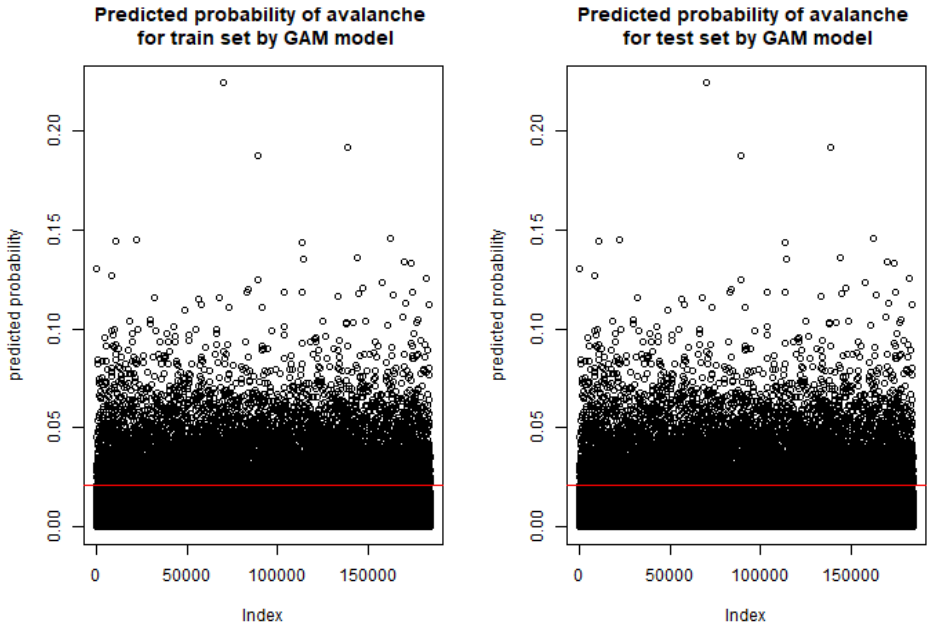


Figure 5.4: Predictions from the trained GAM model for the train and the test set with threshold of 0.021 (red line).

Table 5.5: Confusion matrix of the test set for the GAM model

		Predicted	
		0	1
True	0	175392	8832
	1	138	98

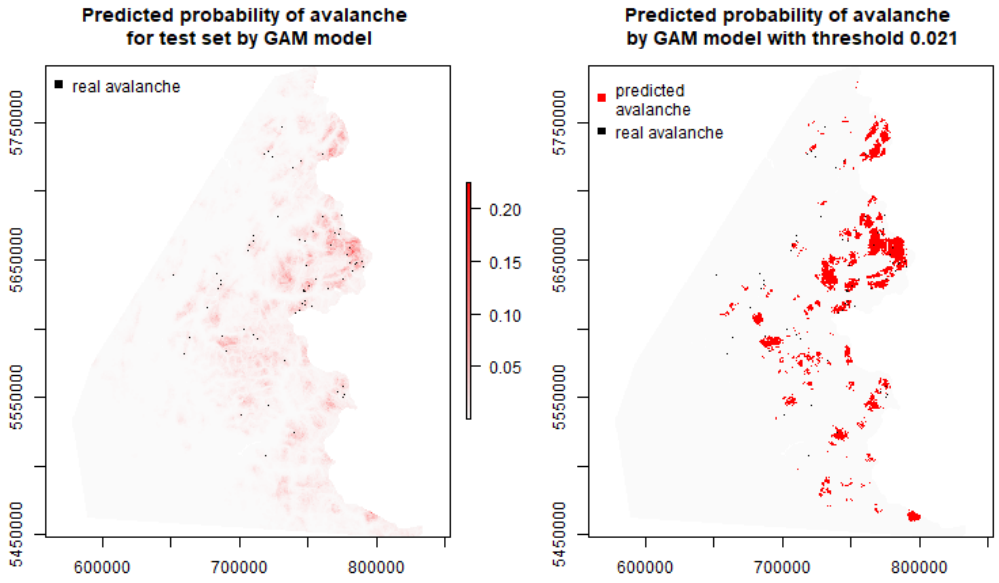


Figure 5.5: Predictions from the trained GAM model for the test set in the French Alps: the predicted values and actual avalanches (left).

When inspecting the confusion matrix in Table 5.5, GAM model has learned to better classify True Negatives as the number of FN has reduced. The number of predicted TP clases has improved by 3 cases. Overall, the GAM model gives better classification results than the GLM. When inspecting the evaluation metrics, accuracy of the GAM model is higher than the GLM, but the MCC is even lower. The negative MCC value indicates that the classifier is not doing well on both of the classes. These metrics have been summarized in the Table 5.6.

Table 5.6: GAM model performance summary

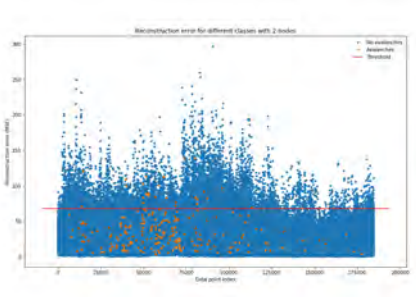
Evaluation metric	Value
Accuracy	95.14 %
MCC	-0.031
FSS	0.0214
FSS, w=1	0.0264
FSS, w=2	0.02764334
FSS, w=3	0.02839876
FSS, w=4	0.02908669

In theory, FSS has to improve when the moving window size gets bigger. This property is present in the described GAM model. The values of the FSS for different moving window sizes are still small. This can be explained with the used probability threshold of 0.021, when indicating the risky avalanche areas. The real number of avalanches in the test set is 236, whereas the GAM model predicts 8930 avalanche locations. If the threshold would have been chosen higher, the FSS would improve as there would be a smaller number of faulty avalanche predictions.

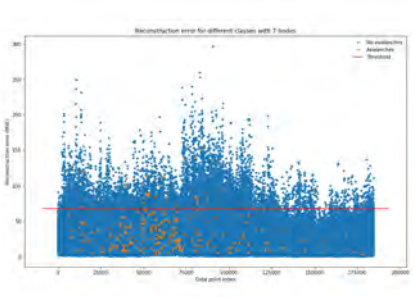
5.3 Autoencoder

The results of the Autoencoder in terms of significance of the predictors are impossible to explain. This is a consequence of the so-called black box algorithm. In the literature, Autoencoder is mentioned as an often-used anomaly detection method. The model was trained as a 3 layer Neural Network: an input layer, 1 hidden layer and an output layer. The first layer is an input layer with 28 nodes - the variables described in Section 3.3.3. The number of nodes in the hidden layer were changed in order to tune the algorithm. The results from autoencoders with 2,7,10 and 26 nodes in the hidden layer are summarized in this section.

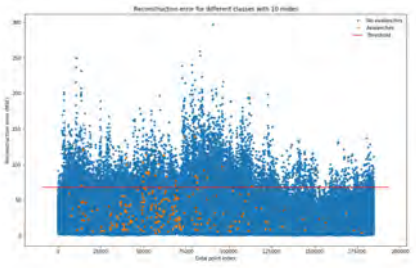
To train the Autoencoder, only the grid cells with no avalanches from the train set were used. In this way, Autoencoder had to find the properties of non-avalanche cells to reduce the dimension and to be able to reconstruct the same pattern. In case of a well trained Autoencoder, the reconstruction error for the avalanche cells from the test set should be higher. The Mean Squared Error was used as the Loss function. The obtained MSE for different sizes of the hidden layers are summarized in Figure 5.6.



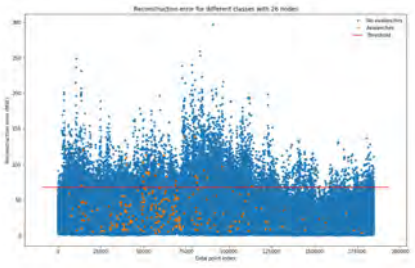
(a) 2 nodes



(b) 7 nodes in the hidden layer



(c) 10 nodes in the hidden layer



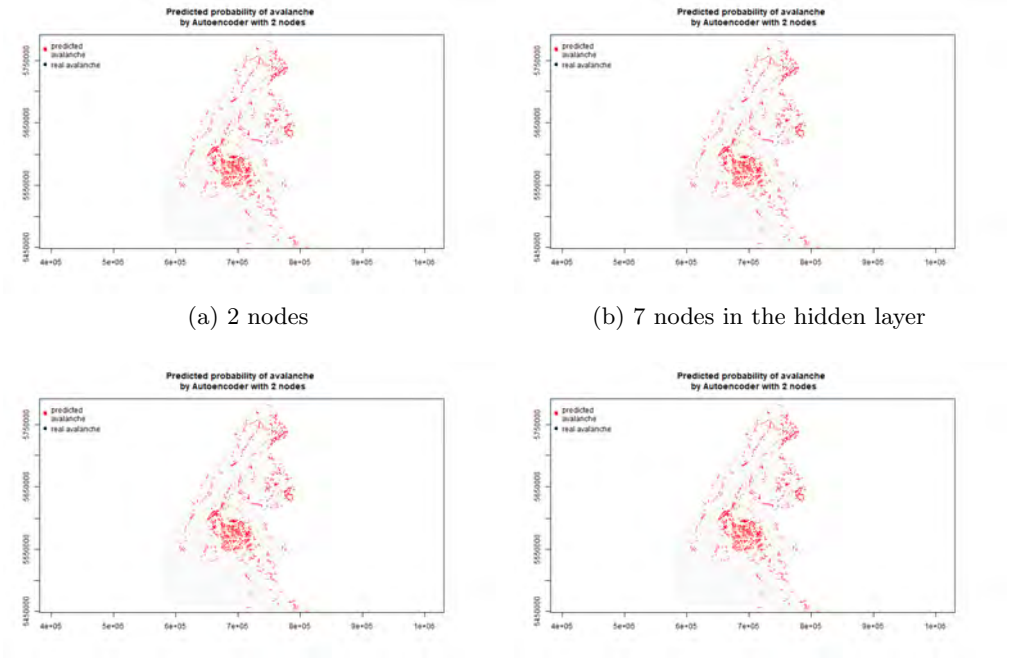
(d) 26 nodes in the hidden layer

Figure 5.6: The MSE of the Autoencoder with 2,7,10 and 26 nodes in the hidden layer - the trained non-avalanche cells (blue) and avalanche cells from the test set (yellow).

We can see that the different number of nodes does not give a big difference in terms of reconstruction error. The plots even look identical. The yellow dots in the graphs represent the avalanche cases from the test set. The Autoencoder should be able to recognize the differences of characteristics for different response classes. As can be seen in Figure 5.6, the positive observations

do not have a specifically higher MSE. The red line indicates a threshold of MSE=75. The observations with higher MSE are all classified as risky avalanches.

The potential risky areas predicted by the various Autoencoders with the threshold MSE=75 are visualized in Figure 5.7.



(a) 2 nodes

(b) 7 nodes in the hidden layer

(c) 10 nodes in the hidden layer

(d) 26 nodes in the hidden layer

Figure 5.7: Predictions from the Autoencoder with different numbers of nodes for the French Alps: the predicted spots and the actual avalanches.

To be able to use the evaluation metrics, the confusion matrices for various autoencoders with different node sizes are computed and presented in Table 5.7.

Table 5.7: Confusion matrices for the Autoencoder with 2, 7, 10 and 26 nodes in the hidden layer.

		Predicted	
		0	1
True	0	174985	9239
	1	207	29

		Predicted	
		0	1
True	0	175028	9196
	1	208	28

		Predicted	
		0	1
True	0	175032	9192
	1	207	29

		Predicted	
		0	1
True	0	175053	9171
	1	208	28

As one would expect from the illustrations above, the results are very similar for the 4 Autoencoders. There are slight differences in the False Negatives but correctly predicted number of avalanches are almost the same. From these 4 models, the Autoencoders with 10 and 26 nodes performs slightly better than the other two.

Now, let's look at the evaluation metrics in Table 5.8. Again, the results are very close for all of the models. When comparing the results from the created GLM and GAM models, the first two outperform the Autoencoder in terms of the FSS. The MCC is positive only for the Autoencoder. It means that the Autoencoder is a better classifier than the GLM and the GAM. This indicates that the Autoencoder captures some properties of non-avalanche cells during the dimensionality reduction that generalized models could not explain. Unfortunately, Autoencoder is a black box algorithm and it is hard to extract, what are these properties. Also, the value is just above the 0, meaning that the classifier is still weak.

Table 5.8: Performance summary for Autoencoders with 2, 7, 10 and 26 nodes.

Evaluation metric	2 nods	7 nods	10 nods	26 nodes
Accuracy	94.88%	94.9%	94.9%	94.92%
MCC	0.0119	0.0113	0.012	0.0113
FSS	0.0061	0.0059	0.0061	0.0059
FSS, w=1	0.0113	0.01118	0.0113	0.0112
FSS, w=2	0.0138	0.0137	0.0137	0.0137
FSS, w=3	0.0155	0.0155	0.0155	0.0155
FSS, w=4	0.0.017	0.017	0.017	0.017
FSS, w=5	0.0181	0.01814	0.0181	0.0182

The under-performing results from the Autoencoder indicate that the variables used in the model are not well describing the avalanche occurrences and the dimension reduction does not really help. Adding multiple hidden layers to the Autoencoder could improve the model, but that would make it even more challenging to explain the characteristics of the risky areas. The suggested improvements for the modelling will be described in the Discussion section.

Chapter 6

Conclusion

The aim of this master's thesis was to investigate, how accurately we can predict the locations of avalanches based on historical avalanches and available open source data. This section answers the sub-questions and the research question presented in Section 1.3. Each sub-question is answered individually and contributes to answer the main question of this research.

What are the main variables that increase this likelihood and accuracy?

Avalanche literature is quite consistent on a few factors, but the opinion for some factors differs from source to source. The significant variables selected by the models match the literature expectations. The main variables to increase the prediction accuracy are elevation of the landscape, proximity of a downhill skiing piste, highways, small paths of the roads and slope steepness. The importance of each of the variables is explained in more detail in Section 5.2. The predictive models were trained on a dataset with historical avalanche occurrences until 01-01-2017. The test set contained avalanches from 01-01-2017 until 02-12-2018.

How these environmental factors should be used?

Elevation and slope steepness are used as a numerical variable. 1 unit of elevation explains 1 km. Numerical values for slope steepness indicate the inclination of the area in percentage. This is done in order to make the output of predictive models more interpretive. The variables - downhill skiing piste, highways, small paths and roads - were transformed so that they can be used as ordinal variables. An exploration of distances between historical avalanches and these factors was performed and depending on the findings, buffers with several distances were created to indicate a proximity of the factor even when the location does not contain the factor. The buffer sizes are summarized in Table 3.5.

What are the recommended methods in this field of subject?

This research looked at a binary classification problem as the target variable in the dataset indicates two classes - an avalanche occurrence or non-occurrence. As the research focuses on discovering the contributing factors to this phenomena and predicting a likelihood of certain locations, two predictive classification models were used to cope with this problem - Generalized Linear model and Generalized Additive model. A type of Neural Network called the Autoencoder was introduced in order to deal with an unbalanced target variable and to make the predictions better.

What validation criteria apply best for this specific problem?

The models were trained on the training set, but performance of these models was evaluated by means of new avalanche data. A literature review was performed in order to find the best evaluation method. Three different performance measures were considered - accuracy, Matthews correlation coefficient and Fraction skill score. Accuracy is the most common binary classification evaluation method. This measure indicates a proportion of correctly predicted locations with the total number of predictions, but the results are misleading when the dataset is unbalanced. To cope with this and understand better, if the model performs well on both binary classes, Matthews correlation coefficient is used. As these two methods evaluate only the exact grids and penalizes the measure twice, when the predictions are incorrect, a neighborhood verification method is introduced. The neighboring cells are rewarded when the model has predicted the same number of occurrences in the neighborhood as the test set contains. The Fraction Skill score is used to evaluate models and make conclusions of their performance.

Which risk model fits the current problem the best?

During the data exploratory phase, elevation and slope steepness indicated a non-linear relationship with the target variable, therefore, GAM model was expected to perform better. The MCC is worse for GAM, but the FSS of GAM gives a slightly better result than GLM. The Autoencoder with 1 hidden layer was expected to perform better than the linear models as it is widely used in dealing with unbalanced response variable. The MCC was indeed positive and higher than the generalized models indicating that the classifier is better. The FSS for the Autoencoder with a different number of nodes in the hidden layer was worse than both GLM and GAM. Therefore, the GAM model with factors - elevation of the landscape, proximity of a downhill skiing piste, highways, small paths of the roads and slope steepness - fitted the current problem best.

How applicable are the results for the usage in the field?

The result of this research is a good first indication of the risky areas and the contributing factors for avalanche initiation. The model is simple, traceable and the risk map is intuitive.

Adding more geospatial factors like snow depth and temperature for the time frame before the incident has happened, would improve the model significantly. However, the whole modelling area is too big to make these models applicable for the application, the grid cells are too wide, and the probabilities are still too small to be reliable. Indicating an area of 3 km^2 (12 cells) as risky does not help to guide the drone and to save the victim faster as that would still take a lot of time. The set thresholds are too low to be accurate. This could be improved by adding more explanatory variables.

Summarizing the answers to the sub questions, the conclusion can be made for the research question:

How accurately we can predict the location of an avalanche/victim caught in an avalanche, based on the historical and available open source data?

The historical avalanche dataset does not contain as much victim data as one would hope for. Therefore, an assumption that every observation from the historical avalanche dataset contained a victim has been made. The reasoning behind was that these avalanche occurrences were reported by people and they could have been caught in the avalanche. This is not necessarily true, but the number of observations with victims was too small to build a predictive model. As the geospatial data was collected open source, the quality of geospatial factors is not clear - it might have missing data. Also, when predicting a natural disaster like an avalanche, it is important to keep in mind that there is a high variety of coincidence which can not be predicted. It is simple to recognize a terrain which could be a potential place for an avalanche, but to recognize when exactly this same terrain will have a snow avalanche is nearly impossible.

Taking all the arguments into consideration, a predictive model and a risk map was created. The Fraction skill score was the highest for a GAM model, but the value 0.028 for a moving window of 7 x 7 cells ($w = 3$) is still a low score. It is close to 0 meaning that the model struggles with recognizing the avalanches for the test set. The potential improvements for the application will be discussed in the next chapter.

Chapter 7

Discussion, reflection and recommendations

The research experienced some challenges and several limitations towards achieving better results. First of all, the data was acquired via open source so the reliability and quality of the data remains unclear. During the literature review, three types of contributing factors were mentioned - terrain characteristics, snow-pack, and meteorological conditions. Acquiring weather and snow pack data matching the same timeline and location from open sources proved to be a challenging task. Therefore, the research focused only on the terrain characteristics as these factors are static and do not change over the time. The exploration of avalanche occurrences indicated a seasonal pattern. The suggestion for further work is to acquire information on the dynamic factors and together with time include them to improve the models.

Some of the factors used for the modelling turned out to be insignificant. Another approach can be introduced - instead of creating buffers around factors and using them as ordinal variables, create raster files with values in meters for a cell that indicates distance to the closest factor. For example, the value for each cell means how far, in meters, the nearest ski resort is. The cells that are located within the ski area gets a score of "0". This would transform all the contributing factors to numerical, and other potential classification models could be used.

Another consideration regarding the data - the size of grid cells in the research were fixed to 500 x 500 meters. A manipulation with different grid sizes could lead to different results. This would only change if the acquired data sets would have small enough resolution and the values to the grid cells would significantly change. The response variable in the research is highly unbalanced, only 0.27% of the observations contain an avalanche occurrence. A suggestion for future work is to focus more on a specific area, like a ski resort and perform this analysis. Then the cell sizes can also be smaller and more applicable for the application.

In order to deal with the unbalanced problem, Autoencoder was introduced. The model was not performing better than the generalized models. Since the algorithm is a type of a Neural Network and, therefore, a black box algorithm, the importance of the contributing factors is impossible to interpret. This research only used three algorithms, while there might be other, better performing algorithms for the specific problem, taking other considerations into account. A further research needs to be addressed in order to deal with unbalanced, spatial data.

Finally, the risk map has been created to guide the drone to the risky areas. As there might be multiple areas with the same level of riskiness, a metric that evaluates the ranking of the predictions could be used. Conceptually, set a threshold that indicates the rank of the number of consecutive grids the drone has to fly, and if the real avalanche is within the first number of cells, the model is good. Future research could include this metric when evaluating the performance of the models as this would compliment the results for the application.

Bibliography

[1] *Avalanche knowledge and prevention portal*. Available online, accessed on 19-03-2019. URL: https://www.whiterisk.ch/en/explore#u=03-02-02.

[2] J. A. Ballesteros-Cánovas et al. "Climate warming enhances snow avalanche risk in the Western Himalayas". In: *Proceedings of the National Academy of Sciences* 115.13 (2018), pp. 3410–3415. ISSN: 0027-8424. DOI: 10.1073/pnas.1716913115. eprint: https://www.pnas.org/content/115/13/3410.full.pdf. URL: https://www.pnas.org/content/115/13/3410.

[3] National Avalanche Center. *Avalanche Encylopedia: Avalanche*. Available online, accessed on 14-03-2019. URL: https://avalanche.org/avalanche-encylopedia/#avalanche.

[4] National Avalanche Center. *Avalanche Encylopedia: Slab*. Available online, accessed on 13-03-2019. URL: https://avalanche.org/avalanche-encylopedia/#slab.

[5] National Avalanche Center. *Avalanche Encylopedia: Wet snow avalanche*. Available online, accessed on 19-03-2019. URL: https://avalanche.org/avalanche-encyclopedia/wet-snow-avalanche/.

[6] Data-Avalanche.org. *Historical avalanches*. data retrieved by e-mail from data-avalanche website, http://www.data-avalanche.org/. 2018.

[7] ESRI. *ArcGIS Resource Center, How Aspect works*. Available online, accessed on 2019-06-18. URL: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/How_Aspect_works/009z000000vp000000/.

[8] ESRI. *ArcGIS Resource Center, How Slope works*. Available online, accessed on 2019-06-18. URL: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/How_Slope_works/009z000000vz000000/.

[9] ESRI. *ArcObjects SDK 10 Microsoft .Net Framework – Working with spatial references*. Available online, accessed on 2019-06-17. URL: http://help.arcgis.com/en/sdk/10.0/arcobjects_net/conceptualhelp/index.html#//0001000002mq000000.

[10] ESRI. "ESRI shapefile technical description". In: *Comput. Stat* 16 (1998), pp. 370–371.

[11] GMES RDA project (EU-DEM). *Digital Elevation Model over Europe (EU-DEM)*. data retrieved from European Environment Agency, `https://opendem.info/OpenDemEU/getData4258.jsp?xmin=4.879882812500005&xmax=16.349609375000007&ymin=42.46855468750001&ymax=49.58769531250001`. 2018.

[12] NATHAN Faggian et al. "Fast calculation of the fractions skill score". In: *Mausam* 66 (2015), pp. 457–466.

[13] Chiba University Geospatial Information Authority of Japan. *Land cover classified in 20 categories*. data retrieved from Land Cover (GLCNMO) - Global version website, `https://globalmaps.github.io/glcnmo.html`. 2019.

[14] Michael Gilleland et al. "Levenshtein distance, in three flavors". In: *Merriam Park Software: http://www. merriampark. com/ld. htm* (2009).

[15] John Andrew Gleason. "Terrain parameters of avalanche starting zones and their effect on avalanche frequency". PhD thesis. Montana State University-Bozeman, College of Letters & Science, 1996.

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[17] Pascal Haegeli and Jürg Schweizer. "Recent developments in applied snow and avalanche research". In: *Cold Regions Science and Technology* 120 (2015), pp. 153–156.

[18] Jan Hauke and Tomasz Kossowski. "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data". In: *Quaestiones geographicae* 30.2 (2011), pp. 87–93.

[19] Edward R LaChapelle. "Snow avalanches: a review of current research and applications". In: *Journal of Glaciology* 19.81 (1977), pp. 313–324.

[20] Joel Lawhead. *Learning Geospatial Analysis with Python*. Packt Publishing Ltd, 2015.

[21] MapCruzin. *Shape files of France*. data retrieved from MapCruzin website, `https://mapcruzin.com/download-free-world-arcgis-shapefile/free-france-arcgis-maps-shapefiles.htm`. 2019.

[22] OpenSnowMaps (Open Street Maps). *Ski extracts from the openstreetmap database*. data retrieved from OpenSnowMap website, `http://www.opensnowmap.org/iframes/data.html`. 2019.

[23] Ski Maps. *Ski Resort Coordinates*. data retrieved from Ski Maps website, `https://skimap.org/SkiAreas/index.xml`. 2019.

[24] Brian W Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2 (1975), pp. 442–451.

[25] David McClung and Peter A Schaerer. *The avalanche handbook*. The Mountaineers Books, 2006.

[26] P. McCullagh and J.A.Nelder. *Generalized Linear Models, Second edition*. Chapman and Hall, 1983.

[27] Scott E McIntosh et al. "Cause of death in avalanche fatalities". In: *Wilderness & environmental medicine* 18.4 (2007), pp. 293–297.

[28] Jason E Nachamkin and Jerome Schmidt. "Applying a neighborhood fractions sampling approach as a diagnostic tool". In: *Monthly Weather Review* 143.11 (2015), pp. 4736–4749.

[29] Simon Nåfält. "Assessing avalanche risk by terrain analysis: an experimental GIS-approach to The Avalanche Terrain Exposure Scale (ATES)". In: *Student thesis series INES* (2016).

[30] Nigel M Roberts and Humphrey W Lean. "Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events". In: *Monthly Weather Review* 136.1 (2008), pp. 78–97.

[31] Andrea Rossa, Pertti Nurmi, and Elizabeth Ebert. "Overview of methods for the verification of quantitative precipitation forecasts". In: *Precipitation: Advances in Measurement, Estimation and Prediction*. Springer, 2008, pp. 419–452.

[32] WSL-Institut für Schnee- und Lawinenforschung SLF. *Avalanche danger level*. Available online, accessed on 15-04-2019. URL: `https://www.slf.ch/en/avalanche-bulletin-and-snow-situation/about-the-avalanche-bulletin/danger-levels.html`.

[33] WSL-Institut für Schnee- und Lawinenforschung SLF. *Avalanche types*. Available online, accessed on 15-03-2019. URL: `https://www.slf.ch/en/avalanches/avalanche-science-and-prevention/avalanche-types.html`.

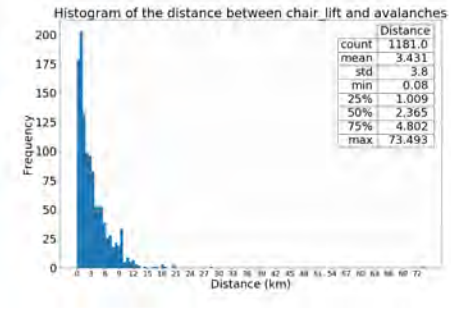
[34] Jürg Schweizer, J Bruce Jamieson, and Martin Schneebeli. "Snow avalanche formation". In: *Reviews of Geophysics* 41.4 (2003).

[35] Science and knowledge service of the European Commission. *Spatial distribution of European forests (Forest Cover Map)*. data retrieved from The European Commission's science and knowledge service website, `https://forest.jrc.ec.europa.eu/en/past-activities/forest-mapping/`. 2006.

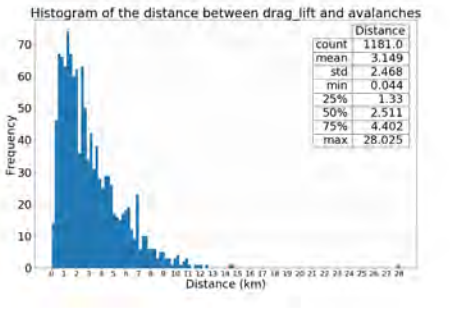
[36] Frank Techel et al. "Avalanche fatalities in the European Alps: long-term trends and statistics". In: *Geographica Helvetica* 71.2 (2016), pp. 147–159.

Appendix A

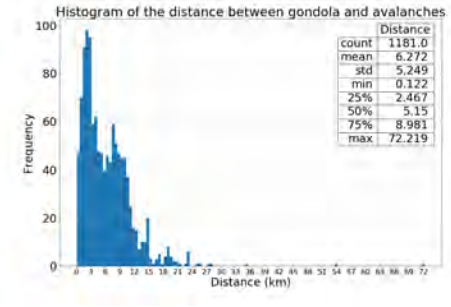
Appendix



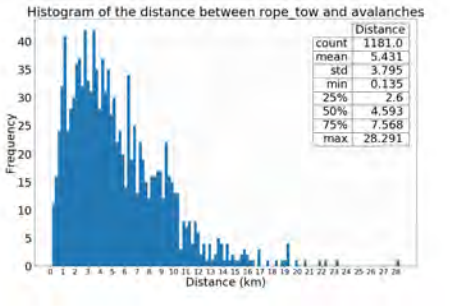
(a) Chair lift



(b) Drag lift

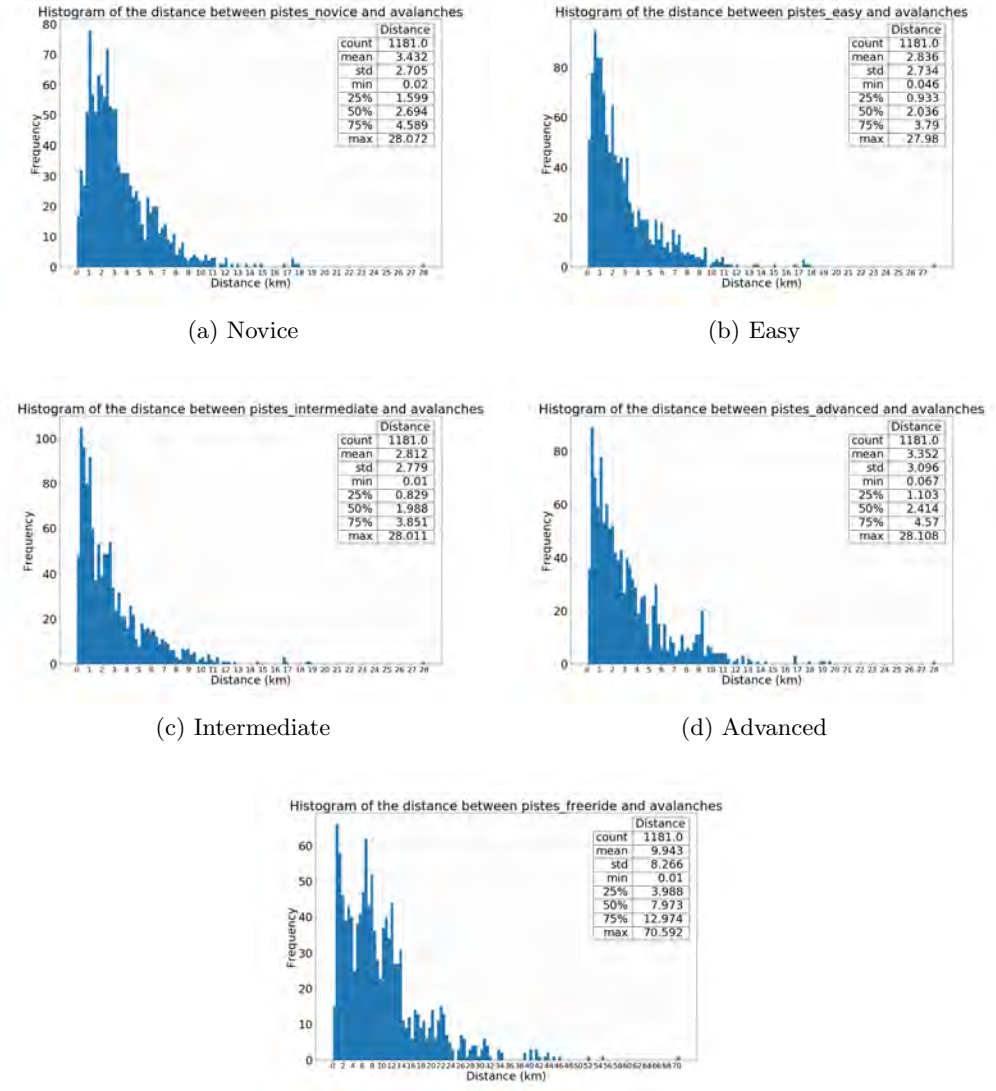


(c) Gondola



(d) Rope tow and magic carpet

Figure A1: Histograms of the distances between types of lifts and avalanches.



(a) Novice

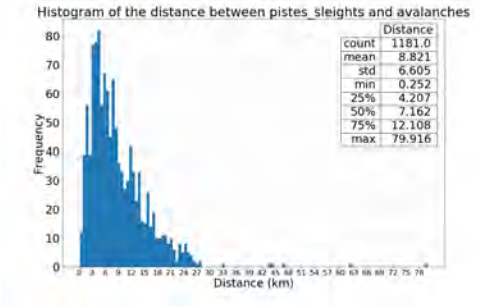
(b) Easy

(c) Intermediate

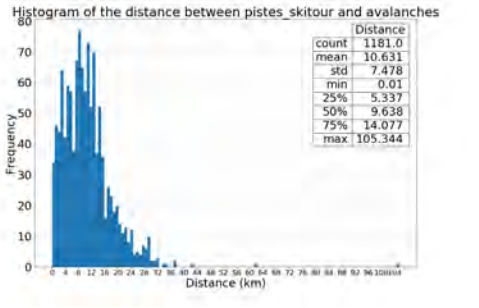
(d) Advanced

(e) Freeride

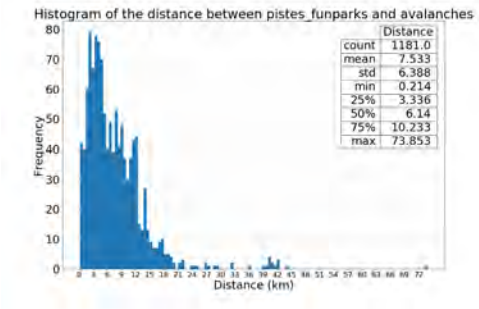
Figure A2: Histograms of the distances between ski piste difficulty and avalanches.



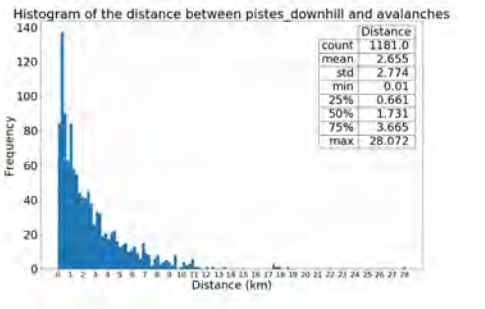
(a) Sleights



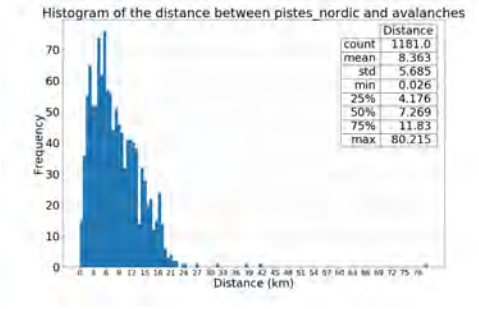
(b) Skitour



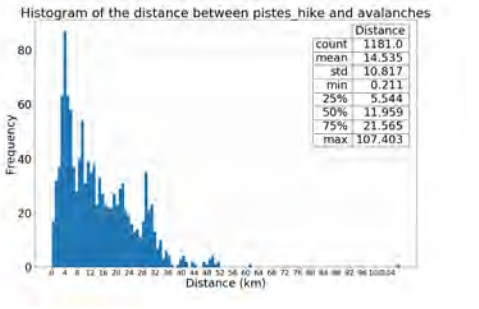
(c) Funparks



(d) Downhill

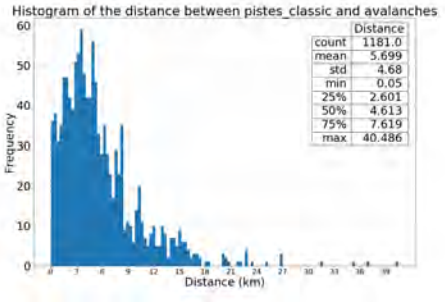


(e) Nordic

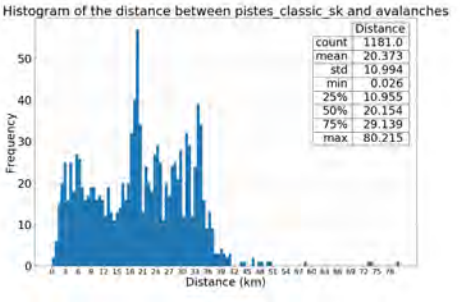


(f) Hike

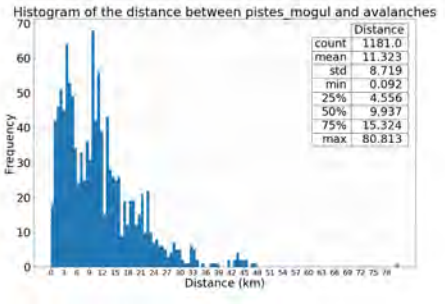
Figure A3: Histograms of the distances between ski piste types and avalanches.



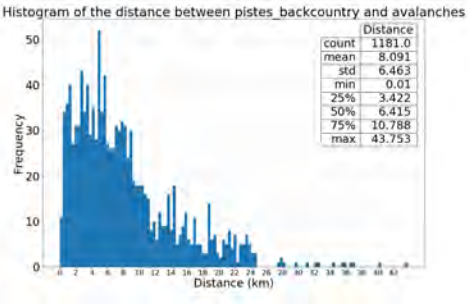
(a) Classic



(b) Classic + skate

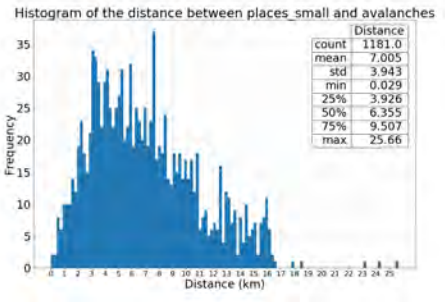


(c) Funparks

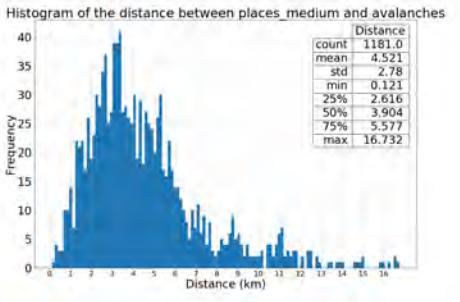


(d) Downhill

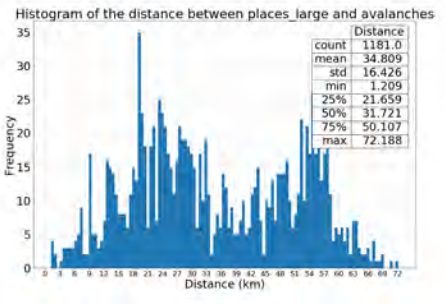
Figure A4: Histograms of the distances between ski piste grooming and avalanches.



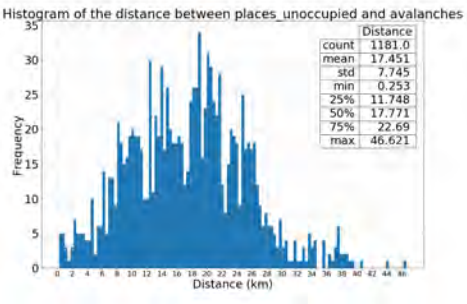
(a) Small



(b) Medium

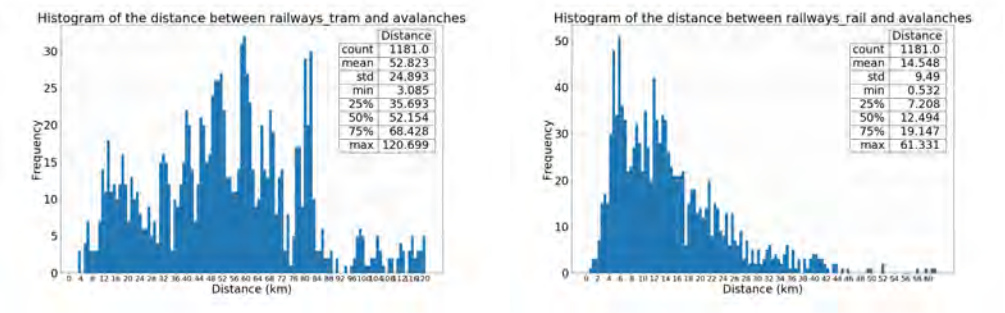


(c) Large

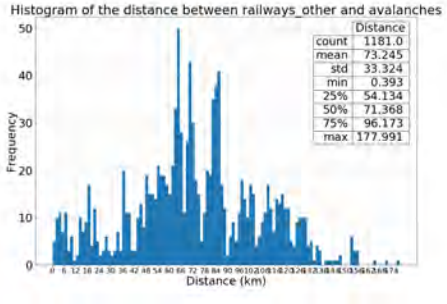


(d) Unoccupied

Figure A5: Histograms of the distances between settlements and avalanches.

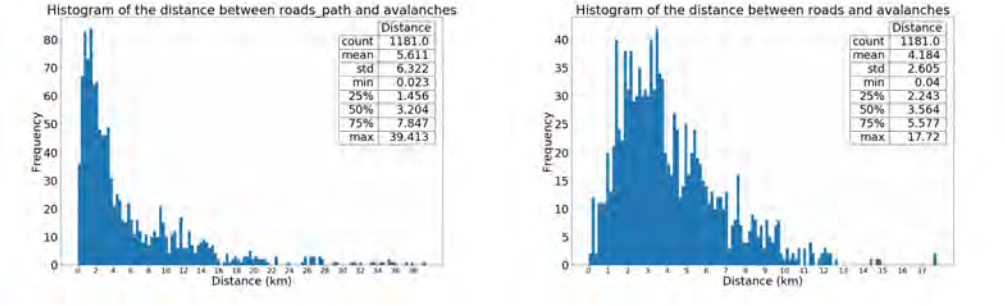


(a) Tram (b) Rail

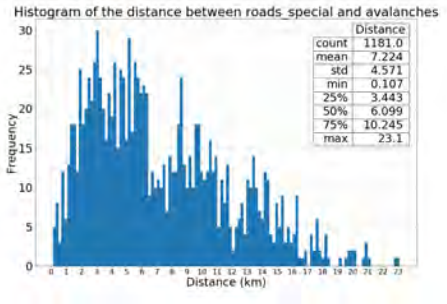


(c) Other

Figure A6: Histograms of the distances between railway types and avalanches.



(a) Path (b) Road

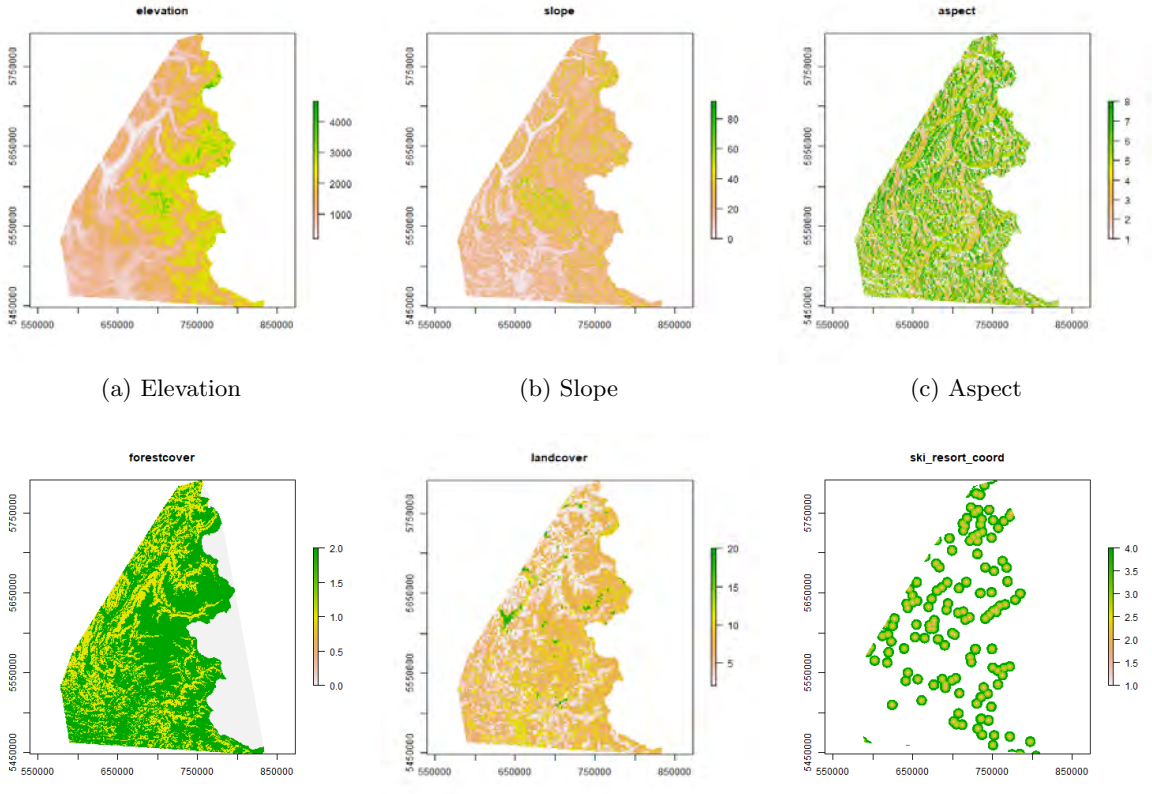


(c) Special

Figure A7: Histograms of the distances between road types and avalanches.

Appendix B

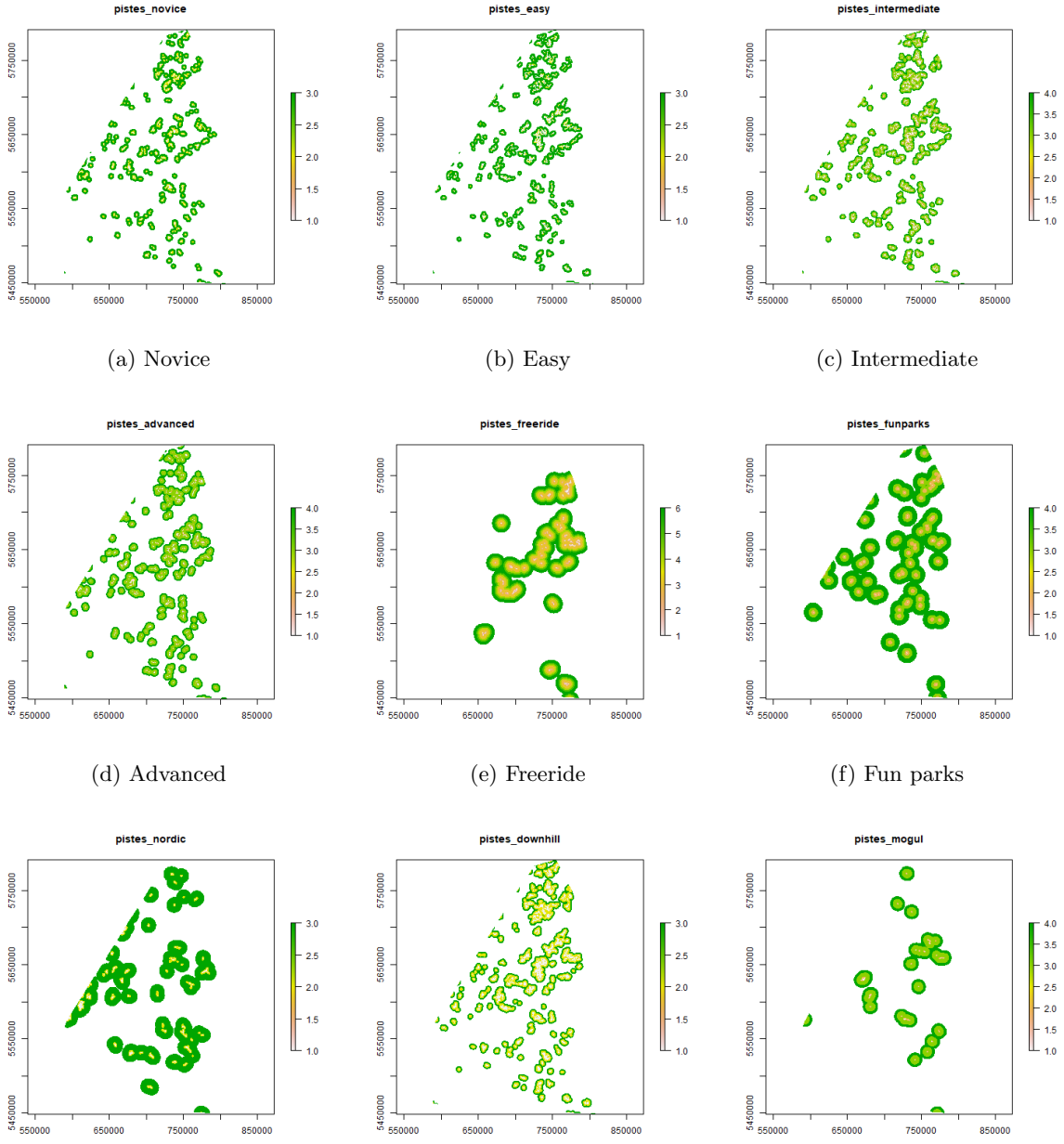
Appendix



(a) Elevation (b) Slope (c) Aspect

(d) Forest cover (e) Land cover (f) Ski resorts

Figure A1: Rasters of factors with same extent and resolution.

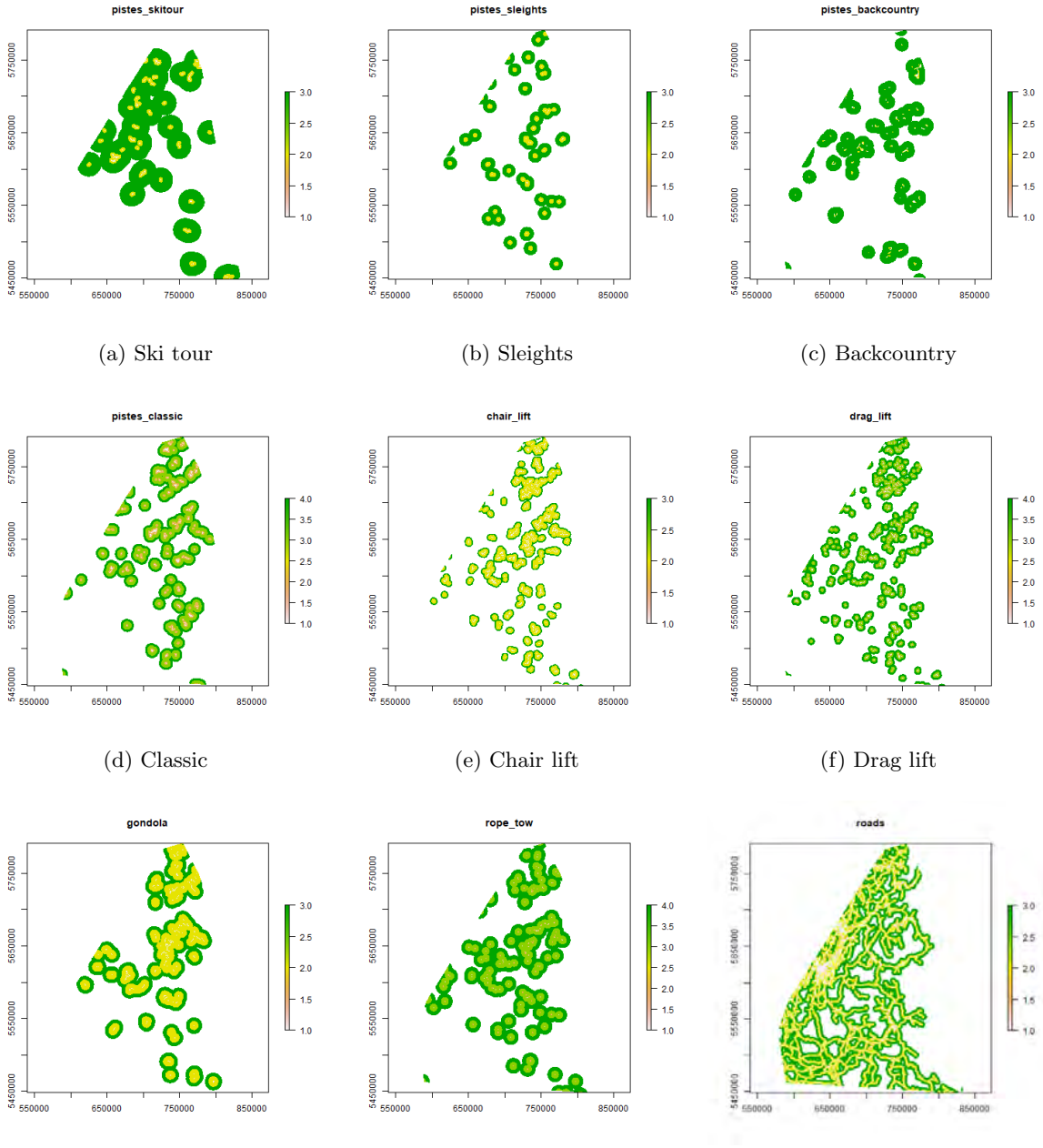


(a) Novice (b) Easy (c) Intermediate

(d) Advanced (e) Freeride (f) Fun parks

(g) Nordic (h) Downhill (i) Mogul

Figure A2: Rasters of factors with same extent and resolution.

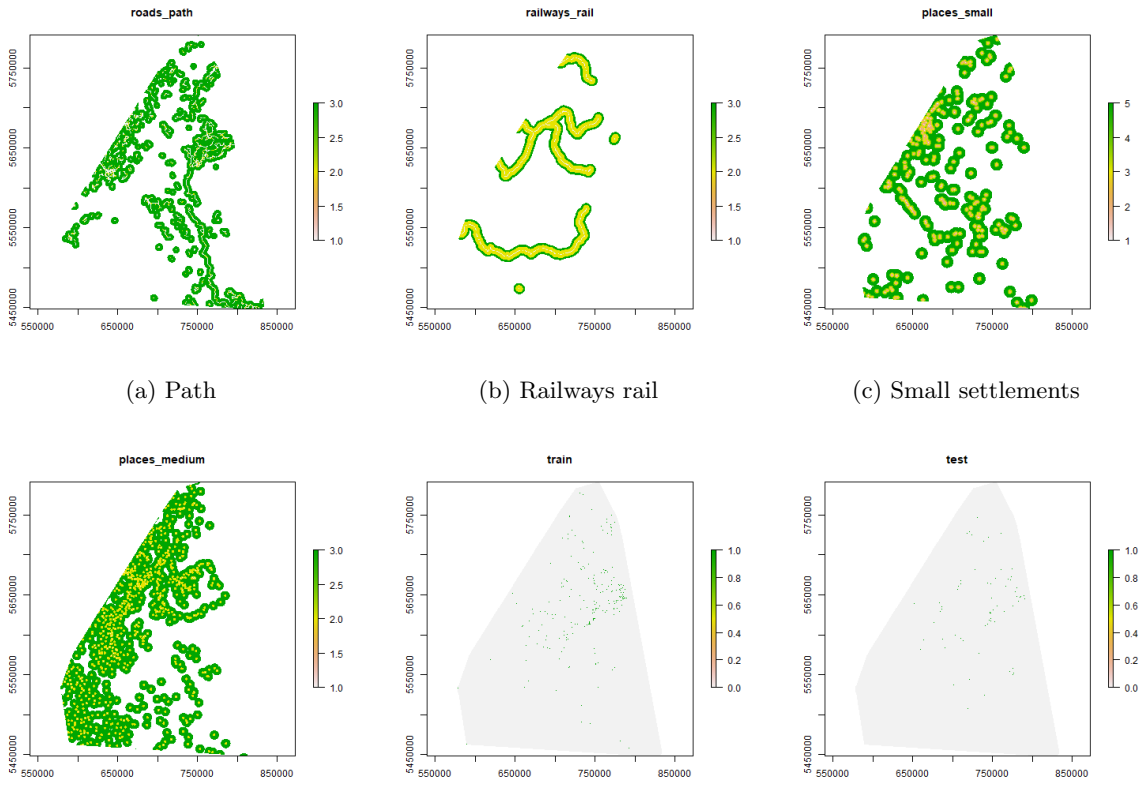


(a) Ski tour (b) Sleights (c) Backcountry

(d) Classic (e) Chair lift (f) Drag lift

(g) Gondola (h) Rope tow + magic carpet (i) Roads

Figure A3: Rasters of factors with same extent and resolution.



(a) Path (b) Railways rail (c) Small settlements

(d) Medium settlements (e) Train (f) Test

Figure A4: Rasters of factors with same extent and resolution.