Statistical forecasting of snow avalanches situations using field measurements

Master's Thesis

Faculty of Science

University of Bern

presented by

Augustine Saugy

2015

Supervisor:

Prof. Dr. Lutz Dümbgen
Institute of Mathematical Statistics and Actuarial Science

Co-Supervisor:

Dr. Robert Bolognesi
METEORISK

Oeschger Centre for Climate Change Research

Acknowledgments

My acknowledgments are first for Prof. Dr. Lutz Dümbgen for his helpful and expert advices in the domain of statistics. Furthermore, his availability and patience for answering my numerous questions was greatly appreciated.

Then, I also need to acknowledge all my colleagues working in the office METEORISK, who have been totally comprehensive, generous and tolerant with me. More particularly, I am very grateful to Dr. Robert Bolognesi for his understanding, his knowledge about snow avalanches which he generously shared with me and his help during all my internship and master thesis work.

Abstract

In mountainous regions, people and different infrastructure like ski resorts are threaten by snow avalanches during winter months. To prevent these catastrophic events, it is important to know if, during a particular day, snow avalanches are likely to occur or not, and which meteorological and snowpack variables are important in triggering these events. Linear discriminant analysis (LDA) and logistic regression (LR) are two statistical methods used to find which meteorological and snowpack variables are important in discriminating days with and without snow avalanches. For more accuracy, typical snow avalanche contexts can be defined, and corresponding main discriminating variables can be found. The system NivoLog is an analytical tool working on the basis of k-nearest neighbours algorithm, and used to help decision makers to define a particular day as snow avalanche day or not. Thanks to statistical analyses, different sets of parameters corresponding to each typical snow avalanche contexts can be constructed on the basis of different weighting of important variables. This leads to an improvement of the system NivoLog, which can be used differently according to typical snow avalanche contexts. Furthermore, the RCR in NivoLog depends on the number of similar nearest neighbours selected for deciding if, during a particular day, at least on snow avalanche is likely to occur. In this way, decision rules related to each set of parameters for typical snow avalanche contexts are delivered.

Keywords:
Snow avalanches - discriminant analysis - logistic regression - NivoLog parameterization

Table of contents

I. INTRODUCTION

a. Risk of snow avalanches in the Alps

Living in mountainous regions means living relatively close to the nature but also close to difficult meteorological conditions and natural hazards. Heavy rainfall, important snowfall, strong winds added to landslides, debris flows, rockslides expose people to natural risk they have to live with.

The usual definition of risk (can be given by the equation:

RISK = PROBABILITY OF OCCURRENCE * POTENTIAL DAMMAGE.

The PROBABILITY OF OCCURRENCE refers to a natural hazard (in the case of this study, snow avalanches) and is expressed as number between 0 and 1 according to the low probability (close to 0) or high probability (close to 1) that the natural hazard occurs. The POTENTIAL DAMMAGE refers either to infrastructure and material goods (in this case, the risk is expressed in units of money), or to humans lives (and in this case, the risk is expressed in units of humans lives). This equation is only valid at a certain place and for a defined period of time, because the two components of the equation can change with place and time. In this way, snow avalanche risk is only present where people can be affected or infrastructure damaged, for a certain place, at a certain time.

Alpine valleys and other mountainous regions in the world are threatened by possible snow avalanches during winter. However, with the previous definition, risk has existed only as the first inhabitants arrived in the valleys and has increased as economy and society developed. In the past, few people were living in the mountains, and their houses were gathered together behind natural protections as forests, moraines, erratic blocs or on flat areas to reduce the risk of snow avalanches (Ancey C., Gardelle F. &C., Zuanon J.-P., 2003). Then, in the early XXth century, tourism has developed, roads, railways and infrastructures have been built in areas where no one would have built in the past because of the common knowledge about the danger of snow avalanches. So, as time goes by, risk has become more and more widespread in Alpine valleys. The consequence of this development is that it is now necessary to develop and improve accurate snow avalanche forecasts to avoid potential disasters.
All winters are different in terms of meteorological conditions and snow avalanches: some are very snowy, leading to disastrous consequences, and some others can be sunny with less precipitation. In the Swiss and French Alps, some years are still in the memory because of the numerous and devastating snow avalanches which occurred. Concerning the Swiss Alps, 1951, 1975 and 1999 have been recorded as exceptional years (Villecrose J., 2001; Rougier H., 1975; Estienne P., 1951). Unusual snow avalanches in Evolène (Valais, 1999), Montroc (France, 1999), Davos, Klosters and Zernez (Grisons, 1950), Disentis (Grisons, 1975) have been observed. For the French part of the Alps, the years 1970, 1978, 1981 and 1999 have been exceptional concerning snow avalanches (Villecrose J., 2001). As explanatory factors, one often finds in the literature: abundant snowfalls, low temperatures and strong winds (Villecrose J., 2001; Marcel J., 1970). The results of such meteorological conditions are more numerous snow avalanches, which develop outside traditional pathways and lead to destruction of villages in the Alpine valleys.
Being aware of the past disasters, the importance of accurate snow avalanche forecasts is obvious. Nowadays, research in this subject is still going on in order to find better prediction models and a better understanding of snow avalanches behaviour.

b. State of the research about the subject

A lot of research has already been done in the domain of the prediction of snow avalanches according to meteorological variables, since the catastrophic snow avalanche years 1950 - 1951.

First, one can classify the studies according to the numerous parts of the world where snow avalanches have been studied. Boyne H.S. and Williams K. (1992) studied the region of Berthoud Pass, Colorado; Floyer J.A. and McClung M.D. (2003) the region of Bear Pass, Canada; the Norwegian Island of Svalbard was studied by Eckerstorfer M., Christiansen H.H. (2011) in two of their papers; Singh A., Srinivasan K., and Ganju A. (2005) studied the Indian Himalaya; Fromm R. (2009) studied the Austria; Jomelli V. et al. (2007) the region of the Valloire Valley in the French Alps; the Swiss and French Alps were studied by R. Bolognesi and F.N. Bouvet (Bolognesi R., 2015; Bellot H., Bouvet, F.N., 2010; Pahaut E., Bolognesi R., 2003), and by Buser O., Schweizer J., in the SLF (WSL Institute for Snow and Avalanche Research in Switzerland). These are only a little overview of all the parts of the world which have been studied concerning snow avalanches prediction.

Secondly, different methods have been used in order to find which variables are important for the occurrence of snow avalanches. The most frequent technique is the *Nearest Neighbour analysis*, which has been used for example by Singh A., Srinivasan K., and Ganju A. (2005), Gassner M., Brabec B. (2002) and McCollister C.M. et al. (2002), and studied in detail by R. Bolognesi in his PhD thesis (Bolognesi R. (1991, 1999). Another technique is the *Discriminant Analysis* which was used in the studies of Fromm R., (2009) and Floyer J.A. (2003). The *Logistic Regression* was used in the work of Jomelli V. et al. (2007) and the method of *Classification and Regression Trees (CART)* was used by Boyne H.S., Williams K. (1992). The list of these different methods is not exhaustive. Furthermore, one can also find a combination of different ones like in the study of Floyer J.A. and McClung M.D. (2003), where both the Discriminant Analysis and the Nearest Neighbour Analysis have been used, or in the study of McCollister C.M. et al. (2002), where the Nearest Neighbour method was combined with GIS information.

Thirdly, both the meteorological and snow avalanche variables vary according to the study and their availability in the different parts of the world. However, the results often emphasise the same explanatory variables. The amount of new snowfall was found to be an important variable by Fromm R. (2009), Jomelli V. et al. (2007), Floyer J.A. and McClung M.D. (2003), and Saemundsson T., et al. (2003); the wind variable by Eckerstorfer M., Christiansen H.H. (2011), Saemundsson, T. et al., (2003), McCollister C.M. et al. (2002); the foot penetration by Fromm R. (2009), Floyer J.A. and McClung M.D. (2003) and Floyer J.A. (2003). Once again, the list of these variables is not exhaustive but only give a brief overview of which variables can be important in predicting snow avalanches occurrence.

c. Snow avalanches: a short description

The more general and easier definition of a snow avalanche is "*a rapid flow of snow along a slope*" (Bolognesi R., 2013). However, the conditions of release, the snow and terrain characteristics for each snow avalanche are extremely diversified, and lead to different and various classifications.

The first differentiation refers to the triggering of snow avalanches. First, *natural snow avalanches* occur naturally, as their name implies, due to the imbalance between traction forces in the snowpack and resistance forces applied by the terrain. Traction forces can be increased by the addition of fresh snow by rainfall or by wind, or by infiltration of liquid water due to rainfall or snow melting (Bolognesi R., 2013). The second type of snow avalanches are *accidental snow avalanches*, triggered accidentally by a person or a group of people. In these cases, traction forces are increased by the weight of the person standing on the snowpack, leading to imbalance in forces and snow avalanches. These kinds of accidents mainly concern people practicing activities out of the secured resorts, as off-piste skiing, ski tours, snowshoeing etc. Thirdly, *artificial snow avalanches* are triggered by people responsible for the security of ski resorts or communication infrastructures, using different kinds of explosives. The blast creates a strong shock in the snowpack, and instable layers of snow flow down along the slope. This third kind of snow avalanches is the one, which is in interest for the present study. The difference between artificial snow avalanches and the two other types is that artificial snow avalanches are triggered regularly, as soon as conditions are instable, to secure infrastructures. On the contrary, if no natural or accidental snow avalanches occur, instabilities may continue over many weeks, without any snow avalanches occurrence, while in secured area, periods of instability are only punctual, before the artificial triggering.

The second differentiation refers to the type of snow avalanches. According to Robert Bolognesi (2013), the classification of snow avalanches can be done using one single criterion: the cohesion of snow in the departure zone, at the time of triggering (Bolognesi R., 2013). This leads to three different types of snow avalanches: powder snow avalanches, slab snow avalanches and wet snow avalanches. Powder snow avalanches mobilize fresh and dry snow, with a low density and with weak cohesion between snow crystals (Bolognesi R., 2013). Slab snow avalanches are more widespread and but their characteristics can be extremely different. However, slabs of snow are generally homogeneous and their crystallographic structure differs from other underlying snow layers (Bolognesi R., 2013). Wet snow avalanches mobilize dense and humid snow. They are generally observed during spring time, but can also occur in winter with heavy rainfall or sharp rising of temperature (Bolognesi R., 2013).

d. Goal of the study

First, the present work will aim at finding meteorological and snowpack variables which are the best to discriminate days with and without snow avalanches. A snow avalanche day is defined by the occurrence of at least one snow avalanche, while during no-snow avalanche day no snow avalanche occurs.
Linear Discriminant Analysis (LDA) and Logistic Regression (LR) will be performed on meteorological and snowpack variables, and those having a high discriminant power will be used to parameterize and validate the system Nivolog of the office METEORISK.
The second goal of this study is to find decision rules to apply in addition to NivoLog calculations. These will help decision makers to choose between options for the prediction of snow avalanches disasters.

e. Hypotheses and questions of research

Based on the review of scientific papers and knowledge in the domain of snow avalanches, three hypotheses can be formulated, and will constitute the basis of this work. Based on these hypotheses, questions of research can also be formulated.

- Snow avalanches occur according to specific meteorological conditions during winter time and sometimes in spring too. In this way, it is possible to find meteorological variables, or indices based on various meteorological variables, which trigger snow avalanches (Fromm, R., 2009).

  Is it possible to find good indices in order to take into account the snow drift (which is not always a direct measurement) and the wind direction (which is not a numerical variable)?

  Are all the variables available significantly useful in order to predict snow avalanches' occurrence? And which one could be neglect?

- All meteorological variables have not the same power in order to trigger snow avalanches. Some have a higher predictive power, as for example the amount of new precipitation, the foot penetration, the present temperature trend and the wind speed. (Floyer J.A., McClung M.D., 2003; Floyer, J.A., 2003; Eckerstorfer M., Christiansen H.H., 2011; Saemundsson, T. et al., 2003).

  Which variables are the most influent in predicting snow avalanches' occurrence?

  Are meteorological variables more influential than snow variables?

  Are the predictive variables similar in the different sites or do they depend on the situation and topography of each site?

  Are the variables new precipitations, foot penetration, present temperature trend and wind speed significantly influential as in other studies?

- If enough data are available, it should be possible to find statistical tools able to find the best explanatory variables in order to predict snow avalanches' occurrence (Fromm, R., 2009). Two methods which can find explanations concerning snow avalanches occurrence are linear

discriminant analysis (LDA) and Logistic Regression (LR) (Floyer, J.A., 2003; Floyer J.A., McClung M.D., 2003; Fromm, R, 2009).

Do the two methods of LDA and LR show similar results? Which variables are significantly influential for each of these two methods? What are the differences?

Which method could be best trusted in order to predict snow avalanches' occurrence? Which one has the lowest uncertainty and the best predictive power?

f. Framework of the project

Today's snow avalanches prediction is the result of a societal, political and juridical evolution since people began to live in the alpine valleys. The always more numerous private offices in the domain, the enhanced role of the media in the society and the more frequent conviction of people dealing with snow avalanches lead to a change in the role of services dealing with snow avalanche prediction. All these evolutions are added to the influence of past vision of responsibilities with respect to natural disasters which is still present in our society.

The first trend concerns the different stakeholders in the domain of snow avalanches prediction. From the XIXth century until now, different stakeholders have handled with risk management of snow avalanches. It was first the State which was responsible of risk management in mountainous regions, until the beginning of the 1980's (Ancey C., 2011). At this time, more and more private offices working in the domain of snow avalanches developed: André Roch, Vincent Koulinski, Claude Charlier and Christophe Ancey, Martin Jaeggi, André Burkard and Robert Bolognesi have all created private offices dealing with risk management of snow avalanches in addition to the State (Ancey C., 2011). Then, after about one decade, private offices have become dominant in comparison to public organisms.

The second change during the past century is the always more important role of media in communication. If a disaster occurs, the media will not only inform, but also add emotions and dramatic tone to attract people's attention. This often leads to misinformation, quick conclusions and sometimes false interpretations, as the case of the snow avalanches which occurred in "La crête du Lauzet" (Descamps P., 2005). In this case, the media did not respect a minimum of presumption of innocence for the mountain guide and did not systematically verify their information (Descamps P., 2005). This important influence of the media increases the psychological and social pressure on the persons concerned by the case, and interferes with the justice. But as Descamps P., 2005 write in his paper, one has to keep in mind that "*le temps des medias n'est pas le temps de la justice*" [*the time of media is not the time of justice*] (Descamps P., 2005, p.123), and people should stay aware that all we hear or read is not always valid from all viewpoints.

The third evolution concerns the juridical context. In recent years, there has been an expansion of convictions of mayors, presidents of municipalities, people responsible of security or mountain guides as a consequence of accidents due to snow avalanches. One can cite two examples among many others to illustrate this tendency. The first occurred on the 21st of February 1999: a snow avalanche killed twelve people in the municipality of Evolène (Valais). The president of the municipality and the person responsible of security were both given suspended prison sentences for *manslaughter* and *obstacle to the traffic through negligence* (Tribunal cantonal du Valais, 2006). The final judgment reviewed in 2006 gave the sentences of 2 months suspended prison to the chief of the

security and 1 month to the president of the municipality (Tribunal cantonal du Valais, 2006). The second example occurred on the 23^{d} January 1998 in the "Hautes Alpes" (France), where a snow avalanche killed nine young people and two adults. In this case, the mountain guide was given two years of suspended prison for *manslaughter*, and additionally to pay a fine of 8000 francs (Descamps P., 2005).

These evolutions are based on a more general idea of what is the responsibility of humans with respect to snow avalanches or more generally, natural disasters. As soon as people established in the valleys, snow avalanches posed the problem of responsibilities. However, responsibility has not always been attributed to the same protagonists: imaginary creatures, Devil or people with ill-repute lives have been successively accused to cause snow avalanches. In past times, snow avalanches were personified through "avalanche beasts" like the "Lauwitier" (a sort of chamois or billy in the Lötschental in Valais), to which people had to be respectful (Reyt M.P., 2000). But from the XIIIth to the XVIth century, the Church transformed these personifications into evil or even Devil (Reyt M.P., 2000). In this way, natural hazards became systematically the consequence and the divine punishment for a fault or a wrong action committed by the society (Reyt M.P., 2000). For example, if a snow avalanche occurred in a village, it was thought to be because of the inhabitants' sins. Thus, the responsibility was shared by all the society, and accepted as the consequence of bad actions. Later, during the XVIIth century, Rationalism developed: new techniques and understandings of natural hazards led to the belief that they would be avoided in the future. Unfortunately, numerous disasters showed to people their misunderstanding, and the culpability was placed again in the centre of natural disasters, as explained by Reyt M.P., 2000: "*aux gens de mauvaise vie les avalanches, aux innocents l'impunité*" [*ill-repute people get snow avalanches, innocent ones impunity*] (Reyt M.P., 2000, p.40). In this way, History shows us that people always have searched for a person responsible of disasters. If this person cannot be designated, the disaster is not understood: this idea is still present in today's collective unconscious (Reyt M.P., 2000).

In this way, the present study integrates in the evolution of the framework described above. It appears to services dealing with snow avalanches that people do not trust only "expert's knowledge and experience" anymore, but are more and more convinced that a choice between different possibilities of action regarding snow avalanche prevention is more accurate. In other words, it may be better to get various elements to take a decision instead of one definitive diagnosis. This evolution has already been observed by Mr. Robert Bolognesi since 2006: « [les responsables de sécurité avalanche] *ont une préférence pour des éléments de diagnostic plutôt qu'un diagnostic définitif* »[people responsible for the security prefer elements of diagnosis rather than a definitive diagnosis] (Robert Bolognesi, Internal document of METEORISK, 2014). Following this same observation, the aim of the present work is to parameterise Nivolog, which will deliver various elements of diagnosis to take into account at the time of a decision, and to define rules which can be used for an optimal use of the system.

II. NIVOLOG PRESENTATION AND DATA

a. Supervision

The present work is supervised by two people: Prof. Dr. Lutz Dümbgen from the Graduate School of Climate Sciences, Bern, and Dr. Robert Bolognesi, director of the private office METEORISK in Sion.

○ *Graduate School of Climate Sciences*

The Graduate School of Climate Sciences belongs to the University of Bern and is run by the Oescheger Centre for Climate Change Research (Grosjean M., 2012). Its role is to offer educational scheme and training opportunities for the future climate scientists and professionals (Grosjean M., 2012). Some pioneers in Climate Research have worked at the University of Bern, as for example Eduard Brückner, Heinrich Wild or Rudolf Wolf (Grosjean M., 2012).
Prof. Dr. Lutz Dümbgen is affiliated with the University of Bern and more particularly with the Institute of Mathematical Statistics and Actuarial Science, where he teaches Statistics. His main research interests are the Nonparametric Statistics, the Multivariate Analysis and the Statistical Computing.

○ *METEORISK*

METEORISK is a private office which has been active since 1999. Its fields of competence are varied: forecast of meteorological risks, research and development, engineering and education (Erard N., 2007). These services are either punctual or regular and concern public authorities, private companies as well as media or individuals (Erard N., 2007).
Dr. Robert Bolognesi is the header of METEORISK. He is graduated of the EPFL (Ecole Polytechnique Fédérale de Lausanne), in the domain of computer science (Thesis *« Modèle de metaraisonnement. Application à la prévision de phénomènes catastrophiques »*) and of the University of Grenoble in the domain of geography (Thesis « *Analyse spatiale des risques d'avalanches* »). In addition, he also obtained a diploma « Habillitation of Conducting Research *»* at the University of Grenoble (DHDR).

b. The system Nivolog

Introduction

In mountainous regions, infrastructure, roads, railways, ski resorts etc. are threatened by snow avalanches during winter season. To secure these sensitive infrastructures, two possibilities can be chosen: either constructing structures like protective dikes, retention ponds or avalanche barriers, or taking temporary measures (road closure, evacuations, preventive snow avalanche maintenance, etc.). Thus, these measures imply snow avalanche forecasting or at least risk estimation. However, snow avalanche forecasting is often difficult, and errors can lead to severe consequences. So, the system Nivolog is designed to help forecasters and people responsible for the security in this difficult task.

NivoLog was created in the 80s at the time of the democratization of computers, by Robert Bolognesi. The first version of Nivolog was called *AvaLog* and was the first expert system able to predict local occurrence of snow avalanches. Since then, it has continuously been developed and improved by further researches, as for example, the adding of Robert Bolognesi's thesis at the EPFL, to reduce the risk of false diagnostic; in 1997, Nivolog was created. The most recent version of this system dates from 2014 and includes new functionalities for the configuration and the backup. This later will be used in the present study.

Functionalities

NivoLog includes archiving and information processing functions.

- *Archiving functions*

These functions allow data collection, archiving and sorting for regional snow avalanche forecasting. Information can have diverse forms: data (observations, measurements, etc.), views (pictures, maps, graphs, etc.), videos (clips) and notes (text files), and refer to geographic, topographic, meteorological and snowpack characteristics, as well as snow avalanches. All these information refer to an "area" of several squared kilometres, where climatic conditions are similar and snow avalanche forecasting can be performed on the basis of some measurement stations in this area. The area is made up of different "locations" where snow avalanches can occur.
Geographic and topographic information are constants and do not need to be updated, except if correction is needed. Meteorological, snowpack and snow avalanche information are variables whose values change with time. In consequence, they need to be updated regularly (each day or twice a day for meteorological and snowpack conditions; whenever a snow avalanche occurs for snow avalanche information).
For this first function of NivoLog, it is important to accurately select number and position of measurement stations, and to be constant when collecting and entering data in the system.

- *Information processing functions*

NivoLog is not only a system used for archiving and data collection, but it is also a powerful analytical tool to predict snow avalanche occurrence. The main idea is that similar meteorological and snowpack conditions generally lead to similar snow avalanches contexts. In other words, distances between the observation to predict and other situations are calculated, and the nearest ones (the most similar ones) are an illustration of what could happen for the observation in consideration. In NivoLog, the user has the possibility to pilot all the procedure of similar contexts determination, by changing and adapting predictors, parameters and restrictions (filters).
The analysis performed in NivoLog analysis is the k-nearest neighbours (kNN) analysis. Distances are calculated between the case to predict and all other observations in the database. Then, it is possible to select the number of nearest observations displayed by the system; for the present work, 10 nearest observations are displayed, and the number of observations being "snow avalanche days" is recorded (see Chapter IV.C). To parameterize the calculation, it is possible to: weight or/and normalize the analogy criteria, to define filters, to choose the results being displayed. For the present work, this step is very important (see Chapter IV.C).

Concrete examples

The following figure shows the parameterization grid in NivoLog (before the analysis):

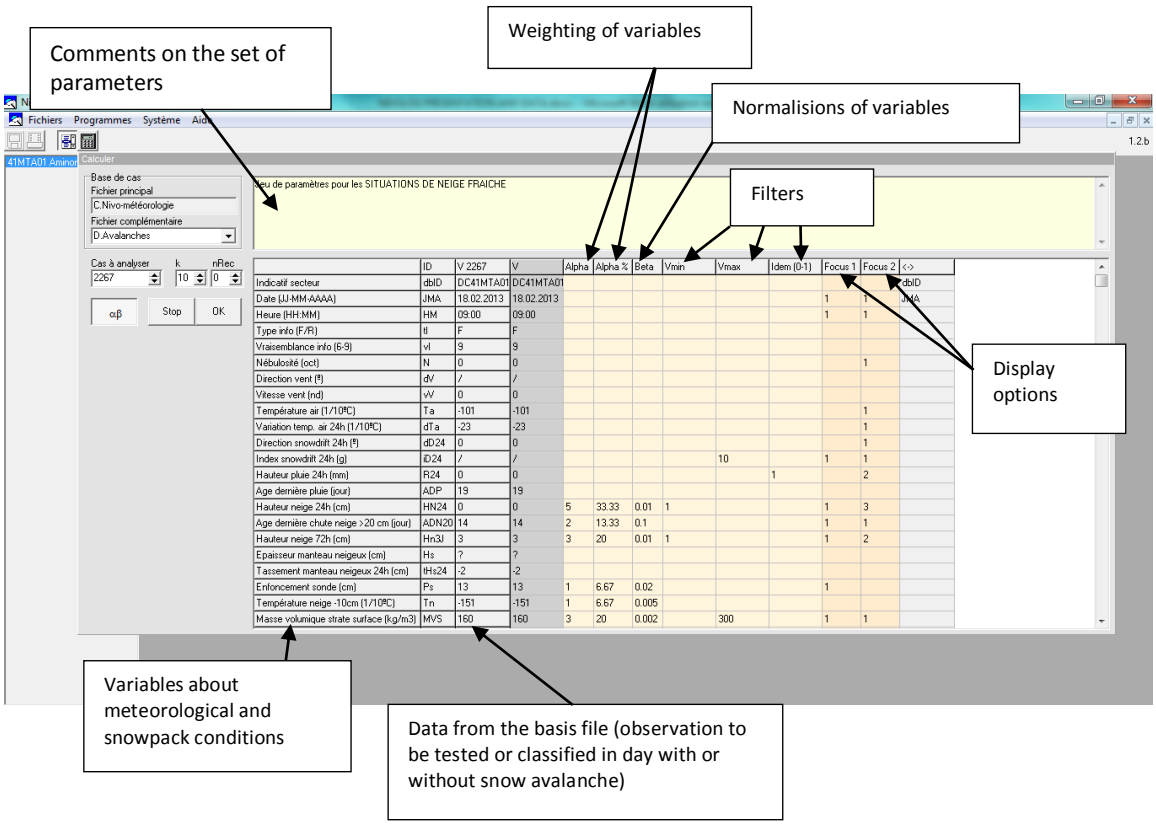

Figure 1: Parameterization window in NivoLog

Once the analysis is performed, results are showed in the following manner:

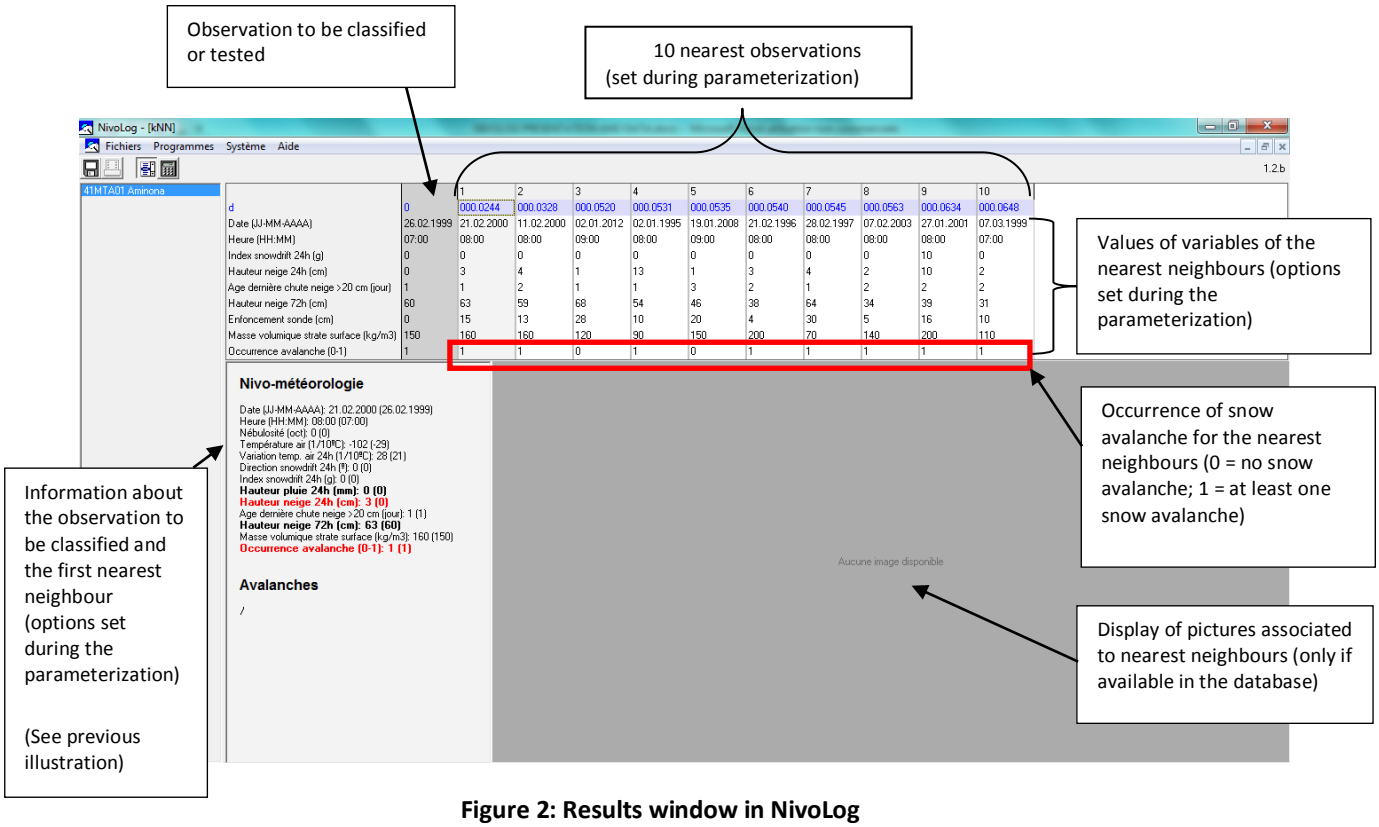

Figure 2: Results window in NivoLog

In this example, 8 nearest neighbours over 10 are snow avalanche days, and only 2 are no-snow avalanche days. So, the observation to be classified is likely to be a snow avalanche day too. This is actually the case (the observation to classify has actually the occurrence of one snow avalanche).

The current use of NivoLog as it is presented above only has one set of parameters by default for all situations, but which can be changed by the user during the parameterization. The improvement expected during this study is to create different sets of parameters related to typical snow avalanche contexts. In this way, when a snow avalanche context is clearly identified, NivoLog users can parameterize the system with the corresponding set of parameters to obtain better reliability in predicting the occurrence of snow avalanches in their particular areas.

c. Region of study

The region of study for this work is located in Switzerland, and more particularly in the canton of "Valais", at the South-western part of the country. The Valais is an alpine canton, where natural hazards related to mountainous regions are of great importance during all seasons, but in a more important way during winter time. Because many people live in valleys, important snowfalls leading to increased snow avalanche activity can threaten roads and railways leading to the villages and even houses located at these places. For example, during the month of February 1999 (Henzen, W., Schönbächler D., Bolognesi R. et al., 2009), many snow avalanches reached human infrastructure in alpine cantons of Switzerland, leading to 17 fatalities. Furthermore, this canton is principally known for the numerous ski resorts and winter sports offered to the public during winter times. In this way, it is necessary to secure ski areas, to build snow avalanche dams and barriers, but also to collect data and observation about this phenomenon for a better comprehension. The data on snow and meteorological conditions are then used to feed models, which can help security managers to take decisions in the ski resorts. For example, the decision to close ski areas or to trigger artificially snow avalanches with explosives, in order to secure the ski slopes.

The ski resort which interests us in this study is "Crans Montana", and more particularly, the neighbourhood of Aminona. This resort is located on the right bank of the Rhône valley, on a plateau, at an altitude of 1500 m. It began to develop in the IXXth Century, when people travelling between Paris and Milan wanted to have a break in their trip. Skiing developed in 1905, the first real ski-lift was opened in 1934, and, since then, the resort has much developed (Emery Mayor D., 2009). For example, the Alpine World Ski Championships were held in Crans Montana during the year 1987, and in 2008, one of the world cup ski races took place in that resort (Emery Mayor D., 2009). Nowadays, Crans-Montana is known for its 140 km of ski slopes, 22 ski lifts and cableways and the tourists coming from various parts of the world.

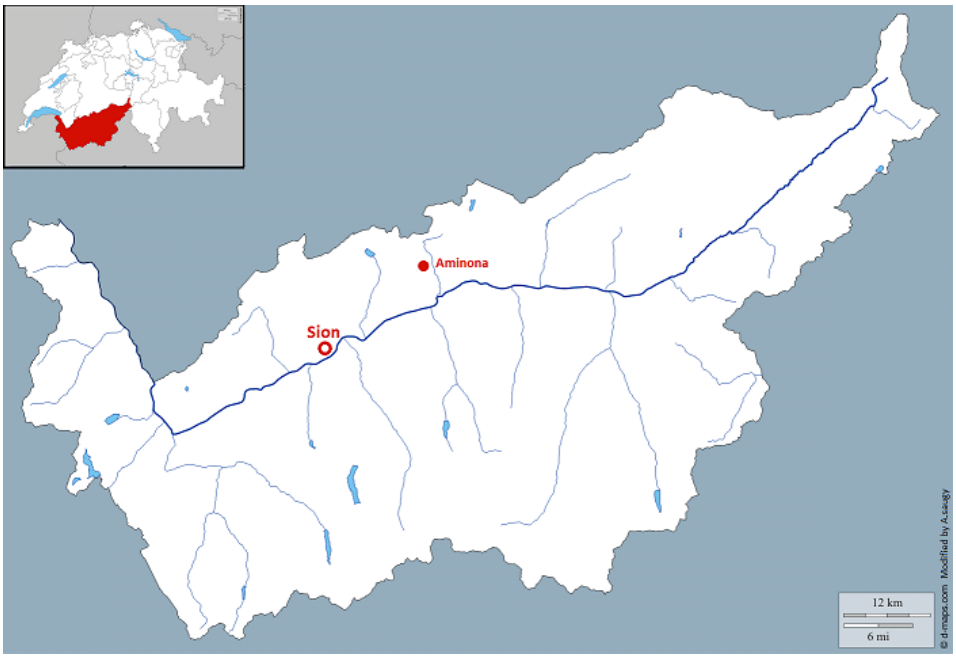

Figure 3: Region of study.

However, Crans Montana has not developed there by chance. In fact, the plateau where the resort is located benefits from a special climate, related to the whole clement climate of the canton of Valais. This part of Switzerland has a lower cloud cover in percentage, lower precipitation rates and a higher sunshine duration than the mean values of the whole country. This can be partly explained by its situation. In fact the canton is located inside the alpine barrier, and protected by the mountains from perturbations coming either from the North, the East, the South or the West. Some statistics can be performed based on the meteorological data of Meteo Swiss, in order to illustrate these differences. In our case, data about sunshine duration are illustrated in the following graph, for the city of Sion and Bern, and the mean for Switzerland.

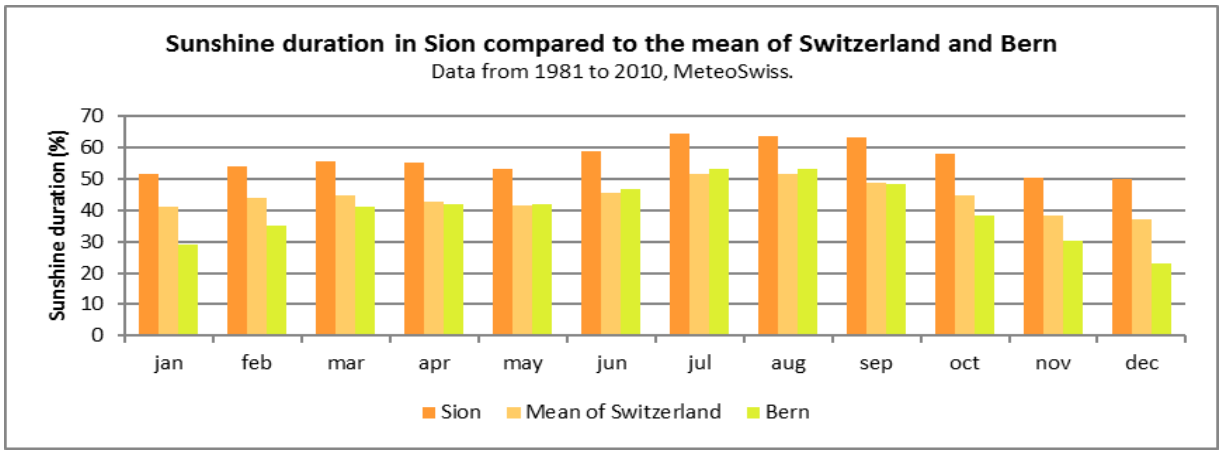

Figure 4: Sunshine duration for Sion, Bern and the mean of Switzerland

The graph above shows that Sion, the capital of the Valais, benefits from a higher percentage of sunshine duration (about 10%) compared with the mean of Switzerland or the canton of Bern.

In Crans Montana, located approximately 20 km away from Sion, a study headed by Robert Bolognesi in 2008 showed that the sun is present in this resort nine days over ten, the wind is generally low and the winters are cold with an adequate snow cover (Emery Mayor D., 2008).
These clement meteorological conditions attract many people in the resort, especially during winter times. In this way, it has become very important to ensure their safety on the ski slopes and in their houses, and to protect them against snow avalanches occurring in the area of the resort.

d. Data

The first idea of this work was to analyse data of various ski resorts in various regions (France and Switzerland). First, data from “L’Alpe d’Huez” in France appeared to be available at the beginning of the study. However, early in the winter, snow cover was very low, and consequently, ski resorts and patrollers working there had an immense work in order to prepare ski slopes for Christmas holidays coming soon. During these holidays, thousands of people come in the ski resort and work is further amplified. After this period, all people having worked continuously since the end of November want to take some days off and the remaining people had, once again, a lot of work. In this way, in January, no data have been sent yet, and the decision was taken not to wait for the data of this ski resort anymore. A second ski resort could have been a data provider: Anzère, located in Switzerland, in the canton of Valais. Here again, the process to obtain data was very complicated due to immense work in the ski resort during the period of Christmas. Furthermore, these data were not complete, and it was also decided to eliminate this ski resort. Finally, the available data selected for the present work are from the ski resort of Crans Montana, presented above.

The data were collected by the patrollers of the ski resort from 1994 until 2013, and stored in the program *Nivolog*, provided by the office METEORISK. First, only manual measurements were done, until the installation of an automatic station of measurements. However, an automatic station can only give information at a specific place, without any adjustment. For example, a measurement of snowfall in 24 hours recorded by an automatic station is equal to 40 cm. However, particularly strong winds blew during the night before and accumulated snow at the measurement location, leading to an excessive value for the snowfall. A patroller can adjust the value, and give better estimation of real conditions than the automatic station. In this way, patrollers have always an important role in collecting data, in delivering precise information thanks to their critical estimation of the values.

Two files are available: one concerning meteorological and snowpack conditions, which is called “*DC41MTA01 – Original*” (cf. chapter IV.a.6), and another concerning only information about snow avalanches characteristics, called “*DD41MTA01 – Original*” (cf. chapter IV.a.6). As the goal of this work is to try to predict the occurrence of snow avalanches based on meteorological and snowpack data, only “*DC41MTA01 – Original*” will be taken into consideration in a first time (cf. chapter IV. a.5).

The location where information about meteorological conditions and snow was collected is located at an altitude of 2300 m, at the upper arrival of one cable car coming from “Aminona”, a neighbourhood of Crans Montana (see figure 4, below).

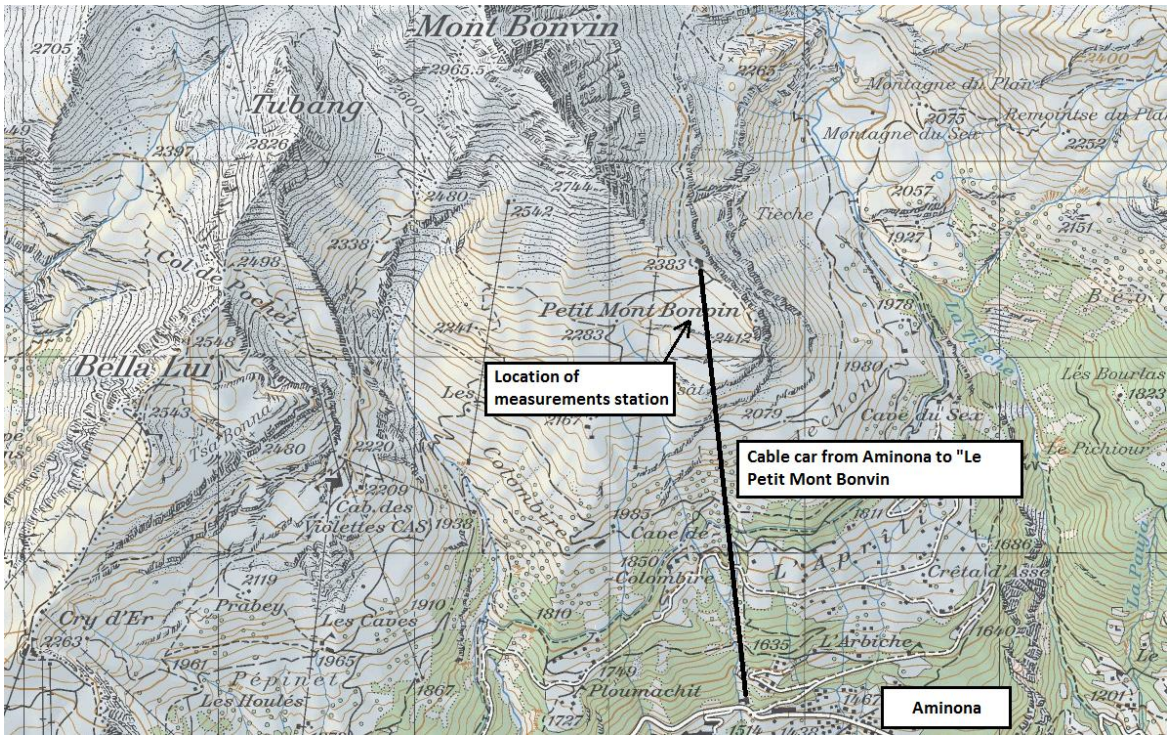

Figure 5: Location of measurement station

This measurements field was installed in 1991 near the bottom of the "Petit Mont Bonvin", chosen for monitoring snow avalanches in the region. The place of measurements is located in a relatively neutral place concerning the wind: it means that, neither accumulation nor erosion of snow due to wind transport should take place. However, it has to represent the wind situation in the principal slopes located above the area to secure. It also has to be representative of the conditions of the whole ski area, meaning that the measuring instruments should not be hidden by infrastructures, located in a depression where accumulation of snow could occur, or too close to the ski slopes, where skiers could interfere with the measurements.

The database is constituted of 19 years of observations for winter months, from 1994 until 2013. Depending on the years and on the date of the first snowfall (which generally marks the beginning of snow, weather and snow avalanches observation for the winter), data are available from November until April of the following year. At least 4 different people worked at this measurement site during these 19 years. In this way, records are not always regular, and they are sometimes effectuated one to three times a day: at 8:00, 11:00 or 14:00; but at least, one record per day is available.

The database of "*DC41MTA01 – Original*" is constituted of 3 types of variables: weather conditions, snowpack characteristics, and snow avalanches observed. All variables are presented in table 1.

Variables		
Weather	**Snowpack**	**Snow avalanches**
Cloud cover (okt)	Cumulative snowfall over the season (cm)	Number of shots in 24h
Wind speed (knot)	Snowdrift direction 24h (°)	Number of artificial snow avalanches in 24h
Wind direction (°)	Snowdrift Index 24h(g)	Number of accidental snow avalanches in 24h
Air temperature (1/10 °C)	Snowdrift direction 72h (°)	Number of natural snow avalanches in 24h
Air temperature variation in 24 h (1/10 °C)	Snowdrift Index 72h (g)	Total number of snow avalanches in 24h.
Humidity of the air (%)	Age of the last rainfall (days)	Mean magnitude of snow avalanches (0-5)
Rainfall (mm)	Age of the last snowfall exceeding 20 cm (days)	Age of the last artificial snow avalanche (days)
Snowfall (cm)	Age of the last snowdrift (days)	Age of the last accidental snow avalanche (days)
	Age of the last hoar (days)	Age of the last natural snow avalanche (days)
	Snow cover thickness (cm)	Regional risk predicted (1-5)
	Variation of snowpack thickness (cm)	Local risk observed (1-5)
	Thickness of probe penetration (cm)	
	Snow temperature at 10 cm depth (1/10 °C)	
	Density of the surface layer (kg/m3)	
	Thickness of surface refreezing (cm)	
	Thickness of surface hoar (cm)	

Table 1: Variables available in the file "DC41MTA01 - Original"

III. STATISTICAL METHODS

The first step of any statistical study is to prepare a reliable database. So, coherence tests, cleaning of errors and corrections are performed to obtain a clean database, potentially leading to better results. For additional information and explanations, see next chapter IV.a. "*Preparation of the database*".

In a second step, two statistical methods are selected to determine the relative importance of each variable in triggering snow avalanches. This will be performed by linear discriminant analysis and logistic regression. These two methods aim at constructing a statistical function, which will classify days in "snow avalanche days" or "no snow avalanche days", based on the variables which best discriminate between these two types of days.

In a third step, *p-values for classification* and *k-nearest neighbours* will be used to compare the assessment of the right classification rate of different statistical methods.

a. Determination of important variables

Linear discriminant analysis (LDA)

Discrimination and classification methods generally have two main goals. First, "*discrimination describes the differential features of objects from several known populations and tries to find "discriminants" whose numerical values are such that the collection is separated as much as possible"* (Johnson R.A. and Wichern D.W., 2007). Secondly, "*classification sorts objects into two or more labelled classes, […] deriving a rule that can be used to optimally assign new objects to the labelled classes*" (Johnson R.A. and Wichern D.W., 2007). In this work, discrimination is used to determine important independent variables (related to meteorological conditions and the state of snow), which separate the dependant variable *"days with and without snow avalanche"* with a high accuracy. Then, classification is used to check the right classification rate (RCR) of days with snow avalanches or not, depending on the discrimination.

LDA yields a linear discriminant function of the form:

$$f(x) = a + \sum_{j=1}^{p} b_j X_j$$

Then *f(x)* is the prediction of Y ϵ {0,1} (in our case 0 is for days without snow avalanches and 1 is for days with at least one snow avalanche), a is the estimated intercept, *p* is the number of independent variables, *j* is the number of covariables and b_j is the coefficient associated to each explanatory variable X_j.

This function separates the two classes as much as possible, such that days with snow avalanches and days without snow avalanches are as far as possible in the p-dimensional space. The interesting part of this function to determine which variables are important in predicting the occurrence of snow avalanches are the coefficients b_j. However, they are dependent on the scale of each variable, and often bias the interpretation. For example, the density of snow varies from 40 to 640 while the cloud cover only varies from 0 to 8; this would give more importance to the density of snow because the

values of this variable are higher than the values of the cloud cover variable. This problem can be encountered by performing LDA on standardized data.
As explained above, the solution to smooth the differences in the ranges of values is to standardize each variable. The method of standardization for j = 1,...,p is the same as the one found in the book of Johnson R.A. and Wichern D.W. (2007):

$$Z_j = \frac{Xj - \mu j}{\sqrt{\sigma jj}}$$

Where X_j is the column vector corresponding to one variable, μ_j is the mean and $\sqrt{\sigma jj}$ is the standard deviation.

After this step, LDA is performed on the new standardized independent variables with a great improvement: coefficients are directly linked to the importance of each variable in discriminating days with or without snow avalanches, and not dependent on the range of values as before. In other words, if a coefficient is high for one particular variable, one can directly say that this variable plays an important role in discriminating days with or without snow avalanches. Furthermore, standardization of variables does not influence the right classification rate. In this way, the same percentage of right classification will appear after linear discriminant analysis performed on initial or standardized data.

Some problems can arise when performing LDA. The first one is linked to the possible collinearity of some variables with each other. In this particular case, two or more variables are too close to each other, and the information carried by these variables is redundant. In such cases, LDA cannot be performed. So, it is necessary to remove one of the two or more variables which are collinear. The second problem which can be faced is linked to the number of variables compared to the size of the sample. If the number of observation is too low compared to the number of variables, analysis cannot be performed. There is no general universal rule for the relationship between the number of variables and the sample size. However, the advice delivered by my supervisor is that the minimal number of observations ($N_{min)}$ should be greater than 5 times the number of variables (p) plus one. The numerical formula for this rule is given below:

$$N_{min} \geq (p+1)*5$$

In the present work, 17 variables will be used in order to discriminate between days with or without snow avalanches. In this way, the minimal number of observations should be greater than 90.

Finally, the reasons why LDA is chosen in the present work and its advantages are listed below:

- It helps to investigate observed differences when causal relationships are not well understood (for example, which variables are the most responsible for the occurrence of snow avalanches?).
- It works with some allocation rules, which can be helpful to define rules for decision makers (see section V).
- It identifies the most important items responsible for differentiation (Castillo-Rivera M. et al., 2000).

Logistic regression

The second method used to determine which of the meteorological variables have the greatest influence in triggering snow avalanches is Logistic Regression (LR). This method is used when one variable takes the binary values of zero and one. Typically, in our case, a day can be classified as a snow avalanche day when at least one snow avalanche occurred and will take the value of one; for a day without snow avalanches, the assigned value is zero. The goal of Logistic Regression is to estimate the probability that the binary variable (day with or without a snow avalanche) takes the value of 1, using a combination of numeric covariables (in our case, the meteorological variables) (Geoffrey J. McLachlan, 1992). More particularly, the logistic regression function calculates the log odds ratio (logit) of the binary variable, given covariables. This is shown in the formula below:

$$f(x) = \text{logit}\,(P(Y=1 \mid X = x))$$

Where $\text{logit}\,(p) = \ln\left(\frac{p}{1-p}\right)$, Y is the binary variable to be estimated, X are the independent covariables.

As for LDA, estimates are indicators of the importance of each independent variable in determining if one day may be "snow avalanche day" or "no snow avalanche day". However, the output of LR also gives the standard error of the estimate, the z-value statistic and more importantly, the p-values of each coefficients (denoted as $\Pr(>|z|)$). They indicate which of the different independent variables are significant in order to separate days with or without snow avalanches. To improve the understanding of the output, some significance signs are used in the software R: "***" are used for statistical significance with confidence level of more than 99.9%, "**" for 99%, "*" for 95%, "." for 90%. Significant variables with a lower confidence level have no sign beside their p-value. This output allows a better selection of explanatory variables than in LDA, because significance is statistically measured, while LDA with standardized variables only gives coefficients indicating the importance of each explanatory variable. As for LDA, the number of variables compared to the sample size can be a problem if the sample has not enough observations compared to the number of variables. In the present work, the same rule based on the advice of my supervisor will be applied (see above).
This method exhibits several advantages:

- "It is highly effective at estimating the probability that an event will occur: [more particularly,] it creates estimates of the likelihood that an event occurs, given a set of conditions" (Sweet S.A. and Grace-Martin K., 1999).
- It is not necessary to have normal distributions of variables to perform a logistic regression (Angillieri M. Y. E., 2010).
- One can take into account the risks associated with the choice of certain variables and calculate confidence intervals for each variable (Keller R. et al., 2007).

Conclusion

LDA and LR are the two statistical methods used to determine which meteorological and snow variables are important in order to separate days with and without snow avalanches. These two methods will be performed on the same dataset, and results concerning important variables should be quite similar. The result of the determination of significant variables triggering snow avalanches is of great importance for the following parameterization of the system NivoLog. In other words, weighting of variables in NivoLog will base on the results of LDA and LR.

b. Right classification rate (RCR) of different methods

After having determined which variables are significant in discriminating days with and without snow avalanches, the second step is to assess the right classification rate (abbreviated RCR in the following parts of this work) associated to the two different groups. The assessment is displayed by a table which presents the membership to groups in the reality, and the membership predicted by the statistical analysis. On the table below, correct classifications are illustrated by cells (1) and (4). This means that the model (LDA or LR) predicted a snow avalanche day and it actually was a day in which at least one snow avalanche occurred (situation (4)). Similarly, a correct classification can also be a day without any snow avalanche which is correctly modelled as day without snow avalanche (situation (1)). Cells (2) and (3) of the table are indication of wrong classifications. In the case (2), the model predicts no snow avalanche, whereas the day into consideration was a snow avalanche day in reality. In the case (3), the model predicts at least one snow avalanche, whereas during the day into consideration, no snow avalanche occurred. With respect to costs related to misclassifications, situation (2) is more dangerous for people and users of the infrastructure than situation (3). In situation (2), snow avalanches should not occur, so infrastructures remain accessible for people, who are endangered because at least one snow avalanche actually occurs. In situation (3), at least one snow avalanche is expected, so security measures are taken to avoid incidents, even if no snow avalanche occurred in reality. However, in the present work, no differentiated costs will be assigned to these two misclassification situations. Indeed, costs of both situations can be elevated, especially if the ski resort is closed due to snow avalanche danger whereas no snow avalanche occurs, because lots of money can be lost for the station.

	Day without snow avalanches (predicted)	Day with snow avalanches (predicted
Day without snow avalanches in reality	**(1)**	**(3)**
Day with snow avalanches in reality	**(2)**	**(4)**

Table 2: Contingency table for days with and without snow avalanches in the reality and after a prediction.

For statistical methods used in the present work, cross-validation is applied before tabulating the RCR as explained above. Cross validation is a technique discovered in the early 30s, which aims at evaluating the performance of different algorithms (Arlot S., Celisse A., 2010). The main idea is that "*Part of data (the training sample) is used for training the algorithm, and the remaining data (the validation sample) are used for evaluating the performance of the algorithm*" (Arlot S., Celisse A., 2010). In our case, a particular case of cross-validation is used: the *leave one out cross-validation*. As explained by Arlot S. and Celisse A. (2010), "*Each data point is successively "left out" from the sample*

and used for validation". This technique is set by default in the software R, when adding the option "CV=TRUE" for discriminant analysis.
Concerning Logistic Regression with the software R, no default option for cross validation is available in the function `glm()`. In the present work, Lutz Dümbgen implemented a function in R, which performs the usual *leave one out cross validation.* This function has two arguments: X, a data matrix of explanatory variables, and Y, a vector of binary response (in our case 0 or 1 for days with or without snow avalanches). The output is a contingency table made up of actual and predicted class label values of Y with cross validation. For more details, see Appendix 1 to 3 at the end of this work.

To summarise, LDA and LR are performed with cross validation. Then, real observations of snow avalanche days are compared with expected values of snow avalanche day (0 or 1) modelled by LDA and LR, respectively. Finally, a contingency table as the one presented above is created, to assess right classifications and misclassifications. Performance of different statistical methods is expressed as a percentage of right classifications compared to the all modelled observations.
Once performance of LDA and LR are assessed, two other methods are used for comparison of statistical methods performance. The first is the method of p-values for classification (Dümbgen, L., Igl, B.-W., Munk A., 2008) and the second is the k-nearest neighbour analysis.

Cross-validated p-values for classification

The method of p-values for classification can be applied when we have a vector of different classes (in our case two classes defined by 0 or 1), and explanatory variables (meteorological and snow conditions variables). The aim of this method is to calculate p-values for each observation (based on the different explanatory variables), and to compute confidence regions for their classification. (Dümbgen, L., Igl, B.-W., Munk A., 2008). In other words, even if there are some classification problems which are intrinsically difficult, there may be cases which can be classified with a high confidence. On the other hand, even in "easy" classification settings, there may be cases in which a unique classification is difficult. The method of p-values for classification is useful for such cases, because it assesses if an observation can be classified in the class [1], [2], in both classes or none of the two classes. In R software, the function "`cvpvs.logreg(X, Y, tau.o= 1, find.tau=FALSE,…)`" developed by Niki Zumbrunnen and Lutz Dümbgen is used in the present work. This function "*computes cross-validated nonparametric p-values for the potential class memberships of the training data* [and] *the p-values are based on 'penalized logistic regression*" (Dümbgen, L., Igl, B.-W., Munk A., 2008).

The output of this function is, for each case, a pair of p-values for the potential class memberships. In the table below, the first row corresponds to the first observation, which has a p-value equal to 0.5 for being classified in class [1], and a p-value equal to 0.47 for being classified in class [2]. With alpha taken equal to 0.05 (a confidence level of 1-alpha, which means 95%), this observation can be classified in the first class because 0.5 is greater than alpha, but in the second class too, because 0.47 is also greater than alpha. In this case, the observation can be classified in both classes, with a slight greater probability of being classified in class [1.] The second observation corresponding to the second line has a p-value equal to 0.01 for being classified in class [1], and a p-value equal to 0.10 for being classified in class [2]. In this case, with alpha also taken equal to 0.05, this observation will not be classified in the first class because 0.01 is lower than alpha.

	[,1]	[,2]
[1,]	0.50	0.47
[2,]	0.01	0.10

Table 3: Probabilities for observations 1 and 2 to be classified in the class [1] or [2].

In order to illustrate graphically the output of the p-values calculation, the function "`analyze.pvs (pv, Y = NULL, alpha = 0.05, roc = TRUE, pvplot = TRUE, cex = 1)`" developed by Niki Zumbrunnen and Lutz Dümbgen is available in the software R. It permits to visualize the p-values, to create ROC curves and to table the classification rate for each class. The difference with other tables of classification rate is that the possibility for an observation to be classified in none of the two classes or in both of the two classes is also indicated. This is due to the calculation of confidence regions, which allows different classifications. In the present work, classifications with 80%, 90% and 95% confidence are computed and compared. For further information about p-values, see Dümbgen L., Igl B.-W., Munk A. (2008) P-values for classification. *Electron. J. Statist.* (2). 468--493. doi:10.1214/08-EJS245.

k-Nearest Neighbours

The method of k-nearest neighbours (kNN) appears as very intuitive to classify objects in specific classes according to the classes of their nearest neighbours. The principal idea of kNN is to calculate distances between the object to classify and the other objects (Thirumuruganathan S. (2010, 17th May); Ripley, B. D. (2002); Geoffrey J. McLachlan (1992)). Different types of distances are available, but in the present work, the Euclidian distance will be used. After having calculated all distances between the object to classify and its neighbours, different options of classification can be chosen. First, it could be decided to take into account only the nearest neighbour, and assign the same class of it to the observation to classify. This is the 1-nearest neighbour method (Thirumuruganathan S. (2010, 17th May)). However, it can also be decided to select more than one nearest neighbour (in the present work 3 and 5). In this case, decision on which class will be assigned to the object to classify is taken on the basis of majority of voting among the classes of the nearest neighbours (Ripley B.D., 2002). For example, on day has to be classified in "snow avalanche day" or "no snow avalanche day". Distances are calculated between this observation and all remaining observations; the 5 nearest neighbours are selected; 4 of them are days with at least one snow avalanche and 1 of them is a day without any snow avalanche. With the majority of voting, the day to classify will be assigned to the class of days with at least one snow avalanche.

As for all other statistical methods used in this work, leave one out cross validation is applied for kNN analysis, with a function programmed by Lutz Dümbgen. This function has three arguments: X, a data matrix of explanatory variables, Y, a vector of binary response (0 or 1), k, the number of nearest neighbours to be considered. The output is a contingency table with cross validated estimates of days with or without snow avalanches and real observation days with or without snow avalanches. For more details about this function, see Appendix 3 (R-code for this function).

Conclusion

The first goal of statistical analysis for the present work is to find which variables are important in triggering snow avalanches in the ski resort of Aminona. However, it is also important to check the RCR of all observations, based on these important variables. In this way, cross validation and contingency tables are performed for LDA and LR, but two other methods are also used to get stronger confidence in classification rates. So, even if kNN and p-values for classification do not allow a selection of significant variables for triggering snow avalanches, they are used to compare the RCR of the different methods.

IV. STATISTICAL LINK BETWEEN SNOW AVALANCHES AND METEOROLOGICAL DATA

a. Preparation of the database

This first part of the work is a very important step before going further in the analysis. In fact, the precision and the validity of further statistical analyses will depend on the quality of the input data. In this way, the first thing to do is to prepare the database which will be used later. This step means selecting the variables used, cleaning errors, constructing new variables on the basis of others and checking if variables are coherent with each other. The basis variables are presented in table 1, part II "NIVOLOG PRESENTATION AND DATA".

a.1. Selection of causal variables

a.1.1 Variables left out

Causal variables are meteorological and snowpack variables which can be a cause of snow avalanches occurrence. In this way, 24 variables are left out of the analysis. First, variables must be quantitative or ordinal to be statistically analysed. So, the 4 variables about pictures and comments are omitted for the statistical analysis, but not deleted because they can be useful in order to illustrate snow avalanches events in a second step. Secondly, variables which cannot be the cause of a snow avalanche event are omitted too. In this way, the local and regional risk, going from 1 to 5 on the European scale, constitute a consequence of specific weather and snow conditions, but not a cause for a snow avalanche to occur. Consequently, these 2 variables are left out of the statistical analyses. Under these same conditions, all variables concerning snow avalanches (number of snow avalanches in 24h, mean magnitude, number of shots, age of the last snow avalanche, etc.) are a consequence of weather and snow conditions. In this way, they constitute control variables for the statistical analysis. For example, if the model predicts a snow avalanche for one particular day in the past, do a snow avalanche really occurred? Consequently, the 9 variables about snow avalanches are not used as causal variables, but grouped in one "effect or consequence variable": "*day with or without snow avalanches*".
Then, only meteorological and snowpack variables remain, but are not all causes of snow avalanches occurrence. Thus, the 2 variables "cumulative snowfalls over the whole season (cm)" and "Thickness of the snow cover (cm)" are useful to check the coherence of other variables like, for example, "Snowfall in 24h (cm)". However, they are not a direct cause of the occurrence of snow avalanches. They are rather a consequence of the variable "Snowfall in 24h (cm)", and give an overall view of the snow cover during one particular winter. In this way they are omitted for the statistical analysis, but kept as "control variables" for the coherence tests.

In a further step, variables containing information which is redundant compared to other variables are omitted too. This is the case for the variables "Snowdrift Index in 72h" and "Snowdrift direction in 72h". These latter are omitted because the information about snowdrift is given by the same variable for 24 hours. Other variables, like hoar, are known to be badly recorded at this measurement site (internal communication METEORISK), and lots of values are missing. In this way, the variables "Thickness of surface frost" and "Age of the last icing" are left out too. The variable "age of the last snowdrift" is also left out of the analysis, because the transport by wind can influence the structure of the snowpack for a long time after the event (Bolognesi R., 2015). In this way, this variable is not predictive for a day by day analysis.

The last variable which is left out of the further analysis is “Humidity of the air”. The decision not to take into consideration this variable was taken in agreement with the office METEORISK, for the reason that the humidity of the air has not already appeared to be predictive for the occurrence of snow avalanches.

a.1.2 Variables considered for the statistical analysis

In his book “Estimer et limiter le risque avalanche” (2013), Robert Bolognesi gives typical contexts leading to the occurrence of snow avalanches. The first one is the input of new snow, by snowfall or by wind transport, and the second one, the input of water, either by rain or by a warming of the snow cover (Bolognesi R., 2013). Furthermore, in the chapter 7 of the book “Guide Neige et Avalanches : connaissances, pratique, sécurité”, some variables are presented in order to assess the stability of the snowpack (Pahaut E., Bolognesi R., 2003). On the basis of these two books, variables which seem important in triggering snow avalanches have been selected. They are listed below.

Snowfall in 24h (cm) – HN24

This variable is measured every day, on a plane surface, which is cleaned after each measurement. It is the most intuitive variable, because snow avalanches can occur when fragile surface snow can be mobilized. A fragile layer of snow means that the crystals and the grains of snow are not well bound with each other. In the case of fresh snow, a large proportion of air is still present in the snow pack (Ancey C., et al., 2003). Snow transformations due to compaction, warming, etc. have not already permitted to the crystals to link with each other, and thus, the top layer after a recent snowfall is often very fragile. In this way, snow avalanches are more likely to occur when new and fragile snow is added to the snowpack. Furthermore, a recent snowfall has the effect to add weight to the snowpack, increase the traction forces, leading to possible snow avalanches. Consequently, this variable is often used when statistics on the occurrence of snow avalanches are performed, and is nearly often statistically significant (Bolognesi, R. (2013), Fromm R. (2009), Jomelli V. et al. (2007), Floyer J.A. and McClung M.D. (2003), Floyer J.A (2003), and Saemundsson T., et al. (2003)).

Rainfall in 24h (cm) – R24

Rainfall is measured thanks to a rain gauge, emptied after each measurement. It is also an important variable in the sense that it brings liquid water in the snow cover, increases the water content and the density of snow. Consequently, the weight of the snow cover is increased too and it can more easily slide downward, leading to wet snow avalanches (Bolognesi, R., 2013). Additionally to the effect of increasing weight, the cohesion between the grains can also become smaller due to heavy rainfalls. In fact, if the liquid water content of the snowpack exceeds 12%, it separates the grains from each other and the snow layer loses stability (Ancey C. et al., 2003).

Variation of the snow cover thickness in 24h (cm) – dHs

This variable is measured by making the difference between snow cover thickness of the previous day and the one of the day of the measurement. This variable refers to 2 phenomena. Either the variation is positive meaning that new snowfall or snow transport by wind has occurred during the last 24h, or it is negative, meaning that partial melting or compaction of snow has taken place. In the first case, this variable is comparable to the variable “Snowfall in 24h”. In the second case, the snow cover can be either stabilized by compaction (crystals of snow are better bound with each other), or weakened by an excessive water content like in the case of a heavy rainfall.

Thickness of probe insertion (cm) - Ps

A graduated probe is inserted in the snow cover with the same force at each measurement. Its penetration gives information about resistance, hardness or lightness of snow. This variable is interesting because it measures the state of snow (light or heavy), and also gives information about the cohesion. If the snow grains are not well bound (like in the case of fresh snow), the probe will easily penetrate the snow layer; in the contrary, when the cohesion between the grains is strong, the probe will only penetrate a small part of the layer. Of course, it is often linked to other variables like density, surface of refreezing, snowfall in 24h or age of the last snowfall exceeding 20 cm. However, it can be viewed as the variable "foot penetration" in other studies, which seems to be a good explanatory variable (Floyer, J.A. (2003), Floyer, J.A. and McClung, M.D. (2003), Fromm R. (2009)).

Wind speed (knots), wind direction (°), snowdrift index (grams) and snowdrift direction (°) – vV, dV, ID, dD

Wind speed is presented by authors as important to study snow avalanches occurrence (Eckerstorfer M., Christiansen H.H., 2011, Saemundsson, T. et al., 2003, McCollister C.M. et al., 2002). However, in reality, this is not directly wind which matters, but the transport of snow by the wind, which is presented by other studies: Bolognesi, R., 2015; Chritin V., Bolognesi R., Gubler H., 1999; Chritin V., Melly T., 1998; Bellot H., Bouvet, F.N., 2010.
Wind speed and wind direction are measured by an anemometer. The wind speed is expressed in knots (conversion to km/h: 1 knot = 1.85 km/h) and the direction is given by the eight cardinal sectors in degrees (North = 360°, North-east = 45°, etc.). Data available for the station of Aminona are collected at least once a day, at a particular point time. For example, the 26th of February 1997 at 8:00, the wind came from the South-West (225°) with a speed of 30 knots.

Figure 6: Driftometer in L'Alpe d'Huez, France.

For snow transport, measuring devices have been continuously developed since the 1980's, as this variable became very important to predict snow avalanches occurrence (Bolognesi, R., 2015). First, mechanical devices were tested (driftometer, prismatic boxes), and then, optical and acoustic sensors were developed in the laboratory of the *Lac Blanc pass,* in the French region of Isère (SPC (Snow Particle Counter), FlowCapt, ABS (Automatic Blowing Snow Station)). In Aminona, snowdrift data (direction and index) are collected by a driftometer, on the last 24 hours. This measuring device was developed by Robert Bolognesi in 1995. It is made up of eight sensors, which are directed to the eight cardinal sectors (North, North-East, East, etc.), and connected to eight corresponding bags. These sensors can be adjusted along a vertical pole, as showed on the figure 5. Snowdrift Index is expressed in grams, after weighing of the bags, and snowdrift direction is expressed in degrees, for the main direction of the last 24 hours.
In a first step these four variables (wind speed, wind direction, snowdrift direction and snowdrift index) are selected to further coherence tests because they seem important in discriminating days with and without snow avalanches. Indeed, when wind is blowing, it can transport snow, and lead to high accumulations in sheltered zones. The consequences of such accumulations are similar to a fresh snowfall, because new snow is added to the snowpack, increasing the traction forces and destabilizing the whole snowpack (Bolognesi R., 2015). Furthermore, snow grains transported by

wind have generally a lower cohesion between each other, which further destabilises the accumulations (Bolognesi R., 2015).

Cloud cover (oktas) - N

Cloud cover is measured in oktas and represents the part of the sky covered (0 okta = clear sky, 8 oktas = completely covered sky). This variable is related to a high range of various situations concerning snow avalanches. On one hand, if the cloud cover is low, sun shines, warms the snowpack, and can either stabilises it if the warmth is not too high (snow crystals partially melt and link with each other), or destabilise it due to snow melting and addition of liquid water. In another hand, if cloud cover is high, it can either mean that snowfall is occurring, leading to a destabilisation of the snowpack (cf. fresh snow), or that the temperature remains low and no melting occurs (destabilisation or stabilisation depending on the context, as explained above). If the cloud cover is low, high infrared radiation takes place, leading to a cooling of the surface of snowpack, even a freezing of the top layer if liquid water is available, and a stabilization of the snowpack. On the contrary, if cloud cover is high during the night, lower infrared radiation takes place and the top layer remains at the same temperature. As a short conclusion, the variable of cloud cover is very difficult to associate to a stable or instable snowpack. However, in some situations, it can be a confirmation of other variables like snowfall in 24 hours or thickness of surface refreezing.

Age of the last rainfall (days) _ ADP

This variable gives an idea of past conditions concerning rainfall. It is selected in the present work because rainfall during one day can have an influence on following days. As it has been showed for the rainfall in 24 hours, liquid water exceeding a certain threshold leads to a destabilisation of the snowpack. This instability can be prolonged on several days if no cooling or refreezing takes place. In this way, the age of the last rainfall in days is selected for further analysis, in order to take into account past conditions.

Age of the last snowfall exceeding 20 cm (days) – ADN20

As presented for the age of the last rainfall in days, this variable also gives information about past conditions of recent snow. However, with respect to snowfall, the present variable also gives information about transformation of snow. As already explained, fresh snow often represents a fragile layer because bonds between the grains are not strong. After some days, fresh snow has been influenced by meteorological conditions (temperature variations, action of the wind and of the sun, etc.) and by compaction, leading to a stronger cohesion between the grains (Ancey C., et al., 2003). In this way, the age of last snowfall exceeding 20 cm is an indicator of the stabilization of recent snow, and selected to represent past conditions for the analysis. The threshold of 20 cm has been defined by Robert Bolognesi: in areas where snow avalanches are regularly triggered by artificial means in order to secure ski resorts, nearly no snow avalanches occur with a recent snowfall lower than 20 cm (internal communication METEORISK).

Thickness of surface refreezing (cm) - RS

The thickness of surface refreezing is measured by making a cut in the snowpack, and then, by measuring the thickness of refreezing on the top of the cut. This is an interesting variable because it gives information on the stability of the top layer (if the surface is frozen, grains of snow are strongly linked to each other), but also on the past conditions (liquid water and temperatures below 0°C are examples of conditions needed for a refreezing). Particularly, this variable could bring interesting information about conditions between two observations (two days). For example, if melting of snow

occurred but the top melted layer refreeze during the night. In this way, this variable is kept for further analysis.

Snow temperature at 10 cm depth (1/10 °C) - Tn

Snow temperature is measured with an electronic thermometer, which has a probe to penetrate the snow pack. After having cut the snowpack vertically, the probe is inserted into it at a depth of 10 cm. The depth of 10 cm has been chosen because the snow is not too much influenced by the sun radiation and the air temperature. This variable is very interesting because snow temperature is in relationship with the structure and transformation of snow grains (Ancey C., et al., 2003). If snow temperature is low, the snow is dry and the transformation of grains is slower than for a snow temperature near 0°C, which permits a partial melting of grains and more bonds between them. In this way the variable of snow temperature is selected in the present work, to give information about the state of snow at 10 cm depth. Furthermore it is considered as important by some authors, when predicting snow avalanches occurrence (Pahaut E.,Bolognesi R., 2003).

Air temperature (1/10 °C) - Ta

Air temperature is measured by a thermometer and recorded each day. This variable is important in studying snow avalanches because it directly influences snow by transformation of crystals and grains (Ancey C., et al., 2003), as well as snow temperature by diffusion through the top layer. As for snow temperature, air temperature acts as a compaction factor if it is near or superior to 0°C (partial melting of snow crystals and more bonds between them), or as a preservative of snow conditions if it is lower than 0°C. Furthermore, this variable has appeared in many studies to be an important one, when studying the occurrence of snow avalanches (Pahaut E.,Bolognesi R., 2003; Floyer, J.A. and McClung, M.D., 2003; Jomelli V. et al., 2007).

Air temperature variation in 24h (1/10 °C) - dTa

This variable is constructed on the basis of the difference between air temperature of the present day and of the previous day at the same time. It is considered as important because it gives an overview of past conditions, which have influenced the snowpack during the last 24 hours. It also gives more detailed information about air temperature because this latter is only measured on a point in time. More particularly, it can indicate either a warming or a cooling, which influences the transformation of snow, as already presented before (Ancey C., et al., 2003).

Density of the surface stratum (kg/m3) - MVS

Density of snow is measured each day by a corer of half a litre, which is inserted into snowpack and weighed. The weight of snow for half a litre is then converted in kg per cubic metre. In the chapter 7 of the book « Guide Neige et Avalanches : connaissances, pratique, sécurité » R. Bolognesi and E. Pahaut emphasise that « *La connaissance de paramètres internes comme le type de grains, la cohésion, la température, la densité, la teneur en eau liquide est essentielle pour évaluer la stabilité du manteau neigeux* » (the knowledge of internal parameters as type of grains, cohesion, temperature, density, liquid water content is essential to assess the stability of snowpack) (Pahaut E., Bolognesi R., 2003). In other words, density gives information about cohesion of snow grains and so, about stability of snowpack. As already explained, fresh snow is considered as fragile because a large proportion of air is present in it and the bonds between snow crystals are weak. When compaction takes place, the space between grains is reduced, the proportion of air decreases, the bonds get stronger and the density increases too (Ancey C., et al., 2003). In this way this variable is very interesting in order to give an overview of the snowpack stability.

Snow avalanches in 24h - AVAL

This last variable is the second type of variable ("response variable" or "dependant variable"), which we will try to explain by the meteorological and snowpack variables during statistical analyses. The total number of snow avalanches in 24 hours includes artificial, natural and accidental snow avalanches in 24 hours, and is reduced to a binary variable: day with (taking the value 1) or day without (taking the value 0) at least one snow avalanche.
At this stage, 17 variables are selected for the statistical analysis, and will be check in the next chapter, in order to eliminate errors or dubious values.

a.2. Coherence tests

Once causal variables are selected for further analysis, it is important to identify wrong values or errors for each of the 17 variables. As already emphasised, the quality of statistical results only depends on the quality of input data. This chapter is dedicated to the identification of errors or dubious values in order to obtain a dataset as clean as possible for further analysis.
The various errors found in the database can have different causes. First, they can be due to typing errors at the moment of entering data in the file: for example, missing of the minus sign, addition of a zero, mixing of numbers, etc. Second type of error is done at the time of the measurement, due to malfunctioning of the measuring device or due to lack of concentration by people doing the measurement. For example, on a 10 cm - graduated pole fixed in the snowpack, confusion between different graduations is possible. Thirdly, errors can also be due to changing in staff. Over 19 years, more than four different people have worked at the measurement site. So, as a new employee arrives, information about how doing measurements, at which time of the day, how many times, with which precision, is not always transmitted correctly. This effect was observed when doing coherence tests in the present work. For example, before the year 2008, the value for the age of the last snowfall exceeding 20 cm never takes the value of zero, even if the snowfall during the last 24 hours was higher than 20 cm. The value of the variable was updated only the day after the snowfall. Then, since 2008, whenever a snowfall exceeds 20 cm in 24 hours, the value of the age of the last snowfall exceeding 20 cm is equal to zero. Finally, error codes are also found in the database, under the form of 999.

Coherence tests have been performed on the following way, for each variable:

1) Testing if some values are out of the domain of definition, or if a contradiction exists with other values of the same observation. In the table below, this corresponds to wrong values (second column).
   Wrong values lead to elimination after complementary verifications.

2) Testing if values outside the norm (98^{th} and 2^{d} percentiles) are still coherent compared with other values of the same observation. In the table below, this corresponds to dubious values (fourth column).
   Dubious values lead to further investigation in order to keep them or not.

All tests are summarized in the table 4, below.

Variable	Wrong value if	Number of errors %	Dubious value if	Number of errors %	Remarks
Total number of snow avalanches in 24 h **AVAL**	Artificial snow avalanches in 24h + accidental snow avalanches in 24h + natural snow avalanches in 24h ≠ total number of snow avalanches	**13** **0.52 %**	Artificial snow avalanches in 24h + accidental snow avalanches in 24h + natural snow avalanches in 24h = 0 and total number of snow avalanches>0 OR Artificial snow avalanches in 24h + accidental snow avalanches in 24h + natural snow avalanches in 24h >0 and total number of snow avalanches=0	**4** **0.16 %**	Only values which change the classification in day with or without snow avalanches occurrence are considered as wrong. Example: a day with a total of 23 snow avalanches, but a sum for all kinds of snow avalanches equal to 24 is considered as possible, and kept in the statistical analysis.
Cloud cover (oktas) **N**	N < 0 or N > 9	**0** **0 %**	/	**/**	No coherence tests because this variable is linked to very different weather situations. Example: one day with cloud cover = 9 and no precipitation and another day with cloud cover = 0 but precipitation during the last 24h.
Thickness of surface refreezing (cm) **RS**	RS < 0	**0** **0%**	RS > 0 and dTa > 0 OR RS > 0 and Ps > 0 OR RS > 0 and Ta > 0 OR When RS > 0, Ps > RS	**7** **2.28%**	/
Variation of snowpack thickness in 24h (cm) **dHs**	dHs J ≠ Hs J – Hs J-1	**84** **3.78 %**	dHs > 0 et Hn24 =0 and dHS>0 et ID=0	**138** **5.52%**	Most of the errors are due to a positive variation of the snowpack thickness even if no snowfall or no snowdrift took place during the last 24h.

Density of the surface stratum (kg/m3) **MVS**	50 > MVS > 600 (Values found in the literature)	**17** **0.68%**	HsJ < HsJ-1 and MVSJ-MVSJ-1 < -20	**125** **5.00%**	The difference of 20 is considered as variations in the place of the measurement.
Snow temperature (1/10°C) **Tn**	Tn > 0 Tn < -30 (Values found in the literature)	**24** **0.96%**	Ta < 0 et Tn > 0 Tn > Ta	**42** **1.68%**	Most of the errors in the second test are due to missing values.
Thickness of probe penetration (cm) **Ps**	Ps < 0	**0 (0%)**	Ps > 98th percentile Hn24 > 0 et Ps = 0 RS > 0 et Ps > 0	**35 (1.40%)** **30 (1.20%)** **21 (0.84%)**	/
Air temperature (1/10°C) **Ta**	/	**/**	Ta > 98th percentile Ta < 2d percentile	**45 (1.80%)** **46 (1.84%)**	/
Variation of air temperature in 24h (1/10°C) **dTa**	dTa ≠ Ta J – Ta J-1	**22 (8.49%)**	dTa > 98th percentile dTa < 2d percentile	**44 (1.76%)** **45 (1.80%)**	First test is performed only on 259 observations. The other 2243 observations were missing and had to be recalculated on the basis of Ta, so, already correct. Most often, the values exceeding the 98th or the 2d percentile are typing errors or error in the calculation of the variation.
Rainfall in 24h (mm) **R24**	R24 < 0	**0 (0%)**	R24 > 0 and Ta < 0°C	**18 (50%)**	Only 36 cases with rainfall over the whole dataset.
Age of the last rainfall (days) **ADP**	ADP > 0 and R24 > 0 ADP = 0 and R24 = 0	**30 (1.20%)** **5 (0.20%)**	/	**/**	917 observations are missing because measurements are only effectuated since the year 2000. Problem noticed for the years 2008 to 2012: For the first observation of the winter season, ADP = 0 even if no rainfall occurred during the last 24h.
Snowfall in 24h (cm) **Hn24**	Hn24 < 0	**0 (0%)**	Hn24 > 98th percentile	**41 (1.64%)**	/

Age of the last snowfall exceeding 20 cm (days) **ADN20**	Hn24 > 20 and ADN> 0	**148 (5.92%)**	AND > 98th percentile	**45 (1.80%)**	About half of the errors are due to a confusion between the value 0 or 1 for the age of the last snowfall exceeding 20cm.
Direction of the wind (°) **dV**	dV is not defined according to 45° steps. (North = 360, North-east = 45, East = 90,…)	**110 (4.40%)**	dV ≠ dD and, conversely, dD ≠ dV.	**1489 (59.12%)**	The majority of the errors are due to undefined direction for the wind or for the snowdrift. EX: dD = 270 and dV = /.
Direction of the snowdrift (°) **dD**	dD is not defined according to 45° steps. (North = 360, North-east = 45, East = 90,…)	**61 (2.44%)**	dV ≠ dD +/- 45° and, conversely, dD ≠ dV +/- 45°.	**1384 (55.32%)**	If tests are performed only on complete data (elimination of undefined values), the percentage of error is of 93 errors over 504 observations (18% of errors).
Wind speed (kn) **vV**	vV < 0	**0 (0%)**	vV > 98th percentile vV < 2d percentile	**47 (1.88%)** **/**	No dubious value for the 2d percentile because it is equal to 0.
Snowdrift Index (g) **ID**	Error value 999	**63 (2.52%)**	ID > 98th percentile ID < 2d percentile ID>0 and vV=0 ID>0 and vV < 6.8 ID>0 and vV < 9.7 ID>0 and vV < 13.6	**63 (2.52%)** **/** **72 (2.88%)** **232 (9.27%)** **298 (11.91%)** **389 (15.55%)**	The 999 values are kept, but highlighted as potentially wrong values. No dubious value for the 2d percentile because it is equal to 0. Assumption: no wind, no snowdrift. Threshold of F. Naaim-Bouvet et al. (2011) Minimum threshold of Vionnet V., 2012. Maximum threshold of Vionnet V., 2012.

Table 4: Summary of cleaning and coherence tests for all causal variables selected.

All the previous cleaning tests show that errors are present in the dataset. However, for snowdrift direction and wind direction, the number of errors is high compared to other variables (see later).

a.3. Cleaning of errors

a.3.1. General variables

In the following, each variable is considered with respect to the previous coherence tests, and the values which need to be corrected, suppressed or accepted after verification, are exemplified.

This chapter only contains practical information about cleaning tests, and is not essential for understanding of the whole work. It can be skipped by readers more interested in results.

Cloud cover (oktas)

For cloud cover, no value appears to be out of the domain (0 to 9). Furthermore, this variable can be related to many different meteorological situations, which makes it difficult to test with respect to other variables. For example, an observation with cloud cover of 9 oktas does not necessarily mean that snowfall or rainfall took place, it cannot give information about the temperature or the density of snow. In this way, as no obvious errors are identified, all the values are kept for the statistical analysis.

Total number of snow avalanches in 24 h

Total number of snow avalanches in 24h is the result of addition of accidental, artificial and natural snow avalanches in 24h. In this way, errors are due to wrong calculation. Two types of errors are possible.
For the first one, wrong calculation leads to a different value, but which remains positive. In other words, the total of snow avalanches for the observation is in any cases positive, meaning that this day is classified as a "snow avalanche day" because at least one snow avalanche occurred.

Example 03-01-2012: for the last 24 hours of this day, 20 artificial, 0 accidental and 5 natural snow avalanches occurred. However the variable "total number of snow avalanches in 24 hours" gives a value of 60. In any cases, snow avalanches occurred this day, and the right value is simply calculated again, leading to a value of 25.

For the second type, wrong calculation can change the day in "snow avalanche day" or "day without snow avalanches". In these cases, values of "total number of snow avalanches in 24h" are deleted, to avoid bias in the analysis.

Example 23-12-2011: for the last 24 hours of this day, 20 artificial, 0 accidental and 40 natural snow avalanches occurred. However the variable "total number of snow avalanches in 24 hours" gives a value of 0. Here, the day could be either with or without snow avalanches. Thus, the value is suppressed.

This particular variable is very important in the present work because it represents the binary dependent variable that we try to explain thanks to other independent variables. In this way, decision is taken to delete all observations for which the occurrence of a snow avalanche is not known, to improve further statistical results. This decision leads to a suppression of 235 observations in which it was not possible to know if snow avalanches occurred or not, and a final dataset of 2267 complete observations concerning the variable of snow avalanches (file "DC41MTA01 - Cleaning a.3").

Thickness of surface refreezing (cm)

Correction: After a comparison with other values of the same observation, errors are identified as typing or measurement errors. In this way, they are rounded or corrected.

Example 30-01-2013: the value for the thickness of refreezing is 0.1 cm, which is very low. Furthermore, it seems incoherent with the value of the air temperature (5.8°C). Considering that the value of 0.1 is dubious, it is rounded to 0.

Acceptance: after further verification and comparison with other variables of the same observation, some values appear to be coherent.

Example 02-02-2002: there is 1 cm of surface refreezing, with a probe penetration of 6 cm. In this case, it is possible that the probe had been able to break the surface of refreezing, and to partially penetrate in the snowpack. In this way, the value of 1 cm is plausible, and kept for further analysis.

Removal: because incoherent with other variables of the same observation, and no identification as typing errors.

Example 05-01-2013: there is 6 cm of surface refreezing for this date, with an air temperature of 2°C, no rainfall or snowfall during the last 24 hours, and no refreezing the day before and the following day. Furthermore, the probe can penetrate 16 cm in the snowpack, which is not possible with a thickness of refreezing of 6 cm. In this way, this value is left out of the analysis.

Variation of snowpack thickness in 24h (cm)

Tests have been performed with the variable "snowpack thickness", which acts as a reference. Difference between the actual observation and the one of the day before should be equal to the value of the variation of snowpack thickness in 24 hours. Errors which are found in the database can lead to three different actions.

Correction: if it is a calculation error or if the variation is very low (possible transport of snow by weak winds)

Example 20-01-2012: the variation of snowpack thickness is 15 cm but the difference between the snowpack thickness of the day before and the one of the considered day is 12 cm. In this case, the variation remains positive, and the error is only due to a wrong calculation. In this way, the value is corrected according to the real difference, and kept for further analysis.

Acceptance: because the errors are due to changing from one season to the other (for example from the 22-04-2004 to the 13-12-2004), or due to two observations during one day (at 8:00 and 14:00). In these cases, values appearing as wrong are accepted, and kept for further analysis. Another case of acceptance is when the variation is positive, no snowfall occurred during the last 24 hours, but snow transport by wind took place. In these cases, it is possible that the snowpack thickness increases, even if no snowfall occurred.

Example 28-02-2001: No snowfall occurred during the last 24 hours, but the variation of the snowpack thickness is 20 cm. however, the value for the snowdrift index during the same period of time is 40 grams. In this way, it seems possible that the increase of snowpack thickness is due to snow transport by wind.

Removal: when the variation of snowpack thickness is largely positive but no snowfall or snow transport by wind occurred during the last 24 hours. This scenario leads to an important number of errors, all occurring when the variation is positive and cannot be explained by other variables. Another situation leads to the removal of the value: the variation of more than 100 cm in 24 hours. Especially at the end of the winter season, values of the snowpack thickness jump from high values to very low values. This may be due to oblivion of measurement during several days by nice weather,

high temperature, and a new measurement after some days, with a sharp decrease in snowpack thickness. However, these cases are removed, because they constitute a bias in the analysis.

Example 07-02-2000: there is +35 cm of variation without any snowfall or snow transport by wind.
Example 12-04-2004: the snowpack is 285 thick. The following day, the value is zero, which gives a variation of -285 cm in 24 hours.

After the tests about variation of the snowpack thickness, too many errors appeared for a positive variation with no obvious reasons. Later in this work, we will modify this variable in order to encounter this problem (cf. Chapter IV.a.4).

Density of the surface stratum (kg/m3)

The domain of definition for this variable is from 50 kg/m3 (fresh and cold snow, with very high air content and very light) to 600kg/m3 (old snow, which has been compacted and partially melted after a relatively long period of time) (Ancey C., Sergent C., Martin E., 2003; Bolognesi R., 2015). As for other variables, three decisions are taken in order to clean the dataset.

Correction: because identified as typing errors.

Example 11-01-2008: density is equal to 20 kg/m3, while no snowfall occurred during the last 24 hours, and conditions of the following and previous days are the same: previous day with a density of 190 kg/m3, following day with a density of 220 kg/m3. Furthermore, a density of 20 kg/m3 for the snow is nearly never observed in the field. In this way, the value 20 appears to be a typing error, where the zero has been omitted. Thus, this value is corrected from 20 to 200 kg/m3, which remains plausible with respect to observations of the following and previous days.

Acceptance: even if values seem dubious, they can be plausible after comparison with other variables of the same observation.

Example 13-04-1994: density is 100 kg/m3 with respect to a density of 160 kg/m3 the previous day, with a packing down of snowpack equal to 2 cm, which seems dubious. However, snowfall in 24 hours is 10 cm, with an air temperature of -8°C (which corresponds to a density of 50 kg/m3 to 150 kg/m3 (Ancey C., Sergent C., Martin E., 2003; Bolognesi R., 2015)). The value of 100 kg/m3 is plausible. It is possible that the snowpack melted of 12 cm the day before, but it snowed during the night, thus the density is lower than the previous day.

Removal: because wrong compared to other variables of the same observation or error code 999.

Example 13-04-2004: density is 999 kg/m3. This value is obviously wrong because even the density of pure ice is about 920 kg/m3. Furthermore, no value with a precision of one tenth is present in the dataset. In this way, this value appears as an error code and is removed for further analysis.

Snow temperature (1/10°C)

The domain of definition for this variable goes from 0°C to -30°C. However, some cases in which the snow temperature is slightly higher than zero are possible, by nice weather, and positive air temperature, in spring. For this variable, many values are missing. This is why many errors in coherence tests with other variables appear. The three decisions to clean the dataset are similar to other variables.

Correction: because identified as a typing error, which has been proved by other variables of the same observation or of the following/previous days.

Example 15-04-2001: snow temperature is equal to 12°C, which is not possible with respect to the domain of definition. Furthermore, the air temperature is equal to -4.5°C and the thickness of refreezing is 2 cm thick. In this way, the value of 12°C is wrong. However, the previous day, snow temperature is equal to -5.6°C and the following day, -7.6°C. Cloud cover is equal to 0, so it is possible that the night was clear too, that radiation was important and that snow temperature decreased. Thus, it seems that the minus sign is missing for the value of 12°C. So, it is corrected in a snow temperature of -12°C, which is more close to the reality.

Acceptance: because close to the domain of definition, and plausible after comparison with other variables of the same observation. Mainly, the values slightly higher than zero are considered as possible in spring time or when the air temperature is positive for the same observation.

Example 22-03-2010: snow temperature is equal to 7 (0.7°C). This value is close to 0°C and can be possible if snow is partially melted, wet, in spring. Here, the date of the observation is the 22d of March and the air temperature is 2°C, so the value of 0.7°c is considered as plausible, and kept for further analysis.

Removal: because out of the domain of definition or error codes 999.

Example 13-04-2004: snow temperature is equal to 999. This code has already appeared in other variables like the density. In these cases, it is not possible to find the value which should be written instead of the error code, and thus, not possible to correct it. In this way, the value 999 is removed for further analysis.

Thickness of probe penetration (cm)

The only restriction for the probe penetration is that it cannot be negative or superior to the snowpack thickness. For this variable, no correction is possible. So, the values are either removed or accepted as they are.

Removal: because out of the domain of definition, considered as too incoherent with respect to other values of the same observation, or error code.

Example 14-02-2011: the probe penetrates 49 cm but the surface of refreezing is 4 cm thick, and no snowfall occurred during the last 24 hours. The probe cannot break the ice recently formed at the top of the layer. Furthermore, no snow transport by wind occurred. So, no fresh layer of light snow can have covered the surface of refreezing, and permit a penetration of the probe. In this way, this value is too incoherent, and removed for further analysis.

Acceptance: because the dubious value has been verified and considered as plausible when compared to other values of the same observation.

Example 20-02-2008: the probe penetrates 17 cm and the surface of refreezing is 0.5 cm. As the thickness of refreezing is close to zero, it is possible that the probe broke the refreezing layer and penetrates the snowpack anyway. In this case, the value of 17 cm is accepted.

Air temperature (1/10°C)

For the air temperature, positive and negative values are correct. So, the only test which can be applied is about dubious values. This means values below the 2d percentile or over the 98th percentile. Then, these values are compared with values of the same observation to see if they are plausible or not. Once this is done, the three decisions are the same as for previous variables.

Correction: because the dubious value has been identified as typing error.

Example 17-01-2002: air temperature is equal to 10.4°C, which seems to be erroneous with respect to the date. Furthermore, air temperature of the previous day was -8.2°C and the following day, -6.4. No refreezing has taken place, meaning that no liquid water was available at the surface of the snowpack. In this way, it seems that the minus sign before the 10.4°c has been omitted. So, the value is corrected in -10.4°C, and kept for further analysis.

Acceptance: because the dubious value has been checked and identified as plausible after comparison with other values of the same observation or values of the previous/following days.

Example 11-02-1999: air temperature is equal to -18.3°C, which is situated below the 2d percentile and can be considered as dubious. However, air temperature of the previous day was -17.8°C and the one of the following day is -17.5°C. So, it can be possible than a cold period was taking place during these times, with very low temperatures, and the values are plausible. Then, they are kept for further analysis.

Removal: No value needs to be removed for the reasons that they are out of the domain of definition. However, some of them are considered as too incoherent with respect to other values of the same observation. So, they are removed for further analysis.

Example 05-03-2003: air temperature is equal to 10°C. The previous and following days, air temperature was equal to -1.3°C and -0.5°C, respectively. So, this value seems wrong compared to values of the previous/following days, and is removed for further analysis.

Variation of air temperature in 24 h (1/10°C)

Test for this variable are effectuated on the basis of air temperature differences between the day under consideration and the day before. However, this test is performed only on 259 observations, because before the year 2008, no records of temperature variations have been done. So, the observations for the years before 2008 have been recalculated on the basis of the air temperature, and test is unjustified. Thus, for the 259 observations, errors lead to three possible decisions.

Correction: because dubious values are identified as typing errors.

Example 17-01-2002: air temperature variation is equal to 18.6°C. However, air temperature of the previous day is -8.2°C and the one of the day under consideration is -10.4°C, giving a difference of -2.2°C. So, the wrong value is corrected and replaced by the new difference.

Removal: because values are considered as too incoherent or impossible to modify because values are missing for the previous day or for the day under consideration.

Example 23.12.2012: air temperature measurement was done at 14:00, but the measurement of the previous day was done at 9:00. In this way, a bias exists in the value of the variation of air temperature in 24 hours. Measurements are not effectuated at the same time, so cannot be compared on a time step of 24 hours. So, all cases similar to the present one are removed for further analysis.

Acceptance: because considered as plausible after comparison with other values of the same observation or values of the previous/following days.

Example 29-12-2011: air temperature variation during 24 hours is equal to -9°C, which is situated in the extreme values of the distribution (cf. percentiles). However, when comparing it with other variables of previous/following days, this value can be viewed as correct, and maintained in the analysis. The two previous days, cloud cover was very low, wind very weak and air temperature equal to 3°C and 4°C, respectively. On the 29th of December at 9:00, wind was blowing from North-West at a speed of 17 knots, cloud cover was equal to 8 (completely covered), and snow transport by wind was also coming from the North-westerly direction. In this way, the sharp decrease in temperature during 24 hours can be assimilated as the arrival of a cold front from the North-West, and is plausible.

Rainfall in 24 h (mm)

For the variable of rainfall, only 36 cases of rain have been recorded over the 2502 observations. So, all the values are verified one by one, and checked with other values of the same observation or with values of the previous/following days. No wrong values have been detected during coherence tests, nether typing errors. However, some values appeared to be dubious when rainfall occurred with an air temperature superior to 0°C. These ones have been verified again and finally accepted. The reason for this is that for each rainfall event, a snowfall event took place too. So, it is possible, that precipitations began under the form of rain, then, rain-snow boundary decreased in altitude, or increased, depending on the nature of the perturbations (cold front, warm front, etc.). In this way all dubious values are kept for further analysis.

Example 5 mars 2001: rainfall in 24h is equal to 1 mm but air temperature at 8:00 is equal to -3.5. However, snowfall occurred too, with a total of 42 cm over the last 24 hours. Thus, the value of 1 mm is plausible. Precipitation began under the form of rain, then, under the form of snow till the measurement.

Age of the last rainfall

As the number of cases with a record of rainfall during the last 24 hours is low (36 observations), the values of the variable "age of the last rainfall" is often high. All values are checked with respect to the variable "rainfall", and two options are possible.

Correction: because dubious values are identified as typing or calculation error. Often, the errors are due to an update of the value for the variable "age of the last rainfall" on the day following the rainfall. In these cases, the value of the age of the last rainfall is set to zero for the day in which a rainfall occurred.

Example 06-01-2001: Rainfall in 24 hours is 1 mm. The day before, no rainfall has occurred for 40 days (value of the variable "age of the last rainfall"). The value for this variable is 41 for the day under consideration, but should be zero. In this way, the value 41 is corrected in value 0, and kept for further analysis, after having checked that the age of the last rainfall for the 07-01-2001 is 1.

Removal: because considered as too wrong and no correction is possible due to unknown values. The cases of removals are mainly due to a misinformation on the value given to the first measurement of the season, for the years 2008 to 2012. For these years, at the beginning of the season, the value for the variable "age of the last rainfall" is set to zero, even if no rainfall occurred during the last 24 hours. So, all the values following this first observation are wrong, because they depend on the first measurement. The only way to have correct values for these years is to have a rainfall occurring during the season, leading to valid values for the following observations. In this way, the measurements done at the beginning of the seasons for these years are removed, until a rainfall event takes place.

Example 24-12-2010: the value for the first observation of the winter season 2010-2011 is zero, but no rainfall occurred in the last 24 hours. Then, all the following values are wrong, till the 06-01-2011, in which a rainfall occurred (value of 5 mm). In this way, values for the age of the last rainfall following this event are valid, and kept for further analysis.

Snowfall in 24h (cm)

The domain of definition for this variable is from zero to more than 100 cm. No values are negative, but coherence tests lead to three options. For the winter seasons 1994-1995 and 2007-2008, two or more observations are effectuated during one day. The first observation (effectuated at 8:00) gives the value of the snowfall during the last 24 hours. Observations effectuated at 11:00, 13:00 or 14:00 only give the snowfall height since the first observation at 8:00 (this was verified with the person who made the measurements for these two seasons). In this way, only values corresponding to the snowfall in 24 hours are kept, and other values are omitted for the further analysis.

Correction: because identified as typing errors. For this option, no dubious values need correction.

Removal: because considered as too wrong. The cases of removal are due to missing measurement during the previous days. When this is the case, it is possible that the variable "snowfall in 24 hours" takes a too high value, cumulated over the days without measurements.

Example 26-11-1996: Snowfall in 24 hours is 100 cm (maximal value for this variable). However, no measurements have been done during the three previous days, and it is possible that the value of 100 cm is a cumulated value. Further investigations were done for this dubious value: in the archives of the office METEORISK, no heavy snowfall is indicated for the date of the 26th November 1996. In this way, the value of 100 cm is removed by precaution.

Acceptance: because appear to be plausible after further verification and comparison with other values of the same observation of values from the previous/following day. The values superior to 40 cm in 24 hours (98th percentile), which have a corresponding total number of snow avalanches higher than zero are identified as plausible.

Example 28-12-1999: Snowfall in 24 hours is equal to 50 cm, and the total number of snow avalanches for this day is 32. So, even if the value of 50 cm is be dubious, it is coherent with the high number of snow avalanches for the same observation. Cases similar to this one are kept for further analysis.

Age of the last snowfall exceeding 20 cm (days)

For this variable, values are considered as wrong if the snowfall in 24 hours is superior to 20 cm and if the age of the last snowfall exceeding 20 cm is not zero. The same three options are chosen for errors.

Correction: For a majority of errors, the cause is confusion between the value 1 or 0 for the day in which a snowfall exceeding 20 cm occurred. But the rule is that if a snowfall exceeding 20 cm occurred, the age of the last snowfall exceeding 20 cm is zero for this observation. In this way, for these cases, the values 1 are replaced by 0. Another type of error is that the update of the variable "age of the last snowfall exceeding 20 cm" is only done the following day. In these cases, the values are set to zero for the day with a snowfall exceeding 20 cm.

Example 01-03-2006: Snowfall in 24 hours is equal to 25 cm, but the age of the last snowfall exceeding 20 cm is equal to 10 days. However, the following day, the value is updated, and the age of the last snowfall exceeding 20 cm is 1, which is correct. So, all errors similar to this case are corrected in the same way, and kept for further analysis.

Acceptance: after verification and comparison with other variables of the same observation or of previous/following days. Dubious values (higher than the 98th percentile), are all situated in spring time, at the end of the winter season (March or April). In this way, it seems possible that no snowfall exceeding 20 cm has occurred for two months or more and that values are high. Furthermore, these values follow each other and thus, represent dry periods or years. For example, spring of the year 2011 had particularly low snow cover, with value for the 24-03-2011 equal to 90 (nearly 3 months without snowfall exceeding 20 cm).

Removal: No values need to be removed for this variable.

a.3.2. Case of wind and snowdrift

Snowdrift is the snow transport by wind. In this way, these two variables seem to have a close relationship, concerning their direction and the wind speed/snowdrift index. In other words, if the wind was high and come from the northerly direction, snowdrift index should also be high, and come from the northerly direction too. However, as shown by the coherence tests, wind and snowdrift are not always similar, and cannot be totally explained by each other. This is partly due the difference of measurement in time. If snowdrift direction and index are measured on 24 hours, wind speed and direction are measured in a point in time, at 8:00 for most of the observations. This difference in the measurement leads to consequences on the relationship between wind and snowdrift direction and on the relationship between wind speed and snowdrift index.

- **Wind and snowdrift direction**

In a first step, the domain of definition for the wind and snowdrift directions is defined according to the eight cardinal sectors (North = 360°, North-East = 45°, etc.). So, all the values outside this domain are rounded to nearest sector (for example, a direction of 40° is rounded to 45°, meaning that the wind is coming from the North-East sector). Then, further coherence tests are performed on these new values.
Most of the time, wind direction is variable from one hour to the other, unless strong perturbation with high wind speed is occurring. But in order to transport snow, wind has to blow in a main direction during a certain time and at a certain speed. In this way, wind can have blown during 12 hours from the North, leading to a strong snow transport in this direction. But then, it can have change direction and blow from the East. In this case, at the time of the measurement, snowdrift will be recorded as northerly snowdrift during the last 24 hours, but as easterly wind will be recorded for this observation.

The problem with these two variables is that wind and snowdrift direction are not numerical. Their values are given in degrees with respect to a circle, where 360° indicates the North, 45° the North-East, etc. In this way, it cannot be said that 360 is greater than 45; in reality, these values only indicate different directions. So, to be able to use these variables in statistical analyses, mathematical transformations should be applied to transform wind and snowdrift direction into numerical variables. In the present work, these transformations will not be done, and wind and snowdrift direction will be left out of the statistical analyses. Several reasons for this are presented below.

Wind direction:

The reason to omit wind direction is that for many studies, this is more wind speed which matters, and wind direction does not seem to be important when studying snow avalanche occurrence in a whole ski area (Floyer J.A., McClung M.D. (2003); Eckerstorfer M., Christiansen H.H. (2011); Gassner M., Brabec B. (2002)). As explained by McCollister C.M., Birkeland K., Hansen K., Aspinall R., Comey R. (2002): "[...] *at the scale of the entire ski area, there is not an obvious relationship between avalanche activity on a given aspect and wind direction".* In fact, in some cases, even if wind blows, it does not transport snow (see below). So, the direction of the wind has not a direct relationship with direction of the snowdrift and its influence on the snowpack and on the possible occurrence of snow avalanches. For example, wind can blow from North, but no snowdrift takes place and no snowdrift direction is recorded. In this way, for the present work, wind direction will be left out of the further statistical analyses.

Snowdrift direction:

Even if snowdrift direction delivers more information than wind direction for the prediction of snow avalanches occurrence, it cannot be used in our work. The reason for this is that data from Aminona are not sufficiently well recorded for this variable. In fact, in 75% of cases, snowdrift direction is unknown, and thus, about 1700 observations over 2267 have missing values for this variable. In this way, even if numerical transformation was applied on snowdrift direction to include it in statistical analyses, we would lose 75% of observations due to these missing values. In this way, we prefer to omit the variable about snowdrift direction.

So, for further statistical analyses, wind direction and snowdrift direction will be omitted. The consequence of this removal is that results of analyses about snow avalanches occurrence concern a larger area and not a local scale anymore. This consequence is compatible with the goal of this study, which is studying snow avalanches occurrence over an entire ski resort. Indeed, if wind direction and snowdrift direction are included in the analysis, local characteristics for each snow avalanche path can be emphasized. But in the present work, results are found for the whole ski area of Aminona.

- **Wind speed and snowdrift index**

Generally, when wind is blowing, it can transport snow. But some precise thresholds from which snow can be transported by wind have been defined. These thresholds vary from 6.8 knots in the study of Naaim-Bouvet F. et al. in 2001, to 9.7 - 13.6 knots in the study of Vionnet V. in 2012. In the coherence tests, all these values are tested, giving a percentage of errors from 9.27% to 15.55%. However, the agreement is not always perfect between the two variables because of the type of measurement in time. At the time of measurement, wind can be weak, but with a high snowdrift index. This is possible if wind was very strong during some hours in the 24 hours period. It transported snow as its speed was high, but weakened at the end of the 24 hours period, just before the time of measurement.
On the graph below, the relationship between the two variables is presented (missing values are omitted). The correlation between wind speed and snowdrift index is shown in the box at the top right of the graph. Here, the coefficient of correlation is equal to 0.0874, which means that only 8.74% of the variability of the first variable (snowdrift index) is explained by the second variable (wind speed). The relationship is not obvious. This can be explained by the difference in the measurements in time, but also by snow characteristics which, in certain cases, do not allow snow transport even if wind is strong (surface of refreezing, completely wet snowpack, etc.).

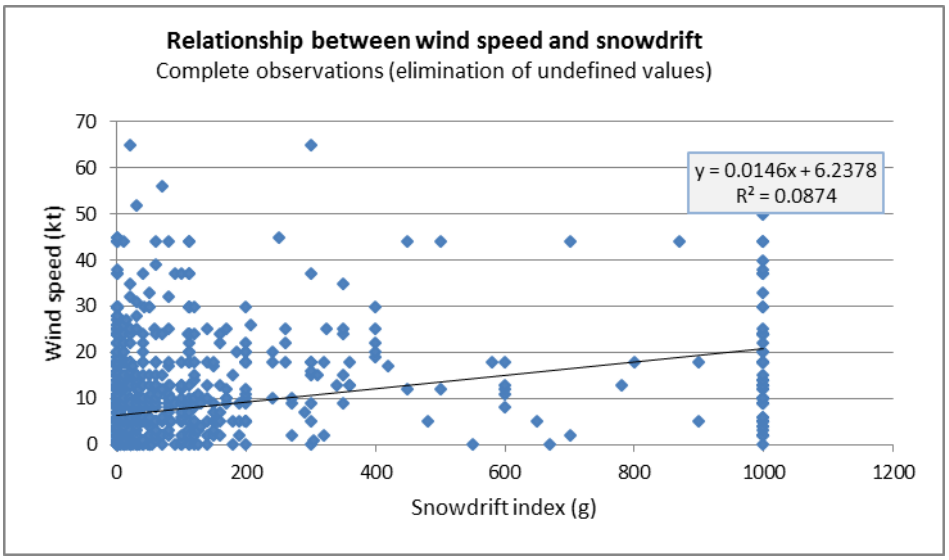

Figure 7: Relationship between wind speed and snowdrift. All values equal to 999 are error codes.

- **Illustration**

A real situation is presented on the following paragraph, to illustrate the mismatching between wind and snowdrift variables. On the 25th March 2006, wind speed at the time of measurement was 9 knots, and it was coming from the westerly direction (value equal to 270°). For the measurement of snowdrift, the value for the direction was 90° (easterly direction) and the index was equal to 10 grams. In this situation, a clear mismatching between these variables appears.
Synoptic situations provided by MeteoSwiss have been found for this date in the archives of the office METEORISK, and are presented to illustrate the mismatching.

The 24th March 2006 **The 25th March 2006**

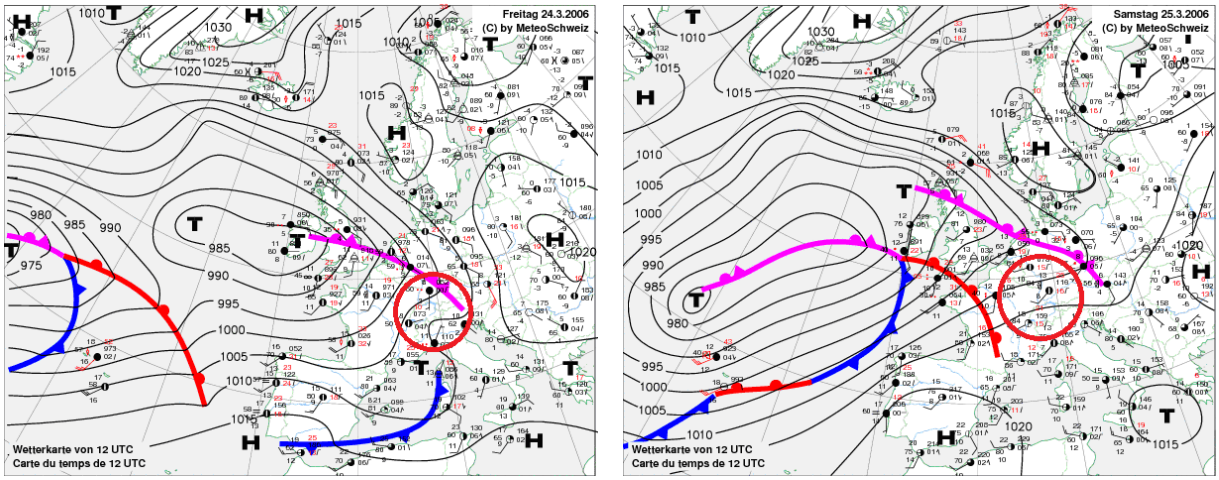

Figure 8: Synoptic situations for the 24th and 25th March 2006, leading to a difference between wind direction and snowdrift direction. The synoptic situations are for 12:00 UTC. The main big circle indicates the position of Switzerland. Little circles are indicators of cloud cover, and the line attached to it indicates the wind direction.

For the 24th of March, sky was completely covered, and wind was coming from the easterly direction at 12:00. One can see that Switzerland was still under a low influence of the cold front situated over the Mediterranean region, which passed over the country on the 23th March and lead to 8cm of fresh snow. So, during the situation of the 24th March, snow was transported by wind in the easterly

direction. On the 25th of March, wind direction has changed since the previous day; it comes from the westerly direction, with the approach of a cold front, situated over France. Furthermore, the cold front over the Mediterranean region has disappeared, and Switzerland is now under the influence of westerly winds. In this way, mismatching between the variables of snowdrift and wind can be explained. On the 25th March, during the last 24 hours, wind blew from the easterly direction, leading to a transport of snow in that way. At the time of measurement (on the 25th March at 8:00), wind was blowing from the westerly direction. So, this actual example illustrates why wind and snowdrift are often weakly correlated.

- **Decision for further analysis**

As previously explained, wind direction and snowdrift direction are left out of the following analyses. The consequence of this decision is that the present study can only focus on snow avalanches forecasting for the whole ski area, and not each snow avalanche path in particular.
Then, as previously shown, wind speed and snowdrift index are not always in good agreement. However, these are two important variables in order to predict snow avalanches occurrence. Wind speed is interesting for the link that it has with refreezing, density of snow, or temperature. Snowdrift index is important because transport of snow often leads to increased snow avalanches activity (Bolognesi R., 2015; Bellot H., Bouvet, F.N., 2010; Pahaut E., Bolognesi R., 2003). In this way, these two variables are kept for further analysis. However, later, parameterization of Nivolog will permit a selection of situations in which wind speed is kept (when no snowdrift is recorded) or is omitted (when snowdrift index is positive during the last 24 hours).

a.4. New variables

In addition to the variables presented before, two new variables are added to the analysis. At this stage, snowfall during the last three days and snowpack compaction are the two new variables.

- **Snowfall during the last three days (cm)**

This variable is created on the basis of the variable "snowfall in 24 hours", which is summed for the last three days before the observation under consideration. No coherence tests are performed for this new variable, because the basis variable "snowfall in 24 hours" has already been checked and errors removed.
Snowfall during the last three days is an interesting variable to add to the analysis because it gives information on the cumulative precipitation which occurred. As already explained, the adding of fresh snow, even if it is often light, leads to increased traction force on the snowpack (Bolognesi R., 2015). The variable of snowfall during the last 24 hours only gives information on one day conditions (according to its definition). However, snowfall equal to 10 cm in 24 hours is not a high value for the stability of the snowpack, but three days with 10 cm of snow give a total of 30 cm, which can destabilise the snowpack. In this way, overview on past precipitation can bring further important information to the analysis. Beyond the threshold of three days, fresh snow is more likely to have been transformed by compaction or weather conditions. So, no other variables summed over more than three days are kept for this analysis.

- **Snowpack compaction in 24h (cm)**

As shown in coherence tests, the variable "variation of the snowpack thickness (cm)" contains a lot of errors, mainly for positive variations (cf. Chapter IV.a.3.1). Furthermore, this variable is redundant with the variable of snowfall in 24 hours, which is the main cause of positive snowpack variation. In this way, the decision was taken to keep only negative values, meaning that snowpack thickness decreased in 24 hours and compaction took place. So, positive values are transformed in "/" (no compaction of snowpack), and negative values are kept as they are. As for the new variable of snowfall during the last three days, coherence tests are not performed, because the variable of snowpack compaction is constructed on the basis of the variable already checked and corrected.
Snowpack compaction in 24 hours is a very important variable because it gives information on the stabilisation of snowpack. As already presented in Chapter IV.a.1.2, fresh snow is very fragile and unstable. But after some times, sun radiation, temperature, wind and other meteorological condition, as well as compaction due the weight of snow act as stabilisers for the fresh snow and the snowpack in general. In this way, the variable of snowpack compaction is important to inform about stability conditions in the snowpack.

a.5. Midway problem and last corrections

The dataset recording meteorology and snow variables was cleaned, and errors removed, as explained in the previous chapters (see chapter IV.a. sections a.1 to a.4). Once satisfying and coherent values were obtained, statistical analysis was performed on this dataset. More particularly, Linear Discriminant Analysis and Logistic Regression were applied (cf. chapter IV.b); however, they gave unsatisfactory results. So, the decision was taken to split the whole dataset in four specific snow avalanche situations, in agreement with Robert Bolognesi: situations with fresh snow, with rainfall, with warming, and with snow transport by wind (cf. chapter IV.b). Once these specific datasets were constructed, same statistical analyses were performed on them, specifically. In addition to linear discriminant analysis and logistic regression, k-nearest neighbours, p-values for classification and quadratic discriminant analysis were also performed in order to assess the rate of right classification concerning days with or without snow avalanches (cf. chapter IV.B).
After these first statistical analyses, important variables, which discriminate between days with and without snow avalanches were found, for each specific situation. In this way, the parameterization of the system NivoLog was started (cf. chapter IV.c). However, the results delivered by the analysis in NivoLog were dubious. For example, several observations recorded as "no-avalanche day" had 9 nearest neighbours over 10 with the characteristic "avalanche day", which is strongly incoherent. In order to find an explanation for these incoherent results, comparisons with a file recording only information about snow avalanches ("*DD41MTA01 – Original"*) were effectuated. At this step, unsuspected errors appeared.

Unsuspected errors

Coherence tests and wrong values cleaning were performed on the dataset recording meteorological conditions and state of the snow, called "*DC41MTA01 – Original*" (cf. chapter IV.a.6). In this file, coherence tests for the occurrence of snow avalanches were performed thanks to information contained in the same file (the total number of snow avalanches in one day must be equal to the addition of accidental, artificial and natural snow avalanches for the day in consideration). Another file called "*DD41MTA01 – Original*" (cf. chapter IV.a.6) only contains information about snow avalanches characteristics (length, height of the rupture, length of the rupture, energy …). Unsuspected errors were that snow avalanches recorded in the file "*DD41MTA01 – Original*" (snow avalanche characteristics) are not systematically recorded in the file "*DC41MTA01 – Original*" (meteorological and snow characteristics). In this way, some observations appearing as "day without snow avalanche" in all the previous analyses (linear discriminant analysis, logistic regression, k-nearest neighbours, p-values for classification, quadratic discriminant analysis) were in reality "day with snow avalanches" in the file "*DD41MTA01 – Original".*

One example: 21st February 1999

In the file "*DC41MTA01 – Original*", this observation appears as "day without snow avalanches"; in other words, the variable <AVAL> takes a value of 0. However, snowfall in 24 hours is equal to 30 cm and snowfall in 3 days is equal to 61 cm, which seems dubious for a day without snow avalanches. When looking at the file "*DD41MTA01 – Original",* concerning characteristics of snow avalanches, one discovers that on the 21st of February, two snow avalanches occurred, at the sites 205 and 301, with both a height of rupture of 20 cm. Unfortunately, this case is not isolated, and new cleaning tests must be done in order to suppress all mistakes.

New cleaning tests

The goal of these lasts tests is to verify that snow avalanches recorded in the file "*DD41MTA01 – Original"* also appear in the meteorological and snow state file "*DC41MTA01 – Original".* Over a total of 385 days with snow avalanches recorded in "*DD41MTA01 – Original",* the results of the tests show that 236 cases are correctly classified as snow avalanches days in both files, 48 are observations of snow avalanches made out of the period of time recorded in "*DC41MTA01 – Original"* (so they can be left out for the present comparison) and 101 observations were recorded as snow avalanches in "*DD41MTA01 – Original",* but do not appear in "*DC41MTA01 – Original"*. Given that characteristics of snow avalanches are recorded in "*DD41MTA01 – Original"*, dates corresponding to these latter should appear as date with snow avalanche occurrence in "*DC41MTA01 – Original".* In this way, all observations in "*DC41MTA01 – Original"* with no occurrence of snow avalanche but with characteristics of at least one snow avalanche in "*DD41MTA01 – Original"* are corrected in snow avalanche days. After this last correction, the cleaned file is called "*DC41MTA01 - Cleaning A.5*" and contains 2267 observations. For the following analysis, "DC41MTA01 - Cleaning A.5" will be the basis file, from which it will be possible to derive subsample for different situations (cf. chapter IV.b.2).

As already mentioned, the problem encountered during this work really emphasizes the importance of testing the coherence and identifying errors in the basis file for each statistical analysis. It is the only way to improve the validity and the accuracy of results.

a.6. Files

For practical work, it is important that all different files used remain clear. In this way, a summary is presented in the table 5, below:

Name	Description	Number of observations
DC41MTA01 - Original	Original file about meteorological and snow variables (Wind, temperature, snow density, etc.)	2502
DD41MTA01 - Original	Original file about snow avalanches only (Length, width, height of fracture, etc.)	2516
DC41MTA01 - Cleaning A.3	File with errors cleaned in section A.3	2267
DC41MTA01 - Cleaning A.5	File with errors cleaned in section A.3 and A.5 (Correction of incoherencies in snow avalanches events between "DC41MTA01 – Original" and "DD41MTA01 – Original".	2267
DC41MTA01 - FreshSnow	File comprising only situations with fresh snow in 24 hours (conditions: HN24>0, ID<50, R24=0)	457
DC41MTA01 – FreshSnow-NivoLog	File composed of situations with fresh snow for the set of parameters "fresh snow" in NivoLog parameterization (conditions: HN24≥20cm, Hn3J≥35cm, RS=0)This file is only made up of observations with no missing values for the variables weighted in NivoLog.	191
DC41MTA01 – Rain	File comprising only situations with rain in 24 hours (conditions: R24>0)	36**!**
DC41MTA01 - Warming	File comprising only situations with milder temperatures or thaw (conditions: R24=0, ID=0, MVS≥350, Hn24<10)	108
DC41MTA01 – Warming - NivoLog	File comprising only observations with no missing values for the variables weighted in NivoLog.	71
DC41MTA01 - SnowTransport	File comprising only situations with snow transport in 24 hours (conditions: ID≥50, Hn24<30)	288
DC41MTA01 – SnowTransport-NivoLog	File comprising only observations with no missing values for the variables weighted in NivoLog.	199
DC41MTA01 – Atypical - NivoLog	File made up of observations belonging to none of the four typical snow avalanche situations (fresh snow, snow transport, warming, and rainfall). File comprising only observations with no missing values for the variables weighted in NivoLog.	777

Table 5: Summary of all different files used in the present work.

b. Determination of important variables – Results

Cleaning and coherence tests of the previous chapter allowed the correction of numerous errors. In this part of the work, statistical methods presented in the chapter "III. STATISTICAL METHODS" are applied to the cleaned database to find important variables in triggering snow avalanches for the ski resort of Aminona. In a first chapter, these statistical analyses are applied indifferently on the whole dataset. In a second chapter, different snow avalanche situations are selected in order to improve the results of the first analyses.

b.1. Interpretation of the results

For all analyses below, the presentation of the results is established as follows:

1. First, results of LDA are presented in the form of a table. The first column indicates which variables are taken into consideration for the analysis (for abbreviations, see chapter IV.a). The second column presents the contingency table with real membership of observations to classes 0 or 1, and estimated membership by the statistical analysis (also 0 or 1). Numbers appearing in bold in the table are right classifications, and numbers in italic are wrong classifications. The third column shows the percentage of right classification with cross validation, for the variables used in the analysis. It is calculated by adding the right classifications (in bold), dividing by all observations (right and wrong classifications), and multiplying by 100 to obtain percentages. The fourth column indicates which variables are important in discriminating between days with snow avalanches and days without snow avalanches. These important variables are boldface when they are greater than 5 in absolute value, and considered as important with this criterion. In following rows, other LDA are performed with variables appearing important during the previous analysis, trying to keep the RCR high. The last row shows the best LDA performed on most important variables with the highest RCR.
2. Secondly, results of LR are presented in the same way as the results of LDA. The only difference is the fourth column. In the case of LR, only important variables are showed, and the corresponding signs indicate the confidence level for their importance in discriminating between days with and without snow avalanches (cf. Chapter III "STATISTICAL METHODS"). As for LDA, the best LR with most important variables and highest RCR is presented in the last row of the table.
3. Thirdly, after these first two statistical analyses, discussion on results and important variables takes place.
4. Fourthly, results of the method of p-values for classification are presented. This method is always performed with all variables unless one of them appeared as collinear with another one or one of the variables is always equal to zero. It allows classifying observations with a high level of confidence, even in difficult cases, and so, is performed with all variables. Three contingency tables are presented, for a classification rate with a confidence level of 95%, 90% and 80% confidence, respectively. The rows indicate the real membership to classes (0 or 1). The columns indicate the estimated membership to classes according to p-values of each observation. The first column indicates the membership to none of the two classes, the second one to the first class (0), the third one to the second class (1), and the fourth one to both of the two classes. For a better interpretation of the results, totally right classifications appear in bold, totally wrong ones appear in italic and classifications, which belong to both or none classes appear neutral. If the numbers in the contingency tables are multiplied by 100, they indicate the percentage of observations wrongly/rightly classified with respect to the class they belong to[1].

[1] NB: Right classifications for class [0] and class [1] cannot be summed because they relate to each class, respectively.

5. Fifthly, results of the method of kNN analysis are presented. As for p-values for classification method, kNN is always performed with all variables unless one of them appeared to be collinear with another one or one variable is always equal to zero. Three contingency tables are presented, for analysis with one (k=1), three (k=3) and five (k=5) nearest neighbours, respectively. Rows indicate the real membership of observations (class 0 or 1), and columns indicate the estimated membership of observations by kNN analysis. Numbers in the table correspond to the number of observations classified according their real membership and their estimated membership. The RCR for kNN analysis is calculated by adding the right classifications, dividing them by the total of observations (right classifications in bold), and multiply by 100 to obtain percentages.
6. Sixthly, a second discussion on RCR of the different methods presented before takes place.

b.2. Analysis on the whole dataset

All cleaning and coherence tests of the previous chapter led to a clean basis file called “DC41MTA01 - Cleaning A.5” and made up of 2267 observations. For statistical analysis, only complete values for each variable and each observation can be taken into account. If a missing value remains present in the dataset, LDA and LR are not able to be performed. In this way, all incomplete observations are removed, which leads to a database of only 1027 complete observations.

Linear discriminant analysis

Variables considered for LDA	Table	Right classification(with cross validation)	Important variables (greater than ¦5¦)
All variables	estimated 0 1 observed 0 578 221 1 37 191	74.88%	RS -0.2811873 N -0.7272642 **MVS 23.1428982** Tn 3.6445342 Ps 0.3022454 **Ta 20.4042556** R24 0.4689347 **ADP 6.6598234** ADN20 0.8952742 **dTa -10.2356343** tHs24 -0.1696056 vV 0.6673565 **ID 22.1960769** **HN24 13.5040496** **Hn3J 42.4178457**
MVS, Ta, ADP, dTa, ID, HN24, Hn3J	estimated 0 1 observed 0 578 227 1 37 185	74.29%	**MVS 20.952969** **Ta 25.629291** **ADP 5.688557** **dTa -10.273283** **ID 21.052254** **HN24 13.399518** **Hn3J 42.731040**
MVS, Ta, dTa, ID, HN24, Hn3J	estimated 0 1 observed 0 580 228 1 35 184	74.39%	**MVS 21.613382** **Ta 23.896973** **dTa -9.588883** **ID 22.203583** **HN24 13.392686** **Hn3J 41.799696**

Logistic regression

Variables considered for LR	Table	Right classification (with cross validation)	Important variables
All variables	estimated 0 1 observed 0 552 63 1 190 222	75.36%	N ** MVS . Ta ** R24 * ADP * dTa . ID * HN24 *** Hn3J ***
N, MVS, Ta, R24, ADP, ID, HN24, Hn3J	estimated 0 1 observed 0 558 57 1 190 222	75.95%	N ** MVS . Ta ** R24 * ADP * ID * HN24 *** Hn3J ***

Discussion on important variables found by LDA / LR

LDA and LR indicate that five variables are very important in discriminating between days with snow avalanches and days without snow avalanches for both analyses. These common variables are:

- snow density (MVS)
- air temperature (Ta)
- snowdrift index (ID)
- snowfall in 24 hours (HN24)
- snowfall during the last three days (Hn3J).

MVS is important for both analyses, which seems coherent with results found in the literature (Pahaut E.,Bolognesi R., 2003; Floyer J.A., 2003). Indeed, MVS can refer to very different snow avalanche regimes, or stability states. If MVS is very low (50 – 120 kg/m^3), it indicates that snow is recent and cold, because transformation of grains have not taken place yet, the air content is high, cohesion between the crystals of snow is low, and snow avalanches can occur due to the lack of stability in the snowpack. At the same time, very dense snow (400 – 600 kg/m^3) can either indicate a very stable snowpack, with high cohesion between the grains, or a very humid snowpack with high content of liquid water, which could also lead to snow avalanches (Ancey C., Sergent C., Martin E., 2003). In this way, it seems coherent that MVS is an important variable in discriminating between days with or without snow avalanches.

Finding Ta as an important variable also seems coherent with literature and studies, in which it is widely used (Boyne H.S., Williams K., 1992; Singh A., Srinivasan K., Ganju A., 2005; Pahaut E.,Bolognesi R., 2003). The importance of Ta can be understood by its indirect role played on the transformation of snow crystals. Even if snowpack is relatively isolated from atmosphere, Ta has a direct influence on snow temperature (Tn) in the upper layer. In this way, if Ta is cold, transformations of snow crystals take place slowly, compared to warmer temperature with corresponding faster transformations. (Ancey C., Sergent C., Martin E., 2003). If light snow crystals

transform rapidly, cohesion between them increases at the same time as stability. However, if Ta is too warm snow melts, liquid water content increases, and destabilizes the snowpack. Indeed, Ta is a variable playing a role at a certain threshold: if it is lower than zero, snow avalanches may or may not occur. However, if it is greater than zero, snow avalanches are more likely to occur. In this way, it seems reasonable to find that Ta is an important variable.

Snowdrift Index also appears as an important variable in many other studies (Bolognesi R., 2015; Bellot H., Bouvet, F.N., 2010; Pahaut E., Bolognesi R., 2003). This variable is more important that wind because this directly measures the transport of snow, without the use of modelling. ID is important in discriminating days with and without snow avalanches. The reason for this is that transported snow leads to high and heavy accumulations, increases traction forces in the snowpack and destabilizes it (Bolognesi R., 2015). Furthermore, cohesion between grains of snow transported by wind is weak, which increases the instability (Bolognesi R., 2015). So, it seems that the occurrence of snow avalanches and snowdrift events are in relationship, and the result of LDA and LR for ID seems coherent.

HN24 and Hn3J both refer to fresh snow added to the snowpack, which appears to be one of the causes for snow avalanches to occur in various studies (Fromm R., 2009, Jomelli V. et al., 2007, Floyer J.A. and McClung M.D., 2003, Saemundsson T., et al., 2003, Bolognesi R., 2013). Indeed, addition of fresh snow increases traction forces due to the weight of snow. But more important for the destabilization of the whole snowpack is the adding of a new layer of snow, which is not always in perfect cohesion with the lower layers (Bolognesi R., 2013). In this way, HN24, Hn3J and the occurrence of snow avalanches also seem in close relationship, and thus, these two variables are important to predict snow avalanches occurrence or not.

To summarize, the results of LDA and LR seem coherent with literature and knowledge in the domain of snow avalanches, as explained above. In this way, these five common important variables derived from LDA and LR will be used in priority for NivoLog parameterization in chapter IV.c. In other words, the weighting will be applied preferentially on these variables, compared to other ones which have not appeared as statistically significant for the discrimination between days with and without snow avalanches.

p-values for classification

With a confidence level of 95% :

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.23414634**	*0.04878049*	0.7170732
1	0	*0.04854369*	**0.41990291**	0.5315534

With a level of 90% confidence :

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.34146341**	*0.09918699*	0.5593496
1	0	*0.09951456*	**0.53883495**	0.3616505

With a level of 80% confidence :

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.5902439**	*0.2000000*	0.2097561
1	0	*0.1990291*	**0.6407767**	0.1601942

kNN

For kNN with k=1:

```
           estimated
              0   1
observed  0 467 148
          1 175 237

RCR: 68.55%
```

For kNN with k=3:

```
           estimated
              0   1
observed  0 522  93
          1 197 215

RCR: 71.76%
```

For kNN with k=5:

```
           estimated
              0   1
observed  0 536  79
          1 222 190

RCR: 70.69%
```

Discussion on different RCR

p-values for classification and kNN are two methods used to assess the classification rate of the observations belonging to the whole dataset, in addition to LDA and LR. The method of p-values for classification has the advantage that even in difficult classification cases, observations can be classified with a certain confidence level. In the case of the whole dataset, all observations can be classified in one class or the other, or in both classes at the same time (no observations belongs to neither of the two classes). However, the differentiation between class [0] and [1] is not optimal, because some observations can be classified in both classes. However, even if classification for the whole dataset is intrinsically difficult, (as indicated by LDA and LR) it is possible, with a confidence level of 95%, to correctly classify 41.99% of snow avalanche days as “snow avalanche days” and 23.41% of no-snow avalanche days as “no-snow avalanche days”. In another hand, the method of kNN permits to adjust the analysis by selecting various numbers of neighbours. With k=3, a RCR of 71.76% is reached. The results of kNN, LDA and LR are quite different, but remain close to each other. LDA and LR give better RCR than kNN (74.39% and 75.95% respectively), with the most important variables taken into consideration.

To summarize, RCR for the whole dataset reach: 74.39% with LDA, 75.95% with LR, and 71.76% with kNN (k=1). Furthermore, when taking a confidence level of 95%, 41.99% of days with snow avalanches are correctly classified and 23.41% of days without snow avalanches too. Even if these results are satisfactory, the whole dataset considered in these previous analyses is, in reality, composed of different typical situations of snow avalanches added to situations more common, with no relevant event. So, the improvement proposed in the present work is to split the whole dataset into four sub-datasets corresponding to well-known snow avalanches situations, and one sub-dataset composed of common observations without relevant event.

b.3. Analysis of typical situations of snow avalanches

b.3.1. Presentation of the datasets

To improve the RCR of the first analysis and find more accurate results concerning significant variables, it was decided to divide the whole dataset into typical meteorological situations potentially leading to snow avalanches. This partition is based on typical snow avalanches situations found in the book of Robert Bolognesi "Estimer et limiter le risque avalanche" (2013). At the page 39 of this book, four meteorological situations potentially triggering snow avalanches are presented: input of snow by fresh snowfall, input of snow by wind transport, input of liquid water by rainfall and input of liquid water by snowpack melting (Bolognesi R., 2013). In this way, the previous dataset ("*DC41MTA01 - Cleaning A.5*") is split into four groups according to the criteria found in the book of Robert Bolognesi.

- Fresh snow

Fresh snow leads to typical situations of snow avalanches occurrence. Indeed, adding fresh snow on the already existing snowpack increases traction forces, and destabilize the whole snowpack by adding new layers of snow, which do not always have a good cohesion with the underlying layers. So, the important point for the selection of fresh snow situations is that snowfall during the last 24 hours needs to be superior to zero, and, at the same time, snow transport by wind is weak (to put aside snow avalanches situations due to snow transport) and no rain occurred during the last 24 hours of the observation (to put aside snow avalanches situations due to rainfall). In this way, the dataset representing only situations with fresh snow is constructed according the following criteria: HN24 > 0, ID < 50 kg/m^3 and R24 =0. This leads to a dataset called "*DC41MTA01 – FreshSnow*", made up of 457 observations.

- Snow transport by wind

Transport of snow by wind leads to the same type of destabilization in the snowpack as fresh snowfall presented above. Indeed, snow is added to the existing snowpack by the wind instead of being added by solid precipitations. However, transport of snow by wind can lead to greater and very irregular accumulations, compared to simple snowfall without wind. So, traction forces can be highly increased in certain places compared to others. Furthermore, grains of snow transported by wind have a lower cohesion than snow crystals deposited without wind and so, accumulations of snow due to wind transport increase the instability of the snowpack. Snowdrift has been presented as a highly relevant variable to predict the occurrence of snow avalanches by various studies (Bolognesi R., 2015; Bellot H., Bouvet, F.N., 2010; Pahaut E., Bolognesi R., 2003). So, the dataset corresponding to snow transport situations is selected according to the following criteria: ID $\geq$ 50 kg/m^3 and HN24 < 30 cm (to avoid taking into account situations with high amount of fresh snow). It is called "*DC41MTA01 – SnowTransport*" and is made up of 288 observations.

- Rainfall

Rainfall situations often lead to snow avalanches occurrence. The reason for this is that liquid water added to the snowpack increases its density and the related traction forces. In other words, if the traction forces acting on the snowpack are increased (due to the more elevated density of rain in the snowpack) and become stronger than resistance forces, snow avalanches occur. Furthermore, high liquid water content (about 12%) weakens the cohesion between snow grains (Ancey C., Sergent C., Martin E., 2003). So, snow avalanches occurrence also seems in close relationship with heavy rainfall.

The dataset for rainfall situations is constructed by selecting observations with the variable “rainfall in 24 hours” greater than zero. In other words, only observations for which rain occurred in the last 24 hours are kept for this dataset. It is called “*DC41MTA01 – Rain*” and is made up of only 36 observations.

- Warming situations

As for rainfall, the principal trigger of snow avalanches due to warming is the addition of liquid water in the snowpack. The consequences are the same as for rainfall situations, the only difference being that liquid water come from melting of snow. This dataset is constructed on the basis of different criteria: ID=0 (to avoid taking in consideration snow avalanche situations due to transport of snow), R24=0 (to put aside situations of snow avalanches due to rainfall), HN24 < 10 cm (to put aside situations of snow avalanches due to snowfall), MVS $\geq$ 350 kg/m^3 (to select situations with elevated densities). These criteria relate to melted snow, very dense due to high liquid water content. So, the dataset concerning warming situations is called “*DC41MTA01 - Warming*” and is made up of 108 observations.

b.3.2. Fresh snow situations

Separation of the whole dataset led to different sub-datasets in the previous chapter. In this part of the work, statistical analyses are performed on the dataset called "*DC41MTA01 – FreshSnow*", made up of 457 observations. However, as already said for the whole dataset, statistical analyses can only be performed if each observation has complete values for each variable. In this way, all observations with at least one missing value for one variable are left out of the analysis. This leads to a dataset made up of only 248 complete observations.

1. Linear discriminant analysis

Variables considered for LDA	Table	Right classification (with cross validation)	Important variables (greater than ¦5¦)
All variables except R24 (equal to zero by definition)	estimated 0 1 observed 0 **83** *41* 1 *29* **95**	71.77%	RS -0.8829016 N -1.9188503 **MVS 70.3133653** **Tn -11.7076693** **Ps 5.6148027** **Ta 10.0404541** ADP -3.2015581 **ADN20 11.0921794** dTa -3.7644611 tHs24 -0.2776319 vV -0.7323918 **ID 7.1736359** **HN24 12.2493435** **Hn3J 23.1844074**
MVS, Tn, Ps, Ta, ADN20, ID, HN24, Hn3J	estimated 0 1 observed 0 **81** *38* 1 *31* **98**	72.18%	**MVS 63.315917** **Tn -13.821730** **Ps 6.304368** Ta 4.579433 **ADN20 10.027642** **ID 7.056113** **HN24 12.346227** **Hn3J 22.878546**
MVS, Tn, Ps, ADN20, ID, HN24, Hn3J	estimated 0 1 observed 0 **82** *38* 1 *30* **98**	72.58%	**MVS 64.093754** **Tn -11.690856** **Ps 6.276304** **ADN20 10.126530** **ID 7.135505** **HN24 12.324330** **Hn3J 23.073207**

For this statistical analysis, R24 needs to be removed because it is always equal to zero, by definition of the dataset "*DC41MTA01 – FreshSnow*".

2. Logistic regression

Variables considered for LDA	Table	Right classification (with cross validation)	Important variables
All variables except dV (collinear with dD) and R24 (equal to zero by definition)	estimated 0 1 observed 0 **83** *29* 1 *34* **102**	74.60%	RS . N * MVS ** Tn . ADN20 ** ID * HN24 *** Hn3J ***
RS, N, MVS, Tn, ADN20, HN24, Hn3J	estimated 0 1 observed 0 **83** *29* 1 *34* **102**	74.60%	RS . N * MVS ** ADN20 ** ID * HN24 *** Hn3J ***

3. Discussion on important variables found by LDA / LR

LDA and LR indicate that five variables are very important in discriminating days with and without snow avalanches for both analyses. These common variables are:

- MVS (density of snow)
- ID (snowdrift index)
- ADN20 (number of days since the last snowfall exceeding 20 cm)
- HN24 (snowfall in 24 hours)
- Hn3J (snowfall during the last 3 days).

For fresh snow situations, MVS appears as a very important variable in discriminating between days with or without snow avalanches, as for other studies in the domain of snow avalanches (Pahaut E.,Bolognesi R., 2003; Floyer J.A., 2003). This directly relate to the type of snow: low densities (50 – 150 kg/m^3) indicate very light and cold snow, with a high air content and many spaces between snow crystals, while higher densities (200 – 350 kg/m^3) relate to heavier and more temperate snow, with a lower air content. As already explained, for light snow, cohesion between snow crystals are generally weak, which creates instabilities in the layer. For more dense snow, cohesion is generally stronger because less air is present, snow grains are better bonded with each other, leading generally to a more stable layer. Thus, finding MVS as an important variable for discriminating snow avalanche situations seems coherent.

Snowdrift index is an important variable too (like in other studies and books: Bolognesi R., 2015; Bellot H., Bouvet F.N., 2010; Pahaut E., Bolognesi R., 2003). Even if only fresh snow situations with a low snowdrift are taken into consideration, this latter variable still remain important. This can be explained by close relationship between the arrival of fronts and snowfall. Often, fronts are accompanied by winds, which transport falling snow. So, it seems coherent that snowdrift index is also an important discriminating variables for fresh snow situations, due to its link with the arrival of fronts.

ADN20 give information on conditions during the days preceding the observation in consideration. If this variable is low, important quantities of fresh snow fell during the days before, stabilization due to snow crystal transformations has not have a long time to take place, and, generally, instabilities are present in the snowpack. On the contrary, if ADN20 is high, weak snowfall can have occurred, but the last important snow fall has have time to stabilize, at least partially. So, with finding ADN20

important for discriminating days with and without snow avalanches, one can point out that knowledge about days before the observation seems important.
Both HN24 and Hn3J are point out as important variables by LDA and LR, as for many other studies (Fromm R., 2009, Jomelli V. et al., 2007, Floyer J.A. and McClung M.D., 2003, Saemundsson T., et al., 2003, Bolognesi R., 2013). This means that quantities of snow fell in 24 hours or 72 hours matter in discrimination days with and without snow avalanches, because they are providers of snow for snow avalanches to occur.

4. p-values for classification

With a confidence level of 95%:

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.34821429**	*0.04464286*	0.6071429
1	0	*0.04411765*	**0.35294118**	0.6029412

With a confidence level of 90%:

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.54464286**	*0.09821429*	0.3571429
1	0	*0.09558824*	**0.44117647**	0.4632353

With a confidence level of 80%:

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.6607143**	*0.1964286*	0.1428571
1	0	*0.1985294*	**0.6102941**	0.1911765

5. kNN

For kNN with k=1:

	estimated	
	0	1
observed 0	**73**	*39*
1	*48*	**88**

RCR: **64.92%**

For kNN with k=3:

	estimated	
	0	1
observed 0	**77**	*35*
1	*55*	**81**

RCR: **63.71%**

For kNN with k=5:

	estimated	
	0	1
observed 0	**74**	*38*
1	*49*	**87**

RCR: **63.71%**

6. Discussion on different RCR

p-values for classification and kNN are two methods used to assess the classification rate of the observations belonging to the dataset for fresh snow situations, in addition to LDA and LR. The method of p-values for classification has the advantage to classify observations with a certain confidence level even in intrinsically difficult cases. In the case of fresh snow situations, all observations can be classified in one class or the other, or in both classes at the same time (no observations belongs to neither of the two classes). However, the differentiation between class [0] and [1] is not optimal, because some observations can be classified in both classes at the same time. However, with a confidence level of 95%, it is possible to correctly classify 35.29% of snow avalanche days as “snow avalanche days” and 34.82% of no snow avalanche days as “no snow avalanche days”. The wrong classifications are both close to 5%, due to the confidence level set to 95%. Concerning the method of kNN, the best classification rate reaches 64.92% when taking k=1.

Contingency tables of LDA/LR with cross validation give better results: 72.58% and 74.60%, respectively, with only important variables. These results are not much better than results for considering the whole dataset. However, it permits to select specific important variables triggering snow avalanches for fresh snow situations.

b.3.3. Snow transport situations

For snow transport situations, "*DC41MTA01 – SnowTransport*" is made up of 288 observations. However, as for all previous analyses, only complete observations for each variable are taken into consideration by statistical analyses. So, the removal of missing values leads to only 92 complete observations.

1. Linear discriminant analysis

Variables considered for LDA	Table	Right classification (with cross validation)	Important variables (greater than ¦5¦)
All variables	estimated 0 1 observed 0 **7** *12* 1 *13* **60**	72.83%	RS -0.5033549 N -2.5322696 **MVS 53.1833811** **Tn 25.5628527** **Ps 12.1314634** **Ta -13.3309560** R24 2.0710023 ADP 2.1650450 **ADN20 -5.3497477** **dTa 23.9860965** tHs24 -2.7795091 vV 0.4175058 **ID 25.4360422** **HN24 7.2167792** **Hn3J 27.6987527**
MVS, Tn, Ps, Ta, ADN20, dTa, ID, HN24, Hn3J	estimated 0 1 observed 0 **5** *7* 1 *15* **65**	76.01%	**MVS 31.672376** **Tn 30.955592** **Ps 12.121218** **Ta -14.234460** ADN20 -4.550141 **dTa 14.189534** **ID 12.580979** HN24 2.865344 **Hn3J 43.859696**
MVS, Tn, Ps, Ta, dTa,ID, Hn3J	estimated 0 1 observed 0 **5** *3* 1 *15* **69**	80.43%	**MVS 39.98422** **Tn 34.64295** **Ps 14.33090** **Ta -13.73775** **dTa 12.74073** **ID 38.03044** **Hn3J 58.57711**
MVS, Tn, Ps,ID, Hn3J	estimated 0 1 observed 0 **5** *3* 1 *15* **69**	81.52%	**MVS 39.24356** **Tn 30.07433** **Ps 13.81109** **ID 38.39146** **Hn3J 59.84065**

2. Logistic regression

Variables considered for LDA	Table	Right classification (with cross validation)	Important variables
All variables	estimated 0 1 observed 0 **10** *10* 1 *10* **62**	78.26%	No significant variable!
MVS, Tn, Ps, ID, Hn3J	estimated 0 1 observed 0 **9** *11* 1 *6* **66**	81.52%	Hn3J ***

3. Discussion on important variables found by LDA / LR

Analysis on snow transport situations are performed with 15 explanatory variables on 92 observations. According to the rule $N_{min} \geq (p+1)*5$ (see chapter III "STATISTICAL METHODS"), $92 \geq 80$, which is close to the limit of the number of variables allowed with respect to sample size. So, the results of this part need to be considered cautiously. This can maybe also explain the difficulty to find important variables with LR. As no significant variables were emphasized by LR, the best discriminating variables from LDA are taken to perform LR. This combination of variables gives the highest RCR.

For both analyses, only Hn3J (snowfall during the last 3 days) seems to be a common important variable. However, MVS (snow density), Tn (snow temperature), Ps (probe penetration), ID (snowdrift index) also appear as important variables for LDA, and they also give the best RCR for LR. In this way, MVS, Tn, Ps, ID and Hn3J are considered as common important variables for both analyses, even if, for LR, they do not appear significant. So, we consider that LDA and LR lead to the identification of five common important variables:

- MVS (snow density)
- Tn (snow temperature)
- Ps (probe penetration)
- ID (snowdrift index)
- Hn3J (snowfall during the last 3 days)

Snow density is an important discriminating variable because it gives information about the structure and the type of snow. If MVS is low, it means that high air content is present in the top layer, and cohesion is not strong between the snow grains. In this way, wind can more easily mobilize snow. On the contrary, if MVS is high, cohesion between snow grains is stronger and snow is more difficult to mobilize. So, it seems coherent that the discrimination between days with and without snow avalanches can be partly done by MVS.

Snow temperature is in close relationship with MVS, and gives information about the structure and type of snow too. If Tn is close to zero, snow can partially melt and cohesion between snow grains increases[2]. On the contrary, if Tn is low and MVS is low too, this can indicate a snow layer with high air content and low cohesion. So, for these latter conditions, mobilization of snow by wind is easier.

[2] NB: If the liquid water content becomes too high, cohesion between snow grains is reduced and the layer of snow becomes unstable.

So, coupled with MVS, Tn also seems coherent for the discrimination between days with and without snow avalanches.

Ps is an important variable too, because it relates, in some extent, to cohesion inside the top snow layer. If probe penetrates very deep in the snowpack, snow grains are generally not well bonded to each other, meaning that snow can be more easily mobilized by wind[3]. If the penetration of the probe is less deep, this generally means that snow grains are strongly bonded with each other and mobilization of snow by wind is more difficult. So, finding Ps as an important discriminating variable seems coherent.

ID is the third important variable emphasized by LDA and LR. Snowdrift index give information about the quantities of snow transported by wind. So, as explained in the discussion for the whole dataset, if high transport of snow takes place, this can lead to great snow accumulations at certain places, and great instabilities due to increased traction forces. Furthermore, grains of snow transported by wind generally have a lower cohesion.

Hn3J appears as an important variable to discriminate days with and without snow avalanches. The reason for this is that, for snow transport situations, snow must be available to be transported, and to be mobilized by wind, snow needs to be taken from snowpack. So, if fresh snowfall occurred during the last three days, snow is generally not very compact and stable, and snow particles can easily be taken and transported by wind, leading to great and instable snow accumulation at certain places. On the contrary, if no snowfall took place during the last 3 days preceding the observation, a lower quantity of snow can generally be mobilized. In this way, it seems coherent to find Hn3J is an important variable discriminating days with and without snow avalanches.

So, discrimination between days with and without snow avalanches when snow transport takes place partially depends on the five variables mentioned previously. This means that for NivoLog parameterization in the following chapters, MVS, Tn, Ps, ID and Hn3J will be weighted preferentially (see Chapter IV.c).

4. p-values for classification

With a confidence level of 95%:

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.05000000**	*0.0500000*	0.9000000
1	0	*0.04166667*	**0.2361111**	0.7222222

With a confidence level of 90%:

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.35000000**	*0.1000000*	0.5500000
1	0	*0.09722222*	**0.4305556**	0.4722222

With a confidence level of 80%:

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.7000000**	*0.2000000*	0.1000000
1	0	*0.1944444*	**0.5972222**	0.2083333

[3] NB : This is not the case in all situations ! With very humid and melted snow, probe can also penetrate deeply in the snowpack, but snow grains cannot be mobilized by wind.

5. kNN

For kNN with k=1:

```
          estimated
              0  1
observed  0 9  11
          1 15 57

RCR: 71.74%
```

For kNN with k=3:

```
          estimated
              0  1
observed  0 8  12
          1 10 62

RCR: 76.09%
```

For kNN with k=5:

```
          estimated
              0  1
observed  0  7 13
          1  7 65

RCR: 78.26%
```

6. Discussion on different RCR

With the method of p-values for classification, all observations can be classified in one class or the other, or in both classes at the same time (no observations belongs to neither of the two classes). However, the differentiation between class [0] and [1] is not optimal, because some observations can be classified in both classes at the same time. However, with a confidence level of 95%, one can correctly classify 23.61% of observations belonging to class [1]. However, the results for class [0] are bad, because only 5% of these observations are rightly classified. Concerning the method of kNN, the best classification rate reaches 78.26% when taking k=5.

Contingency tables of LDA/LR with cross validation give quite similar results: 81.52% for both methods, with only important variables. In the case of snow transport situations, LDA, LR and kNN give quite similar results. Furthermore with p-values for classification one can rightly classify 23.61% of snow avalanche days and only 5% of no snow avalanche days, only taking a risk of 5%.

The RCR of statistical methods applied on the dataset "*DC41MTA01 – SnowTransport*" are about 5% better than RCR for the whole dataset. In this case, separating snow transport situations from the whole dataset appears as a good decision to improve RCR. Furthermore, it allows a better identification of important variables for this typical situation.

b.3.4. Rainfall situations

In this part of the work, statistical analyses are performed on the dataset called "*DC41MTA01 – Rain*", made up of only 36 observations. This means that, over 19 years of observations, only 36 rain events were recorded at 2300 m above sea level, in the ski resort of Aminona. Furthermore, statistical analyses can only be performed if each observation has complete values for each variable. So, after removal of missing values, only 25 complete observations are kept for analysis of rainfall situations.

For this chapter, it is important to be aware that results need to be interpreted cautiously. The very small number of observations for rainfall situations only allows an overview of what could be the important variables in discriminating days with and without snow avalanches, but definitely not highly significant results. Similarly, conclusions and interpretations are dubious too, for the same reason.

1. Linear discriminant analysis

Variables considered for LDA	Table	Right classification (with cross validation)	Important variables (greater than ¦5¦)
All variables except ADP (equal to zero by definition)	estimated observed 0 3 6 1 2 14	68.00%	RS -6.83698044 N -1.00655567 MVS 299.48748997 Tn 22.52593732 Ps 0.09319514 Ta -1.77126251 R24 -0.80656300 ADN20 -27.24958231 dTa -37.48920818 tHs24 16.04964410 vV -2.58453401 ID 7.03506547 HN24 -21.04773479 Hn3J 107.45115418
RS, MVS, Tn, ADN20, dTa, tHs24, ID, HN24, Hn3J	estimated observed 0 4 3 1 1 17	84.00%	RS -6.064126 MVS 293.899307 Tn 25.274390 ADN20 -24.725542 dTa -33.666644 tHs24 15.400717 ID 11.044667 HN24 -29.023094 Hn3J 133.533170
R24, Tn, Hn3J	estimated observed 0 2 0 1 3 19	87.5%	R24 10.32298 Tn 51.13509 Hn3J 88.24020

The first two rows of the table show results that are not valid, because the number of observations is too small compared to the number of variables used. Only the last row showing the best combination of variables can be interpreted with caution.

2. Logistic regression

Variables considered for LDA	Table	Right classification (with cross validation)	Important variables
R24, MVS	estimated 0 1 observed 0 **1** *4* 1 *2* **17**	75%	R24 .
R24, Tn, Hn3J	estimated 0 1 observed 0 **4** *1* 1 *2* **17**	87.5%	No significant variable!

For LR, it is not possible to perform analyses with all variables taken into consideration. So, different runs of the analysis have been performed, and only the best combination of variables is shown here.

3. Discussion on important variables found by LDA / LR

As notified previously, results of LDA and LR need to be interpreted cautiously, due to the small number of observations. For LDA, the combination of R24 (rainfall during the last 24 hours), Tn (snow temperature) and Hn3J (snowfall during the last 3 days) seems to give the best classification rate, with each of these variables being important. For LR, the same best classification rate is found when using the same three explanatory variables. However, none of them are significant at a level of 90% and more. Being aware of the fact that these analyses are dubious, 3 important variables can be emphasized:

- R24 (rainfall during the last 24 hours)
- Tn (snow temperature)
- Hn3J (snowfall during the last 3 days)

R24 appears as important for the discrimination, because addition of liquid water in the snowpack often leads to a destabilization due to increased density and traction forces, but also due to a weakening of cohesion between the grains if the liquid water content exceeds 12% (Bolognesi, R. (2013), Ancey C., Sergent C., Martin E. (2003)). In this way the results of LDA and LR for this variable seem coherent, even if analyses are dubious due to the small number of observations.

Tn also appears as important for the discrimination, because it indirectly gives information on the state of snowpack at the moment of rainfall. If Tn is low, addition of liquid water will either stabilize snowpack by partial melting of snow crystals, or destabilize snowpack slowly, as water will infiltrate progressively deeper layers, weakening cohesion of grains due to melting. If Tn is warmer at the time of raining, cohesion is generally weaker due to a more important liquid water content, and the addition of liquid water will decrease the cohesion further. So, even if precautions need to be taken, it seems coherent that Tn is an important variable in discriminating days with and without snow avalanches.

Hn3J also appears as an important variable when considering rainfall situations. It gives information about the amount of fresh snow, which will possibly be saturated in liquid water due to rainfall. If snowfall during the last 3 days is important, this layer has not always have time to stabilize, and addition of a little liquid water could partially stabilize it. However, if high quantities of liquid water infiltrate this layer, it also can totally destabilize it. So, results about Hn3J also seem coherent, even if it is important to remain cautious with respect to interpretations.

In following parameterization of NivoLog, R24, Tn and Hn3J will be weighted to validate the system performance. However, one knows that these results cannot be totally trusted due to the small number of observations.

4. p-values for classification

With a confidence level of 95%:

```
b   P(b,{}) P(b,{0}) P(b,{1}) P(b,{0,1})
  0       0      0.2        0        0.8
  1       0      0.0        0        1.0
```

With a confidence level of 90%:

```
b   P(b,{})   P(b,{0}) P(b,{1}) P(b,{0,1})
  0       0 0.40000000        0  0.6000000
  1       0 0.05263158        0  0.9473684
```

With a confidence level of 80%:

```
b      P(b,{})    P(b,{0})  P(b,{1})  P(b,{0,1})
  0    0.20000000 0.2000000 0.0000000  0.6000000
  1    0.05263158 0.1052632 0.6315789  0.2105263
```

5. kNN

For kNN with k=1:

```
          estimated
              0  1
observed  0  1  4
          1  3 16

RCR: 70.83%
```

For kNN with k=3:

```
          estimated
              0  1
observed  0  1  4
          1  1 18

RCR : 79.17%
```

For kNN with k=5:

```
Not possible, not enough observations!
```

6. Discussion on different RCR

Because of the small number of observations, interpretations and results on RCR must be interpreted cautiously, as for the importance of variables. With the method of p-values for classification, all observations can be classified in one class or the other or in both classes at the same time, for confidence levels higher than 80%. For this latter confidence level, some estimated observations belong to none of the two classes. With a confidence level of 95%, observations of class [0] have a right classification of 20% while observations of class [1] are systematically wrongly classified. Differentiation between class [0] and [1] is not optimal, because some observations can be classified in both classes at the same time. So, the small number of observations for rainfall situations does not seem suitable for p-values method.

Concerning the method of kNN, only one to three nearest neighbours can be considered by the analysis, but it does not work anymore for five nearest neighbours. This is also due to the too small number of observations. The best RCR is found when considering 3 nearest neighbours and reaches 79.17%, but this value cannot be trusted because kNN analysis is performed on all variables, and the number of variables is too large compared to the number of observations.
RCR of LDA and LR can be interpreted more seriously because only three variables are used to perform analysis. However, 25 observations constitute a very small sample to perform statistical analyses, so the results are dubious anyway. LDA and LR both give the same RCR of 87.5%.

To summarize, all results presented here only give an overview about which variables seem important in the discrimination between days with and without snow avalanches, and about the RCR of different methods. However, the too small number of observations strongly restrains the statistical significance of the results. One can only see that the separation of the rain dataset from the whole previous dataset gives better RCR and allows the determination of three important variables for the discrimination of days with and without snow avalanches.

b.3.5. Warming situations

For these typical situations, statistical analyses are performed on the dataset called "*DC41MTA01 – Warming*", made up of 108 observations. As for previous analyses, LDA, LR, p-values for classification and kNN can only be performed on complete observations. So, after having removed observations with at least one missing value for one variable, only 69 observations are kept.

According to the rule concerning sample size and number of variables, a dataset of 69 observations should be analysed using a maximum of 11 variables. So, results of analyses performed with more than 11 variables need to be interpreted cautiously.

1. Linear discriminant analysis

Variables considered for LDA	Table	Right classification (with cross validation)	Important variables (greater than ¦5¦)
All variables except R24, ID (both equal to 0)	estimated 0 1 observed 0 **43** *13* 1 *6* **7**	72.46%	RS -0.38120604 N 0.19449036 **MVS -24.82665797** **Tn 56.71671952** Ps -2.44640780 **Ta 52.80370375** **ADP 21.67781733** **ADN20 -21.16599575** dTa -1.28959136 tHs24 1.32205017 vV -0.89053554 HN24 0.07557992 Hn3J -3.62772432
MVS, Tn, Ta, ADP, ADN20	estimated 0 1 observed 0 **44** *11* 1 *5* **9**	76.81%	**MVS -26.69987** **Tn 46.92179** **Ta 48.61779** **ADP 22.57196** **ADN20 -17.14202**

LDA is performed on all variables except R24 and ID, which are equal to zero by definition for warming situations.

2. Logistic regression

Variables considered for LDA	Table	Right classification (with cross validation)	Important variables
All variables R24, ID (both equal to 0)	estimated 0 1 observed 0 **43** *6* 1 *12* **8**	73.91%	Tn * Ta . ADN20 * Hn3J .
Tn, Ta, ADN20, Hn3J	estimated 0 1 observed 0 **43** *6* 1 *12* **8**	73.91%	Tn * Ta *
Tn, Ta	estimated 0 1 observed 0 **45** *4* 1 *12* **18**	76.81%	Tn . Ta .

3. Discussion on important variables found by LDA / LR

Results of analyses performed with all variables except R24 and ID cannot be totally trusted because variables are too numerous compared with the number of observations. Analyses performed on 11 variables at a maximum are more valid, according to the rule $N_{min} \geq (p+1)*5$, presented in the chapter III. "STATISTICAL METHODS". LDA and LR both indicate that Tn (snow temperature) and Ta (air temperature) are common statistically significant variables for the discrimination between days with and without snow avalanches. However, for LDA, ADN20 (number of days since the last snowfall exceeding 20 cm) and Hn3J (snowfall during the last 3 days) also seem important because they give the best RCR. In this way, four important variables are considered:

- Tn (snow temperature)
- Ta (air temperature)
- ADN20 (age since the last snowfall exceeding 20 cm)
- Hn3J (snowfall during the last 3 days)

For warming situations, it seems coherent to find that Tn is important in discriminating days with and without snow avalanches. For these particular situations, the more common type of snow avalanches are wet snow avalanches due to partially melted snow at a temperate temperature. In these cases, Tn is a good indicator of stability: if Tn is low during warming situations, melted snow can have refreeze, and so, cohesion between snow grains is strong. On the contrary, if Tn is close to zero, snow is partially melted with liquid water in the snowpack, and thus, cohesion is reduced compared to a lower snow temperature. So it seems coherent that Tn is an important discriminating variable.

But Tn is in relationship with Ta, and snow avalanches can also occur due to a sudden rise of temperature. Ta gives appropriate information because if it is not too warm, melting of snow is reduced, and wet snow avalanches are less likely to occur. On the contrary, if Ta is warm, melting is increased, leading to more possible occurrence of wet snow avalanches. So, finding Ta as a good discriminant variable seems coherent. However, many other factors must be taken into account for the prediction of wet snow avalanches.

ADN20 also seems coherent to discriminate days with and without snow avalanches, because it refers to the past conditions of snow accumulation. As for other situations, a fresh and important

(exceeding 20 cm) layer of snow is generally less stable than old snow, because it has not have time to stabilize. In this way, if ADN20 is low, fresh snowfall took place during the previous days, and this layer is more susceptible to slide and transform in snow avalanche. If ADN20 has high values, the last important snowfall has already have time to stabilize under different meteorological situations, and it is generally more stable. However, very different situations for this variable can lead to days with or without snow avalanches, and it is important to be aware that ADN20 is only one variable among others for discrimination of snow avalanches days.

Finding Hn3J as an important discriminating variable also seems coherent, because, as already mentioned, it refers to past accumulation of snow. In other words, if Hn3J takes high values, recent snow is more likely to be unstable, and to lead to snow avalanches occurrence. This is further amplified if a strong warming takes place and high liquid water content is created in the recent layer of snow.

4. p-values for classification

With a confidence level of 95%:

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.04081633**	*0.04081633*	0.9183673
1	0	*0.05000000*	**0.10000000**	0.8500000

With a confidence level of 90%:

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.1632653**	*0.08163265*	0.755102
1	0	*0.1000000*	**0.25000000**	0.650000

With a confidence level of 80%:

b	P(b,{})	P(b,{0})	P(b,{1})	P(b,{0,1})
0	0	**0.3469388**	*0.1836735*	0.4693878
1	0	*0.2000000*	**0.6500000**	0.1500000

5. kNN

For kNN with k=1:

		estimated 0	estimated 1
observed	0	**36**	*13*
	1	*9*	**11**

RCR: **68.12%**

For kNN with k=3:

		estimated 0	estimated 1
observed	0	41	*8*
	1	*9*	**11**

RCR: **75.36 %**

For kNN with k=5:

```
          estimated
              0  1
observed  0 43  6
          1 14  6

RCR: 71.02%
```

6. Discussion on different RCR

The first thing to note in the results of p-values for classification and kNN analyses is that they are performed with all variables except R24 and ID, and so, the results cannot be totally trusted due to too numerous number of variables.
For the method of p-values for classification, all observations can be classified in one class or the other, or in both classes at the same time (no observations belongs to neither of the two classes). However, the differentiation between class [0] and [1] is not optimal, because some observations can be classified in both classes at the same time. With a confidence level of 95% only 10% of observations belonging to class [1] are rightly classified in class [1], and 4.08% of observations belonging to class [0] are rightly classified in class [0]. The wrong classifications are both close to 5%, because the confidence level is set to 95%.
For the method of kNN, best classification rates reached 75.36% when using three nearest neighbours. However, as notified before, these results need to be interpreted cautiously. For LDA and LR, same RCR of 76.81% is found, when using different variables. RCR of LDA, LR and kNN are quite similar. However, when taking a risk of only 5%, p-values for classification only rightly classify 10% of snow avalanche days and 4.08% of no snow avalanche days. This bad result can be explained by the too numerous number of variables used for this analysis, compared to the number of observations.

The separation of the dataset "*DC41MTA01 – Warming*" from the whole previous dataset do not allow the finding of better classification rates for these typical situations. However, important discriminating variables can be identified more accurately to warming situations, and will be used in later NivoLog parameterization for these typical situations.

c. Parameterization and assessment of NivoLog performance

Previous statistical analyses allow the identification of important variables which discriminate days with and without snow avalanches. The aim of this section is to create different sets of parameters related to the snow avalanche contexts previously defined, to improve analyses in the system NivoLog. The sets of parameters are constructed on the basis of important variables determined by statistical analyses.

1. Fresh snow situations

1) Set of parameters

Different weights are given to variables which appeared significant in predicting the occurrence or not of snow avalanches. These variables are first selected on the basis of statistical analysis performed in the chapter IV.b.3. In this way, density of snow (MVS), snowdrift index (ID), days since the last snowfall exceeding 20cm (ADN20), snowfall in 24 hours (HN24) and snowfall during the last 3 days (Hn3J) get heavier weights in NivoLog. The more weighted variable is HN24, as it appeared as very important for the discrimination in previous statistical analyses. Then, MVS and Hn3J are equally weighted and ADN20 gets a lower weight. Concerning the snowdrift index, even if statistical tests showed that it was important, decision was taken not to weight this variable, because it relates to the typical context of snow transport (internal communication METEORISK). However, a filter is applied for the snowdrift index variable, meaning that only observations with snowdrift index lower than 10 grams can be compared, to avoid snow avalanche situations due to snow transport by wind.

Important variables emphasized by statistical analysis do not permit to classify snow avalanche days and no-snow avalanche days with 100% confidence. In this way, two other variables are weighted in addition to the important variables found during the statistical analysis, based on literature, on the expert knowledge in the domain of snow avalanches and advice of the office METEORISK. The first variable is Ps (probe penetration), because it indirectly indicates the cohesion of the top layer of snow and the depth of instable snow, which can potentially be mobilized to form snow avalanches. The second one is Tn, which gives information about the type of snow (very cold with no liquid water and a low cohesion; or temperate with the presence of liquid water and higher cohesion[4]) The weight of these variables are the same as the one for ID.

2) Application of the set of parameters

Once all variables appearing as important in order to discriminate days with and without snow avalanches are weighted in NivoLog, tests of classification are performed on 80 randomly chosen observations belonging to the fresh snow dataset ("*DC41MTA01 – FreshSnow*"). First, these tests allow the identification of errors, and secondly, the identification of situations for which the present set of parameters gives accurate results.

[4] NB: If the liquid water content become too high (more than 12% according to Ancey C., Gardelle F., Zuanon J.-P., (2003), the cohesion is reduced and instabilities can develop inside the snowpack.

- Identification of Errors

The first kind of error appears for observations with no observed snow avalanches, but eight or nine nearest neighbours over ten with at least one snow avalanche. In this case, the classification by NivoLog is wrong. However, it can also be due to error in the data, because all information was not correctly recorded. This is the case, for example, on the 16th of January 2004. At this date, snowfall in 24 hours is 15 cm, and snowfall during the last 3 days is 105 cm. However, no snow avalanche is recorded in the two files ("*DD41MTA01 - Original*" and "*DC41MTA01 - Cleaning A.5*"). The reason for this is that heavy snowfalls occurred from the 9th of January till the 15th of January, and snow avalanches have been artificially triggered by patrollers of the ski resort from the 9th January till the 15th January (at least one snow avalanche per day). So, on the 16th of January, no further snow avalanches can occur because all the slopes have been cleaned during the 7 precedent days. In such cases, NivoLog cannot take into account the fact that snow avalanches have already occurred and cleaned the slopes during the previous days, because no such variables have been included in the analysis. So, similar cases are left out for NivoLog performance validation, after verification of previous days in the two databases ("*DD41MTA01 - Original*" and "*DC41MTA01 - Cleaning A.5*").

The second kind of error appears for observations in which at least one snow avalanche was observed, but factors leading to snow avalanche occurrence do not seem to be present. The possible explanation for that is a delay in artificially triggering snow avalanches. For example, it is possible that heavy snowfall occurred during the previous days, but people responsible for the security of the ski resort did not have time to trigger avalanches. So, no snow avalanches are recorded during these days with important snowfalls, and then, several days later, when no factors seem important, one natural snow avalanche occur. However, these cases are difficult to distinguish and cannot be removed with certainty. In this way, they are kept as unidentified in the database, but they should be source of classification errors in testing the system performance of NivoLog.

Solutions to these two types of problem leading to bias in the analysis can be to add variables describing the period of time since the occurrence of the last snow avalanche. This would allow a better classification in NivoLog because the system could take into account the cleaning of the slopes during the previous days. Another solution could be to take into account only the snowfall in 24 hours if triggering of snow avalanches took place during the last three days.

- Identification of situations with high classification performance

In two cases, the set of parameters for fresh snow situations gives good results concerning the classification rate of days with and without snow avalanches. The first case is for observations in which heavy snowfall takes place during the last 24 hours and the second case is when heavy snowfalls takes place during the last 3 days (HN24 $\geq$ 20 cm or/and Hn3J $\geq$ 35 cm). Furthermore, no refreezing is present. For these observations, the occurrence of snow avalanches is frequent, and NivoLog shows a good performance in predicting it. However, for situations in which lower quantities of snowfall are recorded (HN24 lower than 20 cm and Hn3J lower than 35 cm), the present set of parameters is not appropriate and gives mediocre results.

To encounter this problem, the characteristics of "fresh snow situations" are redefined: HN24 needs to be either lower than 5 cm or greater or equal to 20 cm, Hn3J must be greater or equal to 35 cm and no refreezing must be present. For observations with these characteristics, the set of parameters for fresh snow situations can be applied and gives good results. Situations with snowfall in 24 hours between 5 cm and 20 cm, with snowfall in 3 days lower than 35 cm and possible refreezing, are not accurately classified. So, these situations are considered as atypical situations, and the set of parameters to apply for prediction in these cases is the one for atypical situations (see Chapter IV.c.5).

3) Validation and decision rules

As explained before, the set of parameters for fresh snow situations can only be applied if HN24 is greater or equal to 20 cm, or if Hn3J is greater or equal to 35 cm and no refreezing takes place. So, a new dataset for these situations is created, called "*DC41MTA01 – FreshSnow-NivoLog*" and made up of 191 observations. For the validation of the set of parameters for fresh snow situations, 55% of the situations of "*DC41MTA01 – FreshSnow-NivoLog*" are selected[5]. The structure of the sample for snow avalanches occurrence is conserved: 50% of observations with snow avalanche and 50% without snow avalanche. This leads to a validation-sample of the file "*DC41MTA01 – FreshSnow-NivoLog*" with 105 observations, 53 being no snow avalanche days and 52, snow avalanche days.

The calculation of RCR in NivoLog depends on the number of nearest neighbours we decide to take into consideration to classify an observation in snow avalanche day or not. In other words, an observation can be classified as a snow avalanche day if 2 nearest neighbours over 10 are snow avalanche days, or if 5 nearest neighbours over 10 are snow avalanche days, etc. For the set of parameters for fresh snow situations applied on the validation sample, the number of nearest neighbours which gives the best RCR is 6. It means that, if 6 or more nearest neighbours over 10 are snow avalanche days, the observation will be classified as a snow avalanche day. With this rule of decision for the set of parameters for fresh snow situations, the table below is obtained:

		Predicted	
		0	1
Observed	0	**49**	*4*
	1	*10*	**42**

91 observations are rightly classified (numbers in boldface) and 14 are wrongly classified (numbers in italic) over 105. This gives a RCR of 86.67%.

In short:

Snow situations are defined as HN24≥20cm, or Hn3J≥35cm and RS=0.
In NivoLog, a day is considered as a snow avalanche day when 6 or more nearest neighbours over 10 are snow avalanche days, leading to a RCR of **86.67%.**

[5] For the validation of each set of parameters, the same percentage of observations is chosen to create validation samples. 55% allows the consideration of a great part of the observations, but it seems necessary due to the small number of observations for each sample.

2. Snow transport situations

1) Set of parameters

The set of parameters for snow transport situations is constructed by giving different weights to variables appearing as important in predicting snow avalanche days or not. They are first selected on the basis of statistical analyses performed in chapter IV.b (LDA and LR) and secondly, according to literature, expert knowledge (R. Bolognesi) in the domain of snow avalanches, and internal communications in the office METEORISK.
Statistical analyses emphasized that MVS (snow density), Tn (snow temperature), Ps (probe penetration), ID (snowdrift index) and Hn3J (snowfall during the last 3 days) are the five most important variables to discriminate days with and without snow avalanches. So, they are given heavier weights than other variables in NivoLog: ID is the more weighted variable, because it is the direct illustration of snow transport by wind. Then, Hn3J and Ps are equally weighted. Finally, MVS gets a lower weight and Tn even a lower weight.
Other variables are important for the discrimination of days with and without snow avalanches for snow transport situations. This is the case of HN24 (snowfall during the last 24 hours), ADN20 (number of days since the last snowfall exceeding 20 cm) and tHs24 (settling of the snowpack in 24 hours). HN24 and ADN20 give information on the amount of snow which can be mobilized by wind, tHs24 indirectly indicate the cohesion of snow in the top layer of the snowpack (if the cohesion is low, snow can easily be mobilized by wind), (internal communication METEORISK).

2) Application of the set of parameters

First, the set of parameters is tested on 100 randomly chosen observations without conserving the structure of the sample. This test is only performed to identify possible errors in the data or in the variables which have been weighted. This first test showed no errors or suspect observations, and the results were satisfying.
Secondly, another test is performed with 150 other randomly chosen observations, to increase the possibility of finding errors. No errors are detected, and the result of this test is only 4% different from the previous test.
The random testing of the set of parameters for snow transport situations do not allow to delete observations which show errors. Furthermore, it seems that variables weighted for the parameterization are coherent and allow a good classification of days with and without snow avalanches. However, the two previous tests were performed on a defined number of observations from the dataset "*DC41MTA01 – SnowTransport*", and the structure for days with or without snow avalanches was not respected. So, validation of this set of parameters needs to be performed on a more strict selection of the sample.

3) Validation and decision rules

First of all, it is important to indicate that NivoLog can only perform analyses when all weighted variables are complete for each observation to classify. In other words, if the observation to classify has a missing value for one of the weighted variable, NivoLog cannot perform the analysis. So, it is important to delete all missing values for the weighted variables in the basis file "*DC41MTA01 – SnowTransport*". This leads to the creation of the file "*DC41MTA01 – SnowTransport-NivoLog*", made up of 199 complete observations.
Secondly, a sub-sample needs to be selected for the validation of the set of parameters for snow transport situations. The same percentage as other typical situations is selected: 55% of the observations from the basis file "*DC41MTA01 – SnowTransport-NivoLog*" are selected, which means a validation-sample of 109 observations. Furthermore, the structure of the sample must be the same. It means that the proportion of snow avalanche and no snow avalanche days is the same in the basis

sample and in the validation sample. 20% of the observations are days without snow avalanches, and 80% are days with snow avalanches. For the validation sample, this leads to 22 observations without snow avalanches and 87 with snow avalanches.

The calculation of RCR in NivoLog depends on the number of similar nearest neighbours we decide to take into consideration to classify an observation in snow avalanche day or not. For snow transport situations, the best RCR is obtained when the following rule is chosen: if 5 or more nearest neighbours are snow avalanche days, the observation in consideration will be classified as a snow avalanche day too. With this decision rule, the following table is obtained:

		Predicted	
		0	1
Observed	0	**14**	*8*
	1	*12*	**75**

89 observations are rightly classified (in boldface) and 20 are wrongly classified (in italic) over a total of 109 observations. This leads to a RCR of 81.65% for situations of snow transport by wind.

In short:

In NivoLog, a day is considered as a snow avalanche day when 5 or more nearest neighbours over 10 are snow avalanche days, leading to a RCR of **81.65%.**

3. Rainfall situations

Rainfall situations are only made up of 36 observations, which lead to dubious statistical results and cautious conclusions (see Chapter IV.b.3). For NivoLog parameterization, we tried to find rules anyway, just for interest, but it is important the keep in mind that statistics and parameterization performed on such a small sample cannot be valid.

1) Set of parameters

Different weights are given to variables which appear significant in predicting the occurrence or not of snow avalanches. These variables are first selected on the basis of statistical analysis performed in the chapter IV.b.3. R24 gets the higher weight because it directly relate to rainfall situations. Then, Tn and Hn3J get lower but equal weights in the parameterization (internal communication METEORISK).

Other variables also are weighted because if only R24, Tn and Hn3J are used to discriminate days with and without snow avalanches, the best classification rate only reaches 87.5% and other variables can potentially increase this number. Furthermore, knowledge of the office METEORISK in the domain of snow avalanches allow the selection of other variables known as important in discriminating days with and without sow avalanches in rainfall situations. These other variables are HN24, ADN20 and MVS (internal communication of the office).

2) Application of the set of parameters

Application of the set of parameters on rainfall situations require that each observation has complete values for each variable weighted in the system. In these typical situations, only 34 observations are complete for each weighted variable. As the number of observations is small for rainfall situations, the set of parameters is applied on the 34 observations, to detect some errors in the parameterization.
For each observation to be predicted by NivoLog, three nearest neighbours are found, all being snow avalanche days, even if the observation in consideration has no snow avalanche. As the three nearest neighbours are always a snow avalanche day, the percentage of right or wrong classification is always 100% (3/3). In this way, the five observations with no snow avalanche are systematically 100% wrongly classified and the 29 observations with snow avalanches are systematically 100 % rightly classified. This is illustrated by the table below:

		Predicted 0	Predicted 1
Observed	0	**0**	*5*
	1	*0*	**29**

With the set of parameters for rainfall situations, 29 observations are 100% rightly classified (3 nearest neighbours over 3 indicate a snow avalanche day) over 34, leading to a RCR of 85% for these situations.

3) Validation and decision rules

On the 34 observations previously considered, a RCR of 85% was reached, with the set of parameters for rainfall situations. However, even if this classification rate seems good, these situations are not predicted with high confidence by NivoLog. In fact, even if the amount of rainfall is very low (1-3 mm of precipitation), the system predicts a snow avalanche day, while the observations in consideration had no snow avalanche. This is the case for all five situations systematically wrongly classified, in which no snow avalanche occurred, but NivoLog predicted a snow avalanche.
This systematic wrong classification of observations without snow avalanches can have three main reasons. First, the set of parameters can be inappropriate for predicting rainfall situations. However, statistical analyses (LDA and LR) have been performed to find important discriminating variables, and the results seemed coherent with literature and knowledge in the domain of snow avalanches. Furthermore, the office METEORISK validated the set of parameters to use in rainfall situations, which appeared as rightly weighted for snow avalanches prediction in such situations. The second reason has already been mentioned in previous chapters, and refers to the very small number of observations for these situations. This first limits statistical analyses as LDA and LR for the finding of important variables, but then, also the algorithm of kNN used in NivoLog, because the number of observations to compare is small too. An improvement would be to increase the number of observation years to increase rainfall situations too. However, at an elevation of 2300 m., rainfall situations are relatively rare, and their number of observations will remain small. The third reason is the lack of appropriate variables to discriminate days with and without snow avalanches. The fact that observations with small amount of rainfall are wrongly classified as snow avalanche days can be due to missing discriminating variables giving information on the depth of infiltration of rain in the snowpack. In other words, if rain infiltrates the 10 first centimetres of snowpack only, this will not lead to a great destabilization if the snowpack is thick. However, if the whole snowpack is fully moistened by rainfall, cohesion between snow grains becomes very weak and snow avalanches are more likely to occur (internal communication by METEORISK). In this way, a variable measuring the depth of moistening of the snowpack could lead to possible better discrimination between days with and without snow avalanches for rainfall situations.
In conclusion, it is not possible to predict snow avalanche occurrence in rainfall situations with a high confidence. NivoLog tends to classify systematically a day in which rainfall occurs as a snow avalanche day, even if the amount of rain is small. However, this prediction inability only concerns 36 days over 2267 (1.59% of observations), and only 5 are wrongly classified over these observations, leading to a wrong classification of 0.22%. So, the basic rule found by NivoLog parameterization is that if rainfall occurs, snow avalanches are likely to occur too.

4. Warming situations

1) Set of parameters

The set of parameters used for warming situations is constructed on the basis of important variables from two information sources: first, statistical analyses performed with the software R, (more particularly Logistic Regression and Linear Discriminant Analysis), and secondly, literature, expert knowledge in the domain of snow avalanches (R. Bolognesi) and internal communications in the office METEORISK.
LDA and LR analyses emphasize that Ta (air temperature), Tn (snow temperature), ADN20 (number of days since the last snowfall exceeding 20 cm) and Hn3J (snowfall during the last 3 days) are the best discriminant variables to distinguish days with and without snow avalanches, because they give the highest RCR. Tn is the more weighted variable, then Ta and ADN20 are equally weighted and Hn3J is given lower weight (internal communication METEORISK).

According to expert knowledge and internal communication in the office, other variables are weighted for warming situations because they are important for the discrimination between days with and without snow avalanches. This is the case of RS (thickness of surface refreezing), N (cloud cover), MVS (snow density), tHs24 (compaction of the snowpack in 24 hours) and ADP (age of the last rainfall). RS and N are important because when they are combined, they give an overview of the evolution of conditions during the day. If RS is high and cloud cover too, snowpack will remain more stable than if cloud cover is low and sun melts rapidly the surface of refreezing. As for other situations, MVS gives information on the state of snow in the top layer, and in some way, on its cohesion too. If Tn is low and MVS high, snow is relatively compact and stable; on the contrary, if Tn is close to zero and MVS high, liquid water can be present in the snowpack and the stability is reduced compared to previous conditions. tHs24 refers to the stabilization of the snowpack during the last 24 hours, and in particular cases of warming, to the melting of snow in the top layer. Finally, ADP gives indirect information on the humidification of the snowpack. Generally, if the snowpack is dry, liquid water due to snow melting or rain will infiltrate and snowpack is not too much destabilized if quantities of liquid water remain low. However, if the snowpack is already saturated in liquid water (due to warm temperatures, strong melting or heavy rainfall), addition of new liquid water will generally lead to a destabilization.

2) Application of the set of parameters

First, the set of parameters previously constructed is tested on 50 randomly selected observations in the dataset for warming situations, without keeping the structure of days with and without snow avalanches repartition. The results of this test are not satisfactory because the RCR only reaches 68%. So, a second test is performed on 100 other randomly selected observations, without keeping the structure of days with and without snow avalanches. This second test gives even lower RCR: 62%.
As these results are not really satisfactory, the set of parameters is reviewed and certain parameters are weighted differently. The two previous tests are performed again with this new set of parameters, but results are not improved and RCR remains close to 65%. Furthermore, no typical characteristics can be found in the data to select only cases for which the set of parameters seems appropriate. In this way, validation is performed with the first set of parameters presented above.

3) Validation and decision rules

As for the validations of previous sets of parameters (fresh snow situations, snow transport situations), the validation sample must have complete observations. This means that NivoLog cannot perform analyses if missing values are present for the weighted variables. In this way, the first step is to create a complete dataset with no missing values for the weighted variables. So, after the cleaning, the complete dataset is called "*DC41MTA01 – Warming-NivoLog*", and only 71 complete observations remain.
The second step before applying the validation of the set of parameters for this situation is to create a sub-sample (or validation sample), on the basis of the complete dataset "*DC41MTA01 – Warming-NivoLog*". To be coherent with previous validations, 55% of the observations are selected, and the structure of the sample concerning days with and without snow avalanches is conserved. This later is the following: 70% of observations are days without snow avalanches and 30% are days with at least one snow avalanche. So, the validation sample has a size of 39 observations, split in 27 in which no snow avalanche occurred and 12 in which at least one snow avalanche occurred.
As already explained, the calculation of RCR in NivoLog depends on the number of nearest neighbours we decide to take into consideration to classify an observation in snow avalanche day or not. For warming situations, the best RCR is obtained when the following rule is chosen: if 4 or more nearest neighbours are snow avalanche days, the observation in consideration will be classified as a snow avalanche day too. With this decision rule, the following table is obtained:

		Predicted 0	Predicted 1
Observed	0	**22**	*5*
	1	*6*	**6**

28 observations are rightly classified (in bold), meaning that prediction by NivoLog corresponds to the real observation, and 11 are wrongly classified (in italic), meaning that the prediction do not correspond to real observation, over a total of 39 observations. This leads to a RCR of 71.79% for warming situations.

In short:

In NivoLog, a day is considered as a snow avalanche day when 4 or more nearest neighbours over 10 are snow avalanche days, leading to a RCR of **71.79%.**

5. Atypical situations

In chapters C.1 to C.4, four different types of snow avalanche contexts have been defined, and for each of them, a particular set of parameters was created. However, there are many days, which cannot be classified in one of the four typical snow avalanche contexts. So, for these atypical situations, a general set of parameters must also be defined. It is recommended to use it for the prediction of observations in which no identifiable or clear event indicates their classification in one of the four snow avalanche situations defined in the chapter IV.b.3.1. In this way, a new file with observations belonging to none of the four typical situations must be constructed. It is called "DC41MTA01 – Atypical - NivoLog" and is made up of 777 complete observations.

1) Set of parameters

The set of parameters for atypical situations is constructed by weighting variables considered as important in discriminating days with and without snow avalanches. The choice of these variables is first made on the basis of statistical analyses performed in chapter IV.b, but also on the basis of literature, expert knowledge in the domain of snow avalanches (Robert Bolognesi) and internal communication in the office METEORISK.
Linear discriminant analysis and logistic regression emphasize that important variables for all situations (analyses performed on the whole dataset) are: snow density (MVS), air temperature (Ta), snowdrift index (ID), snowfall in 24 hours (HN24) and snowfall during the last three days (Hn3J). So in NivoLog, these variables get heavier weights. The heavier weight is given to the variable HN24, because it appeared to be one of the most influent variables in the discrimination between days with and without snow avalanches during statistical tests (for LR it is influent at a confidence level of more than 99.9% - cf. Chapter III.a). Then ID and Hn3J also get important weights because they are important in discriminating days with and without snow avalanches as shown by statistical tests and literature (Bolognesi, R., 2015; Bellot H., Bouvet, F.N., 2010; Pahaut E., Bolognesi R., 2003). MVS and Ta get lower weights, but are also important for the discrimination during atypical situations.
According to expert knowledge and internal communication in the office, other variables are additionally weighted for atypical situations because they are known as important for the discrimination between days with and without snow avalanches. This is the case of ADP (age of the last rainfall), ADN20 (age of the last snowfall exceeding 20 cm) and RS (thickness of surface refreezing). ADP and ADN20 are additionally weighted because they give important information on the past conditions of the observation in consideration (internal communication METEORISK). RS is also weighted because it indirectly indicates the stability of the top layer of snow: if RS is thick, the snowpack is more stable than if no refreezing is present.

2) Application of the set of parameters

The first thing to notice before the application of the set of parameters is that the file for atypical situations is 3 to 7 times greater than files for typical situations (fresh snow, snow transport, rainfall, warming). As the application is only useful to identify errors in the data or an unsuitable weighting, the set of parameters for atypical situations is only applied on 100 randomly chosen observations, without keeping the structure of the sample concerning days with or without snow avalanches. This application leads to two main observations.
First, the classification of days with and without snow avalanches is quite good, around 70% depending on the number of neighbours we considered for the decision. Secondly, days without snow avalanches are better rightly classified than observations with snow avalanches. For example, an observation with no snow avalanche will have 7 to 10 nearest neighbours over 10 with no snow

avalanche too, while an observation with at least one snow avalanche will only have 1 to 4 nearest neighbours with snow avalanche too. This is inherent to atypical situations. Indeed, in previous chapters, we selected typical snow avalanche contexts to increase the RCR of observations belonging to these contexts. In other words, days with at least one snow avalanche are better classified than if a general set of parameter is applied for their classification. So, observations with at least one snow avalanche, but no criteria to belong to typical snow avalanche contexts will be badly classified by the set of parameters for atypical situations. This is what we can see with the number of similar nearest neighbours.
Otherwise, no obvious errors in data or apparent bad weighting appear during the first application of the set of parameters on random observations. In this way, it can be validated, and the decision rule to assess the membership of an observation to the snow avalanche or no-snow avalanche days can be found.

3) Validation and decision rules

As for all typical snow avalanche situations, a sub-sample must be constructed for the validation. This validation sample must have complete values for each observation and each variable, which is weighted in the particular case of atypical situations. If this condition is not fulfilled, NivoLog cannot perform analyses for observations with missing values for weighted variables. However, during the selection of the sample for atypical situations, only complete observations were selected already.
As for other validation samples, 55% of the observations of "DC41MTA01 – Atypical – NivoLog" are randomly selected. However, in the present case, the structure of the sample concerning days with and without snow avalanches is conserved. This leads to a validation sample of 427 observations, with 70% of days without snow avalanche (299 observations) and 30% of days with snow avalanches (128 observations).
As already mentioned, the RCR depends on the number of similar nearest neighbours we decide to consider. In the case of atypical situations, the best decision rule is the following: if 5 or more nearest neighbours are snow avalanche days, the observation in consideration will be classified as a snow avalanche day too. With this decision rule, the following table is obtained:

		Predicted	
		0	1
Observed	0	**289**	*10*
	1	*84*	**44**

317 observations are rightly classified (in dark), meaning that prediction by NivoLog corresponds to the real observation, and 110 are wrongly classified (in red), meaning that the prediction do not correspond to real observation, over a total of 427 observations. This leads to a RCR of 74.98% for atypical situations.

In short:

In NivoLog, a day is considered as a snow avalanche day when 5 or more nearest neighbours over 10 are snow avalanche days, leading to a RCR of **77.98%.**

V. DISCUSSION

Many datasets used for statistical analyses initially include errors at the beginning of a study. In the present work, the meteorological and snowpack data of the ski resort of Aminona contained some errors, which have been cleaned for further analyses. These errors can have various causes, and related solutions to improve the analysis. First, the measurement devices can be a source of errors, if they have an incorrect calibration or a bias in their measurements. To reduce this source of errors, it may be possible to control measurements devices before the study, and find a function to correct related errors in the data. Secondly, people responsible for the collection of data can also be a source of errors, by wrongly reporting the value of measurement, inadequately selecting the place of measurement, estimating a value or rounding it. The reduction of these kinds of errors is difficult unless measurements are done by the person performing the study afterwards. Thirdly, errors can come from the location of measurements station itself, if it is not representative of the whole area, or if it is influenced by terrain, infrastructures or vegetation. For example, the measurement of snowdrift index is biased if the place of measurement is behind a building or in a depression in the terrain. As for the previous type of error, this one can only be corrected at the time of site selection, and so, is difficult to eliminate. Concerning the dataset about snow avalanches characteristics, other kinds of errors can appear. The first error can appear when all conditions for a snow avalanche to occur are present, but no snow avalanche is recorded in the database. For example, during a day with heavy snowfall and wind, it is possible that the ski resort is closed and no one is present to trigger snow avalanches. In this way, conditions for snow avalanches are present, but no snow avalanche occur because the security of the ski resort is not guaranteed when it is closed to the public. The second type of error is that a snow avalanche can be recorded in the database while none of the conditions for a snow avalanche to occur are present. This can also be explained by the fact that, during bad weather conditions, the ski resort can be closed. So, snow avalanches are not artificially triggered during these days. However, when the ski resort opens again, patrollers must secure the ski slopes, and so, trigger snow avalanches. In this way, snow avalanches occur due to past conditions, but conditions of the day are not coherent with a triggering of snow avalanches.
For the cleaning of errors from meteorological and snowpack data, coherence tests have been performed to eliminate obvious errors which were out of the domain of definition, and verify the plausibility of dubious values which were outside the 2d and 98th percentiles. For errors concerning snow avalanches, verification was done with the dataset of snow avalanches called “DD41MTA01 – Original”. Snow avalanches present in this file must correspond to those present in the file of meteorological and snowpack data called “DC41MTA01 – Original”. These tests and cleaning allow the elimination of obvious errors, but some of them are still present and cannot be identified, because they are not dubious enough. In this way, the cleaning step certainly improved further statistical analyses results, but datasets totally free of errors are not reachable in such a study. At a certain point, further cleaning and tests take a lot of time to only eliminate partial remaining errors.

Once meteorological and snowpack dataset was cleaned, statistical analyses were applied in order to determine which variables are important in discriminating days with and without snow avalanches. First, linear discriminant analysis and logistic regression were used to find variables which are important for the discrimination of days with and without snow avalanches, and RCR. Then, the methods of p-values for classification and kNN analysis were also used for an additional assessment of RCR and a comparison with LDA and LR RCR. Furthermore, the method of p-values for classification allowed the finding of a RCR according to a certain confidence level. Results for the whole dataset were satisfactory (RCR of 74.39% for LDA and 75.95% for LR), and coherent discriminating variables for both analyses were found (MVS, Ta, ID, HN24 and Hn3J). However, hypothesis was made that if typical snow avalanche situations are selected, better RCR could be reached, and more accurate important variables could be found. In this way, four different snow

avalanche contexts were defined, on the basis of the book “Estimer et limiter le risque avalanche” (Bolognesi R., 2013). Results about most important variables and best RCR are presented in the table below, for each typical snow avalanche situation; they also answer the question of research about which variables are important in discriminating days with and without snow avalanches. Explanations for each important variable found with LDA and LR are presented in discussions after presentations of the results (see chapter IV.b). In a general manner, the splitting into four different snow avalanche situations led to an improvement of the RCR, and the finding of more accurate important variables, except for fresh snow situations. In this latter case, the RCR remains quite the same, but important variables seem more coherent.

Typical situation	**LDA**		**LR**	
	RCR	Important variables	RCR	Important
Whole dataset	74.39%	MVS, Ta, dTa, ID, HN24, Hn3J	75.95%	N, MVS, Ta, R24, ADP, ID, HN24, Hn3J,
Fresh snow	72.58%	MVS, Tn, ADN20, ID, HN24, Hn3J	74.60%	RS, N, MVS, ADN20, ID HN24, Hn3J
Snow transport	81.52%	MVS, Tn, Ps, ID, Hn3J	81.52%	Hn3J
Rainfall	87.50%	R24, Tn, Hn3J	87.50%	/
Warming	76.81%	MVS, Tn, Ta, ADN20, ADP	76.81%	Ta, Tn

Table 6: Results of LDA and LR with important variables and RCR.

The results of these tests allow the answering to other research questions presented in the introduction part of this work. First, indices for snowdrift index did not need to be constructed, because snowdrift was directly measured in Aminona by a driftometer, and so, no indices or modelling with wind speed were needed. However, it was decided not to take into consideration wind direction and snowdrift direction, because different studies showed that wind direction is not a significantly important variable, and because snowdrift direction was not well represented in the data available for this study. In this way, no mathematical transformation was applied on these two variables; they have only been left out of the analysis. Secondly, statistical analyses showed that not all causal variables selected for this study are important in discriminating days with and without snow avalanches. This is the case of wind speed, air temperature variation in 24 hours and cloud cover. Wind speed never appeared as significantly important in discriminating days with and without snow avalanches, because it is not a direct cause of snow avalanches occurrence; this is the transport of snow by wind which matters, and this quantity is represented by the snowdrift variable. The air temperature variation in 24 hours also never appeared as important, which is understandable: this variable only measures the air temperature difference between two points in time, but does not give an idea of temperature evolution over a period. Cloud cover does not seem important too, with respect to statistical analyses. This is understandable because it does not give information about its evolution; it only indicates the cloud cover at one point in time. However, when it is combined with RS (thickness of surface refreezing) it can indicate the evolution speed of snowpack melting for warming situations.

Thirdly, it appeared that meteorological and snowpack variables are both important for the discrimination of days with and without snow avalanches. In fact, this is the combination of both types of variables which influences the snowpack stability and the occurrence of snow avalanches. Fourthly, no comparisons can be done with other ski resorts as it was expected at the beginning of the study. However, similar work could be performed by the office METEORISK for future customers who will provide adequate data. Fifthly, important variables in discriminating days with and without snow avalanches in literature are not systematically the same as our statistical results. New precipitations and foot penetration, which were found as important in the literature (in the present work HN24, Hn3J and Ps), also appeared as important after our statistical analyses. However, present temperature trend and wind speed (Vv) are variables which did not appear as important for the present study. For Vv, this difference is due to the consideration of snowdrift index, which better discriminate days with and without snow avalanches than wind speed. For the present temperature trend, measurements may be taken differently, and so, dTa and the present temperature trend variable found in literature cannot be compared. Concerning the answer to research questions about statistical methods, one can say that LDA and LR approximately give similar results for the classification rate, with 1% to 3% of difference. However, as already explained, rainfall situations have a too small number of observations to obtain valid results. Generally, LR has a quite higher RCR than LDA. This can be explained by the fact that LR fits a log-function to the data, which better follows the possible patterns, while LDA fits a linear function, which is quite rigid compared to LR. However, the results of these two analyses are very similar. Concerning important variables, only those which were important for both methods are selected for NivoLog parameterization, but quite similar important variables are found for LDA and LR.

Important variables found by statistical analyses have then been weighted in the system NivoLog; other variables have been added to the analysis based on expert knowledge and literature. The aim of this weighting is to create four sets of parameters corresponding to typical snow avalanche contexts previously defined. For each snow avalanche situation, only complete observations are selected, and validation samples are randomly chosen with 55% of the basis observations and the same structure concerning snow avalanche days. Results in NivoLog depend on the number of similar nearest neighbours that the user wants to consider to classify an observation in snow avalanche or no-snow avalanche day. So, decision rules need to be added to the sets of parameters to reach the best RCR. These decision rules and RCR are summarized in the table below:

Situation	Decision rule associated to the set of parameters	RCR
Fresh snow situations	If **6** or more nearest neighbours over 10 are snow avalanche days, the observation in consideration will be classified as a snow avalanche day.	86.67%
Snow transport situations	If **5** or more nearest neighbours over 10 are snow avalanche days, the observation in consideration will be classified as a snow avalanche day.	81.65%
Rainfall situations	/	85% (!)
Warming situations	If **4** or more nearest neighbours over 10 are snow avalanche days, the observation in consideration will be classified as a snow avalanche day.	71.79%

Table 7: Decision rules and RCR of the system NivoLog.

For each snow avalanche situation, if the corresponding decision rule is used to decide if an observation will be classified in snow avalanche day or not, RCR are quite high (71%-86%). For rainfall situations, the too small number of observation did not allow the finding of decision rule, because when such situations occur, NivoLog classifies observations as snow avalanche days in any case. To improve the understanding of snow avalanches triggered during rainfall situations, more numerous observations should be studied. However, rainfall during winter season at an altitude of 2300 m. is relatively rare, and so, the number of observations will remain small in any case. Warming situations are those which have the lowest RCR, because these situations are difficult to classify (internal communication METEORISK). The reason is that the evolution of snowpack conditions is very slow (increasing snow temperature, melting snow grains, reducing spaces between crystals etc.) compared to direct meteorological influences, and thus, the response (snow avalanches occurrence) to a warming situation is not direct due to inertia.

But often, observations cannot be classified in one of these four typical snow avalanche contexts, because no important characteristic event occurs. So, a dataset of atypical situations was additionally constructed, with all observations which cannot be classified in one of the four typical snow avalanche contexts. As for typical situations, a set of parameters is constructed for these atypical situations, and validation is performed in the same way. Decision rule also need to be found to obtain the best RCR. This is illustrated in the table 7 below:

Situation	Decision rule associated to the set of parameters	RCR
Atypical situations	If **5** or more nearest neighbours over 10 are snow avalanche days, the observation in consideration will be classified as a snow avalanche day.	77.98%

Table 8: Decision rule and RCR of the system NivoLog for atypical situations.

Results in NivoLog for atypical situations depend much on the type of snow avalanche day. When observations are snow avalanche days, only 1 to 4 nearest neighbours over 10 are snow avalanche days too; while when observations are no-snow avalanche days, 7 to 10 nearest neighbours over 10 are days without snow avalanche too. In other words, snow avalanche days are badly classified by NivoLog when using the set of parameters for atypical situations. This can be explained by the selection of typical snow avalanche contexts. Indeed, the four typical snow avalanche contexts are good for the classification of snow avalanche days, because they relate to typical conditions leading to the occurrence of snow avalanches. So, for snow avalanche days with no criteria to be classified in one of the four snow avalanche contexts, the classification is not good when using the set of parameter for atypical situations. In fact, this latter should only be used for observations which can really not be classified in one of the four typical snow avalanche context.

Tables 7 and 8 also show that the RCR of warming situations classified by NivoLog is about 6% lower than the RCR of atypical situations. So, the question could be asked if warming situations cannot be included in atypical situations and if it is not better to use the set of parameters for atypical situations when one wants to classify warming situations. Indeed, when applying the set of parameters for atypical situations to the warming situations, the RCR only reaches 64.10%, compared to 71.79% when using the set of parameters for warming situations. In this way, even if the RCR of warming situations is lower than the one for atypical situations, it is in any case better to use the set of parameters for warming situations in these typical situations.

Even if the findings of this work are satisfactory, three different ways can be suggested to improve results and knowledge in the domain of snow avalanches. The first way is to add new variables to the analysis. For example, variables which could describe the thickness of rain percolation in the snowpack, the type of snow grain, the presence or not of people triggering snow avalanches (to avoid errors due to triggering snow avalanches with delay), etc. Furthermore, variables already present in the database but not well recorded could be improved. This is the case for snowdrift direction, which could be a very interesting variable, but is too badly represented in our database. If measurements were available for each day, better results could be found. The adding of snowdrift direction in our analysis could lead to snow avalanches forecasting at a more local scale, because characteristics of each snow avalanche paths could be described according to their orientation. The second way of improvement would be to enlarge the number of data available. Of course, for an analysis performed indifferently on the whole dataset, number of observations in our case is sufficient. But when we decide to divide the dataset in sub-datasets for typical situations, it becomes clear that the number of observations is too small for some of them (case of rainfall situations). In this way, a wider basis sample could lead to more significant results. Finally, results can be improved by a better cleaning of initial errors. In fact, even if coherence tests and cleaning was performed in chapter IV.A, some errors remain in the dataset because they are obviously not out of the definition domain or the norm. However, with additional coherence tests, it would be possible to detect and remove them from the basis data. The only thing to be aware of is that further tests will spend a lot of time to only remove a small number of errors.

VI. CONCLUSION

The present work concerned the analysis of 19 winter seasons in the ski resort of Aminona, Valais, with a particular focus on the discrimination of days with and without snow avalanches based on meteorological and snowpack variables. First, cleaning of the dataset was performed to remove obvious errors and improve the results of further statistical analyses. Secondly, linear discriminant analysis and logistic regression were used to identify important discriminating variables and to assess the RCR of days with and without snow avalanches. The methods of p-values for classification and kNN were used to an additional assessment of RCR of LDA and LR. More particularly, the method of p-values indicates the RCR which can be reached with a certain confidence, even if the problem is intrinsically difficult. These statistical analyses were first performed on the whole dataset, and secondly on sub-datasets corresponding to typical snow avalanche contexts. The aim was to find important variables corresponding to each context, and to improve the RCR of days with and without snow avalanches. Thirdly, for each of the typical snow avalanche situations, a set of parameters was constructed on the basis of statistical results and expert knowledge. In addition, a fifth set of parameters was constructed for observations which cannot be assigned to one of the four snow avalanche situations. So, when people responsible for the security in a ski resort want to predict if a day is a snow avalanche day or not, they should first identify if it corresponds to a typical snow avalanche context, and then, use the set of parameter for this context in the system NivoLog, provided by the office METEORISK.

a. Nivolog improvement

The system NivoLog, developed and provided by Robert Bolognesi (METEORISK) is a powerful tool for the local forecasting of snow avalanche occurrence. This system works on the basis of kNN analysis: for each new observation to predict, NivoLog calculates its distance with other observations in the database, and the k nearest neighbours (set by the user) are displayed in the results widow. The weighting of some variables known as important for discriminating days with and without snow avalanches can be done by the user itself, and forces the analysis to obtain more accurate results for particular well known situations.

Initially, NivoLog only had one general set of parameters present by default in the system, with a weighting of main discriminating variables, to help decision makers to use it (but the user also had the possibility to change the weighting as he wished). The present work provides a great improvement of the system by using different sets of parameters associated to typical meteorological contexts, and by providing decision rules which give the best RCR. A case, which we want to predict, must first be identified as a typical snow avalanche context (fresh snow, snow transport, rainfall, warming or an atypical situation. Then, analysis in NivoLog is run with the corresponding set of parameters (one for each typical snow avalanche situation or atypical situation). The RCR of NivoLog depends on the number of nearest neighbours taken into consideration. So, decision rules associated to each set of

parameters are provided, to select accurately the right number of nearest neighbours for the classification of an observation day.

Example: the person who is responsible of security for the ski resort of Aminona arrives in the morning and observes that strong winds and strong snow transport took place during the last night. He wants to know if the snow avalanche situation is critical and if people are endangered on the ski slopes. So, he runs NivoLog to know if artificially triggering of snow avalanches will give good results and if it is necessary to secure ski slopes for the coming day. First, he identifies this day as a typical snow avalanche situation, and more particularly, a snow transport situation. He loads the set of parameters for such situations, and runs the algorithm. The result is: 6 nearest neighbours over 10 are snow avalanche days. So, he looks at the decision rule for snow transport situations to obtain best results. This rule says that "*if 5 or more nearest neighbours are snow avalanche days, the observation in consideration will be classified as a snow avalanche day*". So, in this example, the responsible for the security of ski slopes in Aminona will consider that the day in consideration is likely to be a snow avalanche day, and he will probably artificially trigger snow avalanches to secure the ski area.

b. Concrete application and benefits

First, as presented above, the present work can be considered as an improvement for the system NivoLog, and a better understanding of typical snow avalanche contexts. However, this study was only made for one ski resort in Valais and could hardly be generalized to other parts of the world, because of different climatic and terrain characteristics. In this way, the present work could be a source of research and development for the office METEORISK, depending on which customers will need to use NivoLog. Similar studies could be performed in the future, to improve NivoLog results, for different regions in Switzerland, or different parts of the world.

Secondly, the present academic work also has concrete applications in the domain of security concerning snow avalanches. The office METEORISK has now the possibility to offer an analysing tool with different sets of parameters and associated decision rules to its customers, to better accompany their decisions during winter seasons. In practical, results of this work are described and explained in the user's manual delivered with the system, and the sets of parameters are included in the software for customers' use.

Thirdly, doing an internship in the same office and in parallel to the study is a great advantage. It allows a better understanding of the phenomenon which is studied by working directly with it, speaking with people who are experts in the domain, seeing direct consequences and implications. It also allows detecting in a more concrete way challenges, problems and limits related to this domain, and, in this way, to be more aware of what is working in the domain of snow avalanches. This domain involves a certain responsibility and a good understanding to detect complexities of snow avalanches. This is also why it is a so fascinating research topic.

VII. REFERENCES

Albertini P. (2000). La responsabilité des élus locaux : nécessité et aberrations. *Pouvoirs* (92). 103 – 116.

Ancey C., Sergent C., Martin E. (2003). Chap.3 Les métamorphoses de la neige, propriétés physiques et mécaniques. In *Guide Neige et Avalanches : connaissances, pratique, sécurité*. 3° Edition. Available at www.toraval.fr/livre/guide.php.

Ancey C., Gardelle F. &C., Zuanon J.-P. (2003). Chap. 1 L'homme face à la neige et aux avalanches dans les temps passés. In *Guide Neige et Avalanches : connaissances, pratique, sécurité*. 3° Edition. Available at www.toraval.fr/livre/guide.php.

Ancey C. (2011). Histoire et controverses de l'ingénierie des risques en montagne. [Introduction au cours de sociologie « controverses » donné par Valérie November au printemps 2001]. Disponible sur www.securiteavalanches-argentiere74.org/IMG/pdf/histoire-ingenierie.

Angillieri M. Y. E. (2010). Application of frequency ratio and logistic regression to active rock glacier occurrence in the Andes of San Juan, Argentina. *Geomorphology* (114). 396 – 405.

Arlot S., Celisse A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys (4). 40-79. DOI : 10.1214/09-SS054.

Bellot H., Bouvet, F.N. (2010). Les capteurs de transport de neige par le vent au banc d'essai. *Sciences Eaux et Territoires* (2). 66-77.

Bolognesi R. (1991). *L'analyse spatiale des risques d'avalanches. Premiers développements d'un environnement informatique d'aide à la décision*. (Thèse de doctorat). Cemagref – division Nivologie – Laboratoire de la Montagne Alpine. Université Joseph Fourier, Grenoble I.

Bolognesi R. (1996). The Driftometer. In *Proc. of the Int. Snow and Science Workshop, Banff*. 144 – 148.

Bolognesi R. (1999). *Modèle de metaraisonnement: Application à la prévision de phénomènes catastrophiques*. (Thèse de doctorat N°1959). Ecole Polytechnique Fédérale de Lausanne (EPFL). Département d'informatique. Laboratoire d'intelligence artificielle.

Bolognesi, R. (2013). *Estimer et limiter le risque avalanche*. Sion : Le vent des cimes.

Bolognesi, R. (2015). Blizzard et avalanches. *Meteo magazine: le blizzard*. (14). 36 – 39.

Boyne H.S., Williams K. (1992). Analysis of avalanche prediction from meteorological data at Berthoud pass, Colorado. *Proceedings* : International Snow Science Workshop. Breckenridge, Colorado, USA.

Castillo-Rivera M., Kobelkowsky A., Chavez M. (2000). Feeding biology of the flatfish Citharichthys spilopterus (Bothidae) in a tropical estuary of Mexico. *Journal of Applied Ichthyology* (16) 2. 73 – 78.

Chritin V., Bolognesi R., Gubler H. (1999). FlowCapt: a new acoustic sensor to measure snowdrift and wind velocity for avalanche forecasting. *Cold Regions Science and Technology.* (30). 125—133.

Chritin V., Melly T. (1998). *Acoustic sensor to measure wind velocity and snowdrift for snow avalanche forecasting*. No.LEMA-CONF-1998-028.144-145.

Descamps P. (2005). L'avalanche de la crête du Lauzet : la mécanique d'un lynchage médiatique. *Les cahiers du journalisme.* (14). 122 – 139.

Dümbgen L., Igl B.-W., Munk A. (2008) P-values for classification. *Electron. J. Statist.* (2). 468--493. doi:10.1214/08-EJS245.

Eckerstorfer M., Christiansen H.H. (2011). Relating meteorological variables to the natural slab avalanche regime in High Arctic Svalbard. *Cold Regions Science and Technology*. (69).184-193.

Eckerstorfer M., Christiansen H.H. (2011). Topographical and meteorological control on snow avalanching in the Longyearbyen area, central Svalbard 2006–2009. *Geomorphology.* (134). 186-196.

Emery Mayor D. (2008, 6 novembre). 100% soleil, info ou intox ?. *Sixième Dimension*. Disponible sur : http://sixieme-dimension.ch/2008/11/06/100-pour-cent-soleil.

Emery Mayor D. (2009, 16 septembre). Crans-Montana, regard sur plus de cent ans d'histoire. *Sixième Dimension*. Disponible sur : http://sixieme-dimension.ch/2009/09/16/crans-montana-regard-sur-plus-de-cent-ans-d-histoire.

Erard N. (2007). METEORISK : Services [Webpage]. Available on http://www.meteorisk.com/.

Estienne P. (1951). Les avalanches des 20 et 21 janvier dans les Alpes suisses, autrichiennes et italiennes. *Revue de Géographie Alpine*. (39) 2. 381-392.

Floyer J.A. (2003). *Statistical avalanche forecasting using meteorological data*. (master thesis). University of British Columbia. Faculty of Graduate Studies. Department of Geography.

Floyer J.A., McClung M.D. (2003). Numerical avalanche prediction: Bear Pass, British Columbia, Canada. *Cold Regions Science and Technology.* (37). 333 – 342.

Fromm, R., Department Natural Hazards and Alpine Timberline, Federal Research and Training Center for Forest, Natural Hazards and Landscape, Austria. (2009). Estimating the forecasting success of artificially triggering of avalanches with the combination of cluster and discriminant analysis. *Proceedings.* International Snow Science Workshop. Davos.

Gassner M., Brabec B. (2002). Nearest neighbour models for local avalanche forecasting. *Natural Hazards and Earth System Sciences.* (2). 247-253.

Geoffrey J. McLachlan (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Brisbane: John Wiley and sons, inc.

Grosjean M. (2012, 10-17). Introduction: from the director. [Webpage]. Available on http://www.climatestudies.unibe.ch/introduction/index_en.html.

Henzen, W., Schönbächler D., Bolognesi R. et al. – Service des Forêts et du Paysage du canton du Valais (2009). *Avalanches ! Les événements de février 1999.* Sierre. ITERAMA.

Johnson R.A. and Wichern D.W. (2007). *Applied Multivariate Statistical Analysis* (Sixth ed.) Upper Saddle River: Pearson International Edition.

Jomelli V. et al. (2007). Probabilistic analysis of recent snow avalanche activity and weather in the French Alps. *Cold Regions Science and Technology.* (47). 180 – 192.

Keller R. P., Drake J. M., Lodge D. M. (2007). Fecundity as a Basis for Risk Assessment of Nonindigenous Freshwater Molluscs. *Conservation Biology.* (21). 191 – 200.

Marcel J. (1970). Note sur l'hiver remarquable 1969-1970 dans les Alpes françaises. *Revue de Géographie Alpine.* (58) 3. 505-513.

McCollister C.M., Birkeland K., Hansen K., Aspinall R., Comey R. (2002). A probabilistic technique for exploring multi-scale patterns in historical avalanche data by combining GIS and meteorological nearest neighbors with an example from the Jackson Hole Ski Area, Whyoming. *Proceedings of International Snow Science Workshop 2002*. Penticton. BC. Canada. 109-116.

Pahaut E.,Bolognesi R. (2003). Chap.7 : Prévision régionale et locale du risque d'avalanches. In *Guide Neige et Avalanches : connaissances, pratique, sécurité*. 3° Edition. Available at www.toraval.fr/livre/guide.php.

Raymond B. (1958). Un volumineux et précieux recueil de météorologie alpine. *Revue de Géographie Alpine.* (46) 1. 213-218.

Ren S., Schultz T. W. (2002). Identifying the mechanism of aquatic toxicity of selected compounds by hydrophobicity and electrophilicity descriptors. *Toxicology Letters.* (129). 151 – 160.

Rencher A.C. (1995). *Methods of Multivariate Analysis*. New York: Johhn Wiley and Sons.

Reyt M.P. (2000). La représentation du risque dans l'imaginaire des altitudes. *Revue de Géographie Alpine.* (88) 4. 35-46.

Rheinberger, C.M. (2012). Learning from the past: statistical performance measures for avalanche warning services. *Natural Hazards.* (65). 1519-1533.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge.

Ripley B.D. (2002) *Statistical data Mining*. Springer-Verlag, New York. Available online at http://www.stats.ox.ac.uk/pub/bdr/SDM2002/DM2002.pdf.

Rougier H. (1975). Les chutes de neige et les avalanches du printemps 1975 dans les Grisons (Suisse). *Revue de Géographie Alpine*. (63) 4. 555-560.

Saemundsson T., Petursson H.G., Decaulne A. (2003). Triggering factors for rapid mass movements in Iceland. *Debris-Flow Hazards Mitigation: Mechanics, Prediction, and Assessment*. Rickenmann & Chen (eds). 167-178.

Singh A., Srinivasan K., Ganju A. (2005). Avalanche forecast using numerical weather prediction in Indian Himalaya. *Cold Regions Science and Technology.* (43). 83-92.

Sweet S.A. and Grace-Martin K. (1999). Chapter 8: Multivariate Analysis with Logistic Regression. In *Data Analysis with SPSS: A First Course in Applied Statistics*. Second edition. Available on http://www.colorado.edu/ibs/pop/jyoung/socy3301/assignments/week11.pdf .

Thirumuruganathan S. (2010, 17th May) A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm [Webpage]. Available on: https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/ (consulted on th 21st of April 2015).

Tribunal cantonal du Valais (2006). Avalanche d'Evolène : Le Tribunal cantonal confirme la condamnation d'André Georges et de Pierre-Henri Pralong. [Communiqué de presse]. Sion.

Villecrose J. (2001). Les avalanches de janvier et février 1999 dans les Alpes du nord françaises. *La Météorologie 8ème série.* (32). 11 – 22.

Zuanon J.P. (1998). Libres propos sur les avalanches. *Revue de Géographie Alpine.* (86) 1). 89 – 91.

VIII. APPENDICES

Appendix 1: R-code for statistical analyses

```
## Preparation of the database
##_____________________________

data.tot = read.csv("wholedataset.csv", header=TRUE, sep=";",
na.strings=c("?","/"))
data.red=data.tot[rowSums(is.na(data.tot.red))==0,] # delete all rows with at least
one missing value

## DA on standardized data
##_____________________________


# 1) standardize the data

standardize.data <- function(X){
if(is.null(nrow(X))) return( (X-mean(X, na.rm=TRUE))/(var(X, na.rm=TRUE))^1/2 )
}
RS=standardize.data(data.tot[,3])
N=standardize.data(data.tot[,4])
MVS=standardize.data(data.tot[,5])
Tn=standardize.data(data.tot[,6])
Ps=standardize.data(data.tot[,7])
Ta=standardize.data(data.tot[,8])
R24=standardize.data(data.tot[,9])
ADP=standardize.data(data.tot[,10])
ADN2O=standardize.data(data.tot[,11])
dTa=standardize.data(data.tot[,12])
tHs24=standardize.data(data.tot[,13])
vV=standardize.data(data.tot[,16])
ID=standardize.data(data.tot[,17])
HN24=standardize.data(data.tot[,18])
Hn3J=standardize.data(data.tot[,19])
AVAL=data.tot[,20]

data.std=as.data.frame(cbind(RS,N,MVS,Tn,Ps,Ta,R24,ADP,ADN2O,dTa,tHs24,vV,ID,HN24,
HN3J, AVAL))
data.red.std=data.std[rowSums(is.na(data.std))==0,] # delete all rows with at least
one missing value

# 2) LDA on standardized data


library(MASS)
DA.2=lda(AVAL ~ RS+N+MVS+Tn+Ps+Ta+R24+ADP+ADN2O+dTa+tHs24+vV+ID+HN24+Hn3J,
data=data.red.std, CV=TRUE)
table(DA.2$class, data.red.std[,16])
R=mean(DA.2$class != data.red.std[,16], na.rm=TRUE)
Good.classification = (1-R)*100
Good.classification

## Logistic Regression with CV
## _____________________________


library(boot)
require(boot)
# 1) find important variables
glm1 = glm(AVAL ~ RS+N+MVS+Tn+Ps+Ta+R24+ADP+ADN2O+dTa+tHs24+vV+ID+HN24+Hn3J,
data=data.red, family=binomial(link=logit))
summary(glm1)
head(data.red)
# 2) cross-validation with the model of Lutz:
source("LD.log.regr.cv.R")
n0 <- 983
p <- 15
79
X = as.matrix(data.red[,c(4,5,8,9,10,15,16,17)])
Y <- data.red[,18]
T1=LD.log.regr.cv(X,Y)
T1
risk1 <- (T1[1,2] + T1[2,1])/sum(T1)
```

```
risk1
risk=100-(risk1*100)
risk

## p-values for classification
## ____________________________


library("pvclass")
pv1=cvpvs.logreg(data.red[,c(3:17)], data.red[,18], find.tau=FALSE, tau.o=1)
analyze.pvs(pv1,Y = data.red[,18], alpha = 0.2, roc = TRUE, pvplot = TRUE, cex = 1)
# same code is run for alpha = 0.1 and alpha = 0.2.

## cross-validated kNN with the function of Lutz Dümbgen
## ____________________________________________________


source("LD.knn.cv.R")
n0 <- 983
p <- 15
X = as.matrix(data.red[,c(3:17)])
Y <- data.red[,18]
LD.knn.cv(X,Y,k=5) -> res1 # same code is run for k=3 and k=1.
res1
risk1 <- (res1[2,1] + res1[1,2])/sum(res1)
risk1
risk=100-(risk1*100)
risk
```

Appendix 2: Cross validated logistic regression function of Lutz Dümbgen

```
LD.log.regr.cv <- function(X,Y)
# X : data matrix of covariables,
# Y : 0-1-response vector,

{
n <- length(Y)
Y.hat <- rep(NA,n)
for (i in 1:n)
{
Xi <- X[-i,]
Yi <- Y[-i]
res <- glm(Yi ~ Xi, family="binomial")$coefficients
Y.hat[i] <- (sum(res*c(1,X[i,])) > 0)
}
Table <- table(Y,Y.hat)
dimnames(Table)[[1]] <- c("0","1")
dimnames(Table)[[2]] <- c("0","1")
return(Table)
   }
```

Appendix 3: Cross validated kNN function of Lutz Dümbgen

```
LD.knn.cv <- function(X,Y,k=1)
# X : data matrix of covariables,
# Y : 0-1-response vector,
# k : number of nearest neighbors.

{
n <- length(Y)
Y.hat <- rep(NA,n)
for (i in 1:n)
{
Xi <- X[-i,]
Yi <- Y[-i]
mu.hat <- colMeans(Xi)
sigma.hat <- apply(Xi,2,sd)
Xi <- t((t(Xi) - mu.hat)/sigma.hat)
Y.hat[i] <- knn(Xi,(X[i,] - mu.hat)/sigma.hat,Yi,k=k)
}
Table <- table(Y,Y.hat)
dimnames(Table)[[1]] <- c("0","1")
dimnames(Table)[[2]] <- c("0","1")
return(Table)
   }
```

Appendix 4: Declaration

Declaration

under Art. 28 Para. 2 RSL 05

Last, first name: Saugy Augustine

Matriculation number: 10-404-770

Programme: Climate Sciences

Bachelor ☐ Master ☐ Dissertation ☐

Thesis title: Statistical forecasting of snow avalanches situations using field measurements

Thesis supervisor: Prof. Dr. Lutz Dümbgen

Dr. Robert Bolognesi

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person, except where due acknowledgement has been made in the text. In accordance with academic rules and ethical conduct, I have fully cited and referenced all material and results that are not original to this work. I am well aware of the fact that, on the basis of Article 36 Paragraph 1 Letter o of the University Law of 5 September 1996, the Senate is entitled to deny the title awarded on the basis of this work if proven otherwise. I grant inspection of my thesis.

.....

Place, date

.....

Signature