Application of nonparametric pattern recognition algorithms in the characterization and forecasting of geophysical events

Candidata:
Dott.ssa Laura Sandri

Tutore:
Chiar.mo Prof. Warner Marzocchi

Coordinatore:
Chiar.mo Prof. Maurizio Bonafede

Bologna, Marzo 2003
Ad Azzo.

Tanto sappiamo tutti che sei da qualche parte.
Those who can, do; those who can't, simulate.

In a forest a fox bumps into a little rabbit, and says, "Hi, Junior, what are you up to?" "I’m writing a dissertation on how rabbits eat foxes," said the rabbit. "Come now, friend rabbit, you know that’s impossible! No one will publish such rubbish!" "Well, follow me and I’ll show you." They both go into the rabbit’s dwelling and after a while the rabbit emerges with a satisfied expression on his face.

Comes along a wolf. "Hello, little buddy, what are we doing these days?" "I’m writing the 2’nd chapter of my thesis, on how rabbits devour wolves." "Are you crazy? Where’s your academic honesty?" "Come with me and I’ll show you." As before, the rabbit comes out with a satisfied look on his face and a diploma in his paw. Finally, the camera pans into the rabbit’s cave and, as everybody should have guessed by now, we see a mean-looking, huge lion, sitting, picking his teeth and belching, next to some furry, bloody remnants of the wolf and the fox.

The moral: It’s not the content of your thesis that is important — it’s your PhD advisor that really counts.

Laura Mikla: small but powerful she is.

(David Jón Fuller)
Contents

1 Introduction .......................................................... 11

2 The Pattern Recognition technique ................................. 15
   2.1 Logic and Statistical PR .................................... 16

3 Testing the performance of some nonparametric Pattern
   Recognition algorithms in realistic cases ...................... 19
   3.1 Summary of the chapter ..................................... 19
   3.2 Introduction .................................................. 19
   3.3 SPR algorithms ................................................ 21
   3.4 Checking the performances of the SPR .................... 23
      3.4.1 Generation of synthetic data .......................... 24
      3.4.2 Application of SPR codes ............................. 32
      3.4.3 Classification Results ................................. 34
   3.5 Discussion of the results .................................. 35
      3.5.1 Comparison of the 4 SPR techniques in the case
            of uncorrelated and not binned datasets .......... 35
      3.5.2 Correlation and binning of the data ................ 38
      3.5.3 Voting of the learning dataset: overfit ............ 38
      3.5.4 KNN: influence of the K-value used ................ 40
      3.5.5 Case of different numbers of learning objects be-
            tween the two classes ................................. 40
   3.6 Final remarks of the chapter ............................... 42

4 Pattern Recognition applied to a dataset of worldwide
   dataset of seismic swarms recorded in volcanic areas ...... 45
5 Recognition of possible recurrent patterns in seismicity preceding recent flank eruptions at Mount Etna volcano

6 On the validation of earthquake forecasting models: the case of Pattern Recognition algorithms
6.6 Pattern Recognition applied to earthquake forecasting:
   M8 and CN algorithms .......................... 101
   6.6.1 Algorithm CN ............................. 103
   6.6.2 Algorithm M8 ............................. 105
   6.6.3 Test of PEFM based on M8 ............... 106
   6.7 Final remarks of the chapter ................ 111

7 Final remarks 113

8 References 117

A K-Nearest Neighbors 129

B Binary Decision Tree 131

C Fisher’s Discriminant Analysis 133

D Linear Discriminant Analysis 137

E Branch-And-Bound Technique 139
Chapter 1

Introduction

Geophysics is a scientific discipline aiming to describe quantitatively the natural phenomena occurring on the Earth. The main characteristic of the physical processes considered in Geophysics is the presence of a high number of degrees of freedom. The relative weights of such degrees of freedom, and the types of relationship among them, are of prominent importance in determining the nature of the process itself. For instance, if a few degrees of freedom play a predominant role in the process, with linear relationships, we can describe the process by means of classical deterministic equations. For example, in describing the trajectory of a bullet shot by a gun, we can approximate the process through a deterministic Newtonian model, neglecting an infinite number of less important degrees of freedom (defects in the bullet shape, air friction, etc...). Instead, if a few degrees of freedom play a predominant role in the process, but the relationship among them are nonlinear, the system might show a chaotic behavior due to the high sensitivity to the unavoidable measurement errors on the initial conditions. In this case, the system will rapidly evolve in a chaotic way, hardly distinguishable from “random noise”. Finally, if there are many variables having a considerable influence on the process, the system might be statistic or complex, depending on the type of relationship among the variables. In the former case, the system can be usually described by Gaussian statistic, while in the latter case power laws are more appropriate.
The empirical approach allows to achieve important information both on the relative weight of the degrees of freedom and on the type of relationship among them. This fundamental step needs to be carried out by means of suitable techniques. In particular, they should be able to deal simultaneously with many parameters. In this Ph.D. thesis, I use a multivariate statistical approach to characterize some relevant Geophysical processes. The empirical methods I have chosen are multivariate analysis techniques belonging to the class of Pattern Recognition algorithms. The potentiality of this type of analysis is due to its ability in identifying possible repetitive schemes \textit{(patterns)} among objects belonging to distinct categories, by extracting information from any possible combination (linear or not) of variables that are supposed to have an influence on the process (see chapter 2).

The intrinsic peculiarities of Pattern Recognition have led to successful applications in several and diverse research fields characterized by large, normally distributed datasets, such as waveform analyses in engineering, brain modeling in biology and psychology, stock market behavior in economy, medical records in medicine, just to mention a few examples (Duda and Hart, 1973; Fukunaga, 1990). Nevertheless, so far the applications of Pattern Recognition in Geophysics have been numerically scarce. The only remarkable efforts in this direction are M8 and CN algorithms for earthquake prediction (Keilis-Borok and Kossobokov, 1990; Keilis-Borok \textit{et al.}, 1988), and very few applications to volcanology (Mulargia \textit{et al.}, 1991; Mulargia \textit{et al.}, 1992; Vinciguerra \textit{et al.}, 2001).

Remarkably, in Geophysics, the most employed Pattern Recognition algorithm is not commonly used in the other scientific fields.

Generally, in fact, there is an important obstacle in the application of Pattern Recognition to Geophysics, with respect to the other research fields mentioned above. In fact, the data available in Geophysics often display large differences if compared to the datasets of other disciplines. First of all, we often deal with few data, often strongly correlated and/or discretized. Most of the times, the
data cannot be assumed to be normally distributed. Rather, they follow other probability density functions \((pdf)\), such as the poissonian, the binomial or the uniform. Finally, as mentioned above, we rarely know which are the variables playing a predominant role in the processes studied. As a consequence, very often we need to include in the datasets all the variables that we are able to measure, thus introducing a noise in the data due to all those variables that, in reality, are not important.

Until now, very few efforts have been made in the past to evaluate the performance of the available techniques in analyzing this kind of data. Certainly, this is a main point. For this reason, a large part of this Ph.D. work is devoted to test the behavior of some of the most common Pattern Recognition algorithms. In particular, before applying Pattern Recognition to some Geophysical problems, I implement 4 statistical nonparametric Pattern Recognition algorithms widely used in other scientific research fields, named \textit{K-Nearest Neighbors, Binary Decision Tree, Fisher's Analysis} and \textit{Linear Discriminant Analysis} (for their description, please refer respectively to the appendices A, B, C and D). I test their performance on synthetic datasets that I simulate. The synthetic datasets are generated in order to realistically mimic the most common dataset properties that we face in Geophysics, such as those mentioned above. The dataset generation, the strategy of simulation and the results obtained are discussed in chapter 3. As shown in that chapter, the simulations allow to identify two algorithms, Fisher's Analysis and Binary Decision Tree, as the most suitable to be applied to datasets with the typical properties of Geophysics datasets.

Successively, I apply these two algorithms to analyze two real sets of volcanological data, in order to characterize the behavior of some variables, mainly seismicity, before a volcanic eruption. A relevant portion of the research topics in Geophysics is related to the study of the occurrence of some disastrous events, such as earthquakes and volcanic eruptions. For obvious reasons, the problem of the predictability
of these devastating phenomena has both a scientific and a social relevance. In this sense, I think that the application of Pattern Recognition to characterize the phenomena preceding a volcanic eruption can greatly help in the comprehension of the mechanisms that promote or unfavor the occurrence of these catastrophic events. In particular, in chapter 4 I analyze a dataset of seismic swarms recorded during unrests in different volcanic areas of the world. Since socially dangerous volcanoes, such as Vesuvius, have generally a low eruptive frequency, the datasets relative to these volcanoes are usually very scarce. Thus, the goal of this application is to find possible repetitive and common patterns among pre-eruptive seismicity, if compared to isolated seismicity, in several volcanic areas. The ultimate idea is to use these patterns, common to different volcanoes, as a primitive forecasting rule for those volcanoes, such as Vesuvius, for whom there are too few quantitative data available about past unrests. In chapter 5 I apply again the two Pattern Recognition algorithms to a dataset of seismic clusters recorded in the Etnan area in recent decades (from 1974 to 2001). I also analyze the same dataset by means of another multivariate statistical technique, i.e. a multivariate regression fit. The goal of this second application is to identify a possible link, suggested by previous works on this area, between the tectonic stress regime and the occurrence of flank eruptions at Mount Etna volcano.

Finally, in chapter 6, I critically review the two most used Pattern Recognition algorithms applied in Geophysics, called M8 and CN. These two codes are still currently used to predict relevant seismic events. I propose a statistical formal procedure to validate a generic prediction algorithm, and apply it to specifically estimate the statistical significance of the only forward prediction experiment currently conducted to forecast earthquakes with magnitude $M \geq 7.5$, based on M8 algorithm.

The final remarks of this Ph.D. study are summarized in chapter 7.
Chapter 2

The Pattern Recognition technique

Pattern Recognition is a very powerful multivariate analysis technique allowing, in principle, the identification of possible repetitive schemes among the objects belonging to distinct categories. Whilst usual data analysis takes into account only one variable of the process at a time, PR is able to extract information from any possible combination (linear or not) of variables that are supposed to have an influence on the process. Moreover, PR does not need the construction of a theoretical model, but it is usually based on a basic and sole hypothesis, i.e., the assumption that the phenomenon under study is governed by a finite number of complex, but repetitive patterns of the variables.

These appealing features led PR to be applied with success in several and diverse disciplines which share the study of complex systems (e.g. Fukunaga, 1990). For this reason, in spite of the scarce results obtained so far in geophysics (partially discussed in chapter 6), we believe that PR might be very useful also in this discipline.

From a technical point of view, the main goal of PR is to classify objects. Every object is represented by an array of qualitative or quantitative variables. The procedure of analysis consists of 3 different steps: the learning phase, the voting phase, and the control experiments. In the learning phase, a set of known and classified objects is analyzed in order to recognize all the possible patterns that char-
acterize each category, i.e., the combination of variables that allow to
discriminate the objects belonging to different classes. This step turns
out to be very useful also from a theoretical point of view, since it al-
 lows to recognize, among all the variables, the ones that really count
to discriminate the classes.

In the voting phase, the patterns identified during the learning
are used to classify other new objects, whose category is unknown.
Finally, the control experiments allow to check the stability of the
results by repeating the learning and the voting phase on independent
data sets.

2.1 Logic and Statistical PR

PR may be roughly divided into two branches: logic (LPR) and sta-
tistical (SPR). In LPR, the arrays of variables describing the objects
consist of a series of "0" or "1", often corresponding to "NO" or "YES"
answers to a questionnaire (Mulargia et al., 1992). Thus, LPR is fea-
sible and ready-to-use when dealing with qualitative or categorical
data; instead, the presence of quantitative data requires to subdivide
the measured range of each variable, for example into two subintervals,
so that a given value is then represented by a "1" ("LARGE") if
the value is contained in the upper subinterval, and "0" ("SMALL")
if it is contained in the lower subinterval. One of the most commonly
used LPR technique is the CORA-\textit{n} family. Calling \textit{trait} an array
consisting of values of \textit{n} variables, CORA-\textit{n} algorithm identifies the
relatively frequent traits in the objects belonging to a specific category
that are relatively infrequent among the objects of all the other cat-
egories. Such traits are often called \textit{characteristic features} (or simply
\textit{features}) of that category. An object is then classified as belonging
to a category if it owns at least a certain number of characteristic
features of that category, and much less of any other.

The main advantage of such technique is the potential large appli-
cability to many different cases. Its main disadvantage, instead, is the
large number of degrees of freedom of the LPR model. For instance, the choice of the kind of discretization of the continuous variables, the critical difference in the number of traits belonging to distinct classes necessary to classify an object, and the dimensions of the traits are only few examples of the large number of subjective choices to be done.

In SPR, each object is represented by an array consisting of the measured values (continuous or discrete) of each variable. The classification of the objects is made by setting up rules that are mainly based on probabilistic and mathematical considerations. If the statistical distribution of the values of the variable in each class were known, the recognition would be easy. Unfortunately, this is never the case in practice, so nonparametric techniques have been developed, such as K-Nearest Neighbors (KNN), Fisher’s Analysis (FIS), Linear or Quadratic Discriminant Analysis (LDA and QDA), and Binary Decision Tree (BDT).

A detailed description of the different PR codes is not a goal of the present chapter. For such information, the reader is remanded to the Appendices A, B, C and D.

All the LPR and SPR techniques so far proposed have their merits and weakness. As a matter of fact, however, the “best” technique does not exist, because any procedure can be the best one depending on the problem under study. Unfortunately, very few efforts were made in the past to evaluate the performance of the methods so far proposed in analyzing different kinds of data. This limits drastically a critical and complete evaluation of the efficiency of the PR techniques in general.

Another issue that must be reminded is that PR always gives a result. Yet, it is necessary to evaluate how good the result is, that is its statistical reliability.
Chapter 3

Testing the performance of some nonparametric Pattern Recognition algorithms in realistic cases

3.1 Summary of the chapter

The success obtained by Statistical Pattern Recognition in many disciplines is certainly related to the quality and availability of many data, normally distributed. However, in other disciplines, the datasets consist of few measurements, often binned, correlated, and not normally distributed. Usually, we do not even know which features have an influence on the process. The main goal of this chapter is to evaluate the performance of some nonparametric Pattern Recognition algorithms when applied to such data. The Pattern Recognition algorithms tested have been selected because of their simplicity and because they are based on different approaches.

3.2 Introduction

Statistical Pattern Recognition (SPR) is a very powerful multivariate analysis technique that allows, in principle, the identification of possible repetitive schemes among objects belonging to distinct categories. The properties of SPR make it a very appealing tool in the study
of complex and/or stochastic systems, where the classical theoretical modeling usually fails. Until now, SPR techniques have been successfully applied to many scientific problems of this type, such as waveform analysis in engineering, brain modeling in biology and psychology, stock market behavior in economy, medical records in medicine, just to mention a few examples (Duda & Hart, 1973; Fukunaga, 1991 and references therein). The success achieved in these cases is certainly linked to the availability of large datasets (e.g. Hattori & Takahashi, 2000; Hattori & Takahashi, 1999; Poston & Marchette, 1998), usually following a normal distribution (e.g. Lotlikar & Kothari, 2000).

Unfortunately, in many complex systems, as in Earth sciences, the available datasets do not have these “nice” characteristics. First of all, the data are often very few, as, for instance, in the study of the occurrence of large earthquakes and volcanic eruptions. Secondly, the statistical distributions of the data are often different from the Gaussian distribution, being the datasets usually binned and/or asymmetric. Also, the data might be characterized by correlations between some of the features. Last, but not least, the datasets usually consist of a certain number of measurements for each of several features, but often it is not clear which ones of these features have a real influence on the process under study. In other words, often real datasets include both useful and irrelevant features. In such cases, one of the main goals is to recognize which is the subset of features (often called optimal, see Raudys & Jain 1991; Jain & Waller, 1978; Raudys & Pikelis, 1980; Raudys, 1979) actually playing a predominant role in the process.

As a consequence, in order to apply profitably SPR techniques to the study of such complex systems, we need to check the general performance of SPR algorithms when applied to this “not-so-nice” kind of dataset. In the past, many papers have investigated on the problem of small datasets (e.g. Raudys & Jain, 1991; Raudys & Pikelis, 1980; Raudys, 1979; Skurichina & Duin, 1998 and references therein), showing that the scarce availability of data can easily contaminate the design and the performance of classifiers. In most of the cases (Raudys
& Jain, 1991; Raudys & Pikelis, 1980; Raudys, 1979), however, the
data were only normally distributed, or they came form a combination
of Gaussian data (Skurichina & Duin, 1998). Here I focus the at-
tention on some aspects poorly investigated in the past. In particular, I
evaluate the discrimination ability of some SPR codes as a function of:

(i) the kind of statistical distribution governing the measurements;

(ii) the size of the dataset;

(iii) the correlation among some of the features;

(iv) the binning of the measurements.

Also, I test the potential ability of these SPR codes in extracting the
“optimal” subset of features from the data.

In the end, I also check how the performance of the algorithms
change if the number of learning objects is significantly different be-
tween the classes. This is a case that is often encountered in real
datasets.

In order to test SPR behavior, I simulate realistic datasets and
I apply to these synthetic data some of the most common SPR al-
gorithms. For simplicity, I explore the 2-class problem. As a sake
of clarity, due to the considerable number of acronyms and variable
names that will be introduced in the chapter, I refer the reader to
table 3.1 for a summary of their meaning.

3.3 SPR algorithms

For this study, I select four nonparametric SPR algorithms to be
tested: K-Nearest Neighbors, Fisher’s Analysis, Linear Discriminant
Analysis (respectively named KNN, FIS and LDA; see Duda & Hart,
1973; Fukunaga, 1991 and references therein) and Binary Decision
Tree (BDT, see Rounds, 1980; Mulargia et al., 1992). These codes
<table>
<thead>
<tr>
<th>SPR</th>
<th>Statistical Pattern Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>K-Nearest Neighbors</td>
</tr>
<tr>
<td>BDT</td>
<td>Binary Decision Tree</td>
</tr>
<tr>
<td>FIS</td>
<td>Fisher’s Analysis</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Classification Error</td>
</tr>
<tr>
<td>RV</td>
<td>Random Variable</td>
</tr>
<tr>
<td>G, E, B, U</td>
<td>Gaussian, Exponential, Binomial and Uniform distributions</td>
</tr>
<tr>
<td>$pdf$</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Kendall’s measure of correlation of RVs</td>
</tr>
<tr>
<td>$N_{lear}$</td>
<td>Number of learning objects used</td>
</tr>
<tr>
<td>$N_{fu}$</td>
<td>Number of features used</td>
</tr>
<tr>
<td>$K$</td>
<td>Number of neighbors used in KNN algorithm</td>
</tr>
<tr>
<td>$\gamma^{(V)}$</td>
<td>Mean $\gamma$ obtained on the voting dataset</td>
</tr>
<tr>
<td>$\gamma^{(L)}$</td>
<td>Mean $\gamma$ obtained on the learning dataset</td>
</tr>
<tr>
<td>$\gamma^{(V1)}$</td>
<td>Mean $\gamma$ obtained on the class1-objects of the voting dataset</td>
</tr>
<tr>
<td>$\gamma^{(V2)}$</td>
<td>Mean $\gamma$ obtained on the class2-objects of the voting dataset</td>
</tr>
<tr>
<td>$\gamma^{(L1)}$</td>
<td>Mean $\gamma$ obtained on the class1-objects of the learning dataset</td>
</tr>
<tr>
<td>$\gamma^{(L2)}$</td>
<td>Mean $\gamma$ obtained on the class2-objects of the learning dataset</td>
</tr>
</tbody>
</table>

Table 3.1: Summary of the acronyms and variable names used in the chapter.
have been selected because they are all based on different nonparametric approaches, implying very few and relaxed assumptions on the statistical distribution of the data. This makes them particularly flexible, and suitable to deal with datasets consisting of random variables (RVs) with non-Gaussian distribution.

Algorithm KNN (see appendix A) is a completely nonparametric method, because it doesn't need any assumption on the statistical distribution of the data and on the form of the discriminant functions. Algorithm BDT (see appendix B) is based on the assumption that the patterns in the data are hierarchically ordered. The discriminant analysis algorithms FIS and LDA (see appendices C and D) assume that the forms of the discriminant functions are known, and the data are used to estimate the parameters of the discriminant functions. Note that, in dealing with few real data, it is usually more difficult to make assumptions on the statistical distributions of the data than to assume a form of the discriminant functions. Here, I use only hyperplanes as discriminant functions, in order to keep as low as possible the number of parameters to be estimated from the data. This is a very important point in dealing with few data. A large number of parameters, as in the case of quadric discriminant functions (e.g. Sierra, 2002), increases the possibility to achieve an overfit of the data.

I use the stochastic approximation (Duda & Hart, 1973) method for the LDA algorithm, therefore, the discriminant rules are different from those obtained with FIS (see Duda & Hart, 1973).

3.4 Checking the performances of the SPR

The check of the performances of the nonparametric SPR algorithms can be divided into three main steps:

1. Generation of synthetic datasets (box 1 in figure 3.1).

2. Application of SPR codes to the generated synthetic dataset: performance of the learning and voting phases (boxes 2 and 3 in figure 3.1).
3. Evaluation of the algorithm performances through the classification error ($\gamma$), i.e., the number of voted data misclassified by the algorithm out of the total number of voted data (box 4 in figure 3.1).

I shall explain each of these steps in the following.

### 3.4.1 Generation of synthetic data

In order to check the performances of the SPR, I need to simulate synthetic datasets which mimic the basic features of a real dataset (see e.g. Skurichina & Duin, 1998). In particular, in Earth sciences we often deal with small datasets, composed by RVs with statistical distributions that can be drastically different from the usually assumed Gaussian distribution. For instance, we commonly face strongly asymmetric distributions. Moreover, we often deal with correlated and/or discrete RVs. Such discretization can be due either to the intrinsic discreteness of the RV (for instance, the number of events) or to the binning of the measurement, due to its finite precision (e.g. the measure of the earthquake magnitudes in Earth sciences, see Tinti & Mulargia, 1987). In order to be realistic, a synthetic dataset needs to retain all of these factors.

In the following, I describe in detail the main features of the synthetic datasets, i.e., the number of data (or objects) contained, the statistical distributions of each component of the objects, the possible correlation between the components, and their binning.

#### The objects

Each synthetic dataset I generate is composed of 11000 objects, half belonging to Class 1 and half to Class 2. In this way we can use up to 10000 objects for the learning phase, and still have 1000 new and independent data to be used for the voting phase. The $i$-th object is represented by a 10-component vector, $\vec{x}_i$. Considering $\vec{x}_i$, its $k$-th component should resemble the measure of the $k$-th physical feature, whose measurements follow one selected probability distribution.
Figure 3.1: Flow chart of the simulation strategy.
Probability distributions

I generate the objects according to the following probability distributions: Gaussian (G), Exponential (E), Uniform (U) or Binomial (B), because they are certainly the most diffuse in nature. For example, they describe satisfactorily different kinds of processes such as the Poisson process (E), or they explain random fluctuations and errors on the measure of a quantity around its mean value (G). Moreover, they cover a range of different properties usually contained in experimental samples and that may affect the performance of the algorithms: some are continue (G, U, and E) and one is discrete (B); some are symmetric (G, U) and others not or not necessarily (E, B). For a definition of the parameters describing such distributions, see Kalbfleisch, 1985.

Important and irrelevant features

In the application of SPR to Earth science, often the researcher doesn’t know which are the physical features having a real influence on the process under study (I call these features “important”). In such cases, the usual procedure is to take into account all the features “suspected” to be important in the process, hoping that the “irrelevant” ones will be (a) excluded automatically by the algorithm (e.g. BDT), or (b) distinguished from the “important” ones with the help of some specific techniques (for KNN, LDA and FIS), for example by forward or backward selection (Fukunaga, 1991), or by some branch-and-bound techniques (Fukunaga, 1991; Raudys & Jain, 1991).

Both (a) and (b) implicitly assume “monotonicity” in the classifier (see Fukunaga, 1991), that is they rely on the assumption that the presence of “irrelevant” features in the analysis does not alter drastically the ability of the algorithms to recognize the real patterns in the data; in other words, they are based on the assumption that the “noise” introduced in the analysis by the “irrelevant” features does not deteriorate the “signal” given by the “important” ones. Here, I test the validity of this assumption for the nonparametric SPR codes used.
In practice, in the synthetic objects that I generate, only a few components are “important”. This is realized by setting up different parameters of the probability distributions in the two classes only for the “important” features. The other features, being governed by exactly the same probability distribution in the two classes, do not allow any discrimination among the objects, and, thus, they are “irrelevant”.

**Explored cases**

I explore two possible cases:

**Case GEB.** In this case only the first 3 components of each object, distributed respectively as G, E and B, are “important”. The other 7 components are “irrelevant”. The parameters of each distribution in each class are shown in table 3.2.

**Case EU** In this case only the first 2 components of each object, distributed respectively as E and U, are “important”. The other 8 components are “irrelevant”. The parameters of each distribution in each class are shown in table 3.3.

In each case, it is clear that the parameters of the distributions of the “important” features in the two classes have a great importance on the discrimination results. For example, if an “important” feature follows a G distribution with parameters in the two classes (μ₁ and σ₁ for class 1, μ₂ and σ₂ for class 2) such that the overlapping is as in figure 3.2a, the discrimination will obviously be much more difficult and worse than in the case in which μ₁, σ₁, μ₂ and σ₂ give no overlapping (figure 3.2b). In order to have comparable overlapping between the two classes for each “important” feature, whatever distribution it follows, I define

\[ \eta = \frac{μ₁ - μ₂}{\sqrt{σ₁^2 + σ₂^2}} \]  

(3.1)

where μ₁, μ₂ are the means in the two classes, and σ₁, σ₂ are the standard deviations. The quantity \( \eta \) gives an idea of the overlapping between the distributions in the two classes: the smaller it is, the larger
<table>
<thead>
<tr>
<th>Type</th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>$\mu_1 = 0$, $\sigma^2_1 = 1$</td>
<td>$\mu_2 = 1.27$, $\sigma^2_2 = 1$</td>
</tr>
<tr>
<td>E</td>
<td>$\lambda_1 = 1$</td>
<td>$\lambda_2 = 10.4$</td>
</tr>
<tr>
<td></td>
<td>($\mu_1 = 1$, $\sigma^2_1 = 1$)</td>
<td>($\mu_2 = 10.4$, $\sigma^2_2 = 108.2$)</td>
</tr>
<tr>
<td>B</td>
<td>$p_1 = 0.30$</td>
<td>$p_2 = 0.83$</td>
</tr>
<tr>
<td></td>
<td>($\mu_1 = 0.3$, $\sigma^2_1 = 0.21$)</td>
<td>($\mu_2 = 0.83$, $\sigma^2_2 = 0.14$)</td>
</tr>
</tbody>
</table>

Table 3.2: Case GEB: parameters of the distributions of the "important" and "irrelevant" features fixed for the generation of synthetic datasets. The consequent values of the mean and variance of each distribution are given in brackets.
"Important" features

<table>
<thead>
<tr>
<th>Type</th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(E)</td>
<td>(\lambda_1 = 1) (\mu_1 = 1, \sigma_1^2 = 1)</td>
<td>(\lambda_2 = 10.4) (\mu_2 = 10.4, \sigma_2^2 = 108.2)</td>
</tr>
<tr>
<td>(U)</td>
<td>(a_{\text{min}} = 0, a_{\text{max}} = 1) (\mu_1 = 0.5, \sigma_1^2 = 0.08)</td>
<td>(a_{\text{min}} = 0.4, a_{\text{max}} = 1.3) (\mu_2 = 0.85, \sigma_2^2 = 0.0675)</td>
</tr>
</tbody>
</table>

"Irrelevant" features

<table>
<thead>
<tr>
<th>Type</th>
<th>(\forall) class</th>
</tr>
</thead>
<tbody>
<tr>
<td>(G)</td>
<td>(\mu = 0, \sigma^2 = 1)</td>
</tr>
<tr>
<td>(E)</td>
<td>(\lambda = 0.5) (\mu = 0.5, \sigma^2 = 0.25)</td>
</tr>
<tr>
<td>(B)</td>
<td>(p = 0.4) (\mu = 0.4, \sigma^2 = 0.24)</td>
</tr>
<tr>
<td>(U)</td>
<td>(a_{\text{min}} = 0.5, a_{\text{max}} = 2.5) (\mu = 1.5, \sigma^2 = 0.33)</td>
</tr>
<tr>
<td>(G)</td>
<td>(\mu = 1, \sigma^2 = 0.25)</td>
</tr>
<tr>
<td>(E)</td>
<td>(\lambda = 2) (\mu = 2, \sigma^2 = 4)</td>
</tr>
<tr>
<td>(B)</td>
<td>(p = 0.6) (\mu = 0.6, \sigma^2 = 10.24)</td>
</tr>
<tr>
<td>(U)</td>
<td>(a_{\text{min}} = 1, a_{\text{max}} = 3) (\mu = 2, \sigma^2 = 0.33)</td>
</tr>
</tbody>
</table>

Table 3.3: Case EU: parameters of the distributions of the "important" and "irrelevant" features fixed for the generation of synthetic datasets. The consequent values of the mean and variance of each distribution are given in brackets.
Figure 3.2: Overlapping between the classes corresponding to 3 different values of $\eta$ for Gaussian distribution: a-almost complete overlap, b-no overlap, c-intermediate overlap (this is the value I used in these simulations). In the figure I report, for each case, the mean and standard deviation used.
the overlapping is. Its meaning is similar to the Mahalanobis distance (Duda & Hart, 1973), although \( \eta \) can take into account different variances in the two classes. For the synthetic datasets, I fix \( \eta = 0.9 \), which guarantees a partial overlapping between the distributions in the two classes (see figure 3.2c for the G distribution).

**Correlation**

In these simulations, I take into account a possible correlation between the first two “important” features (for both cases GEB and EU). Let’s consider the objects represented by 2-dimensional vectors, whose two components are measurements of the first two “important” features, as RVs coming from a bivariate distribution.

A measure of the correlation, or association, between the RVs is given by the Kendall’s \( \tau \) coefficient (Gibbons, 1971):

\[
\tau = \pi_c - \pi_d
\]

where \( \pi_c \) and \( \pi_d \) are, respectively, the probabilities of concordance and discordance for any two independent pairs. In particular, selecting randomly from the sample two objects \( (\bar{x}_i, \bar{x}_j) \), represented by the RVs \( (x_{i,1}, x_{i,2}) \) and \( (x_{j,1}, x_{j,2}) \), we have that

\[
\pi_c = p[(x_{i,1} - x_{j,1})(x_{i,2} - x_{j,2}) > 0]
\]

\[
\pi_d = p[(x_{i,1} - x_{j,1})(x_{i,2} - x_{j,2}) < 0]
\]

Perfect association between the RVs (i.e. perfect accordance or discordance) produces \( |\tau| = 1 \) while completely uncorrelated RVs have \( \tau = 0 \).

In these simulations, for both cases GEB and EU, I generate 3 different types of datasets corresponding to 3 different values of the Kendall’s \( \tau \): \( \tau = 0 \) (no correlation), \( \tau = 0.3 \) (weak correlation) and \( \tau = 0.75 \) (strong correlation).

**Binning**

I explore the possible effects of the binning of the data on the performance of SPR by generating datasets with continuous (G, E, U) and
<table>
<thead>
<tr>
<th>&quot;Important&quot; features</th>
<th>Correlated &quot;Important&quot; features</th>
<th>Kendall’s ( \tau )</th>
<th>&quot;Irrelevant&quot; features</th>
<th>Binning (y/n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G,E,B</td>
<td>G,E</td>
<td>0.3</td>
<td>U,G,E,B,U,G,E</td>
<td>n</td>
</tr>
<tr>
<td>G,E,B</td>
<td>G,E</td>
<td>0.3</td>
<td>U,G,E,B,U,G,E</td>
<td>n</td>
</tr>
<tr>
<td>G,E,B</td>
<td>G,E</td>
<td>0.75</td>
<td>U,G,E,B,U,G,E</td>
<td>n</td>
</tr>
<tr>
<td>G,E,B</td>
<td>G,E</td>
<td>0.75</td>
<td>U,G,E,B,U,G,E</td>
<td>n</td>
</tr>
<tr>
<td>G,E,B</td>
<td>none</td>
<td>0</td>
<td>U,G,E,B,U,G,E</td>
<td>n</td>
</tr>
<tr>
<td>G,E,B</td>
<td>none</td>
<td>0</td>
<td>U,G,E,B,U,G,E</td>
<td>n</td>
</tr>
<tr>
<td>E,U</td>
<td>E,U</td>
<td>0.3</td>
<td>G,E,B,U,G,E,B,U</td>
<td>n</td>
</tr>
<tr>
<td>E,U</td>
<td>E,U</td>
<td>0.3</td>
<td>G,E,B,U,G,E,B,U</td>
<td>y</td>
</tr>
<tr>
<td>E,U</td>
<td>E,U</td>
<td>0.75</td>
<td>G,E,B,U,G,E,B,U</td>
<td>n</td>
</tr>
<tr>
<td>E,U</td>
<td>E,U</td>
<td>0.75</td>
<td>G,E,B,U,G,E,B,U</td>
<td>n</td>
</tr>
<tr>
<td>E,U</td>
<td>none</td>
<td>0</td>
<td>G,E,B,U,G,E,B,U</td>
<td>n</td>
</tr>
<tr>
<td>E,U</td>
<td>none</td>
<td>0</td>
<td>G,E,B,U,G,E,B,U</td>
<td>y</td>
</tr>
</tbody>
</table>

Table 3.4: Summary of the characteristics of the synthetic datasets: each line in the table corresponds to a type of dataset consisting of 11000 objects.

discrete (B) features, and successively applying the SPR algorithms both on these data and on their binned version. The latter is derived from the former by rounding up the values of the continuous features (G, E, U) to a numerable set of allowed values.

To sum up, in order to take into account possible correlations and binning in experimental data, for both cases GEB and EU I generate three kinds of datasets of 11000 objects each, corresponding to three different correlation degrees between the first two "important" features. From each dataset, a binned version is then derived. Table 3.4 reports a summary of all the explored cases.

### 3.4.2 Application of SPR codes

Following the scheme displayed in figure 3.1, each of the four SPR algorithms (KNN, BDT, FIS, LDA) is applied to each type of synthetic dataset shown in table 3.4. In order to have reliable estimates of
the behavior of SPR algorithms, for each of these cases I generate 1000 datasets, thus applying each SPR code 1000 times. Each time, I perform several learnings (box 2 in figure 3.1) on different numbers of data; after every learning, I perform the voting (box 3 in figure 3.1), based on the patterns identified in the learning.

Learning

First of all, I select the number of learning objects ($N_{lear}$). By performing several learning phases, each on a different number of learning objects, but performing the voting phase on the same dataset every time, I evaluate the performance of SPR algorithms as a function of the size of the learning dataset; as mentioned above, we are particularly interested in the cases with few learning data, i.e. $N_{lear} \leq 50$. In the calculation, I use 7 possible values of the learning dataset size: $N_{lear} = 20, 50, 100, 500, 1000, 5000, 10000$.

Secondly, for each $N_{lear}$, I select the number of features used ($N_{fu}$), that is I consider only the first $N_{fu}$ features. Therefore, $N_{fu}$ is the dimension of the vectors representing the objects considered in the analysis. I use 10 possible values: $N_{fu} = 1, 2, \ldots 10$. In this way I evaluate the performance of SPR algorithms as a function of the number of features used. This relationship gives useful hints on the validity of the monotonicity assumption (see section 3.4.1). Note that, when $N_{fu} = 2$ (in case EU) or $N_{fu} = 3$ (in case GEB) I am considering all of and only the “important” features. I refer to this particular subset of features as the “optimal” subset of features (Raudys & Jain, 1991; Jain & Waller, 1978).

For the KNN algorithm, I also explore the influence of the parameter $K$, because, in the finite sample-size case, it can significantly alter the performance of the classifier (Raudys & Jain, 1991). For this purpose, I apply KNN algorithm for 6 different values of $K$: $K = 1, 11, 51, 101, 501, 1001$. Clearly, I can try all of these 6 $K$-values only when 10000 learning objects are used, while in the other cases
the largest $K$-value used is the largest one satisfying the condition

$$K \leq \frac{N_{\text{test}}}{10}$$

In this way, for BDT, FIS and LDA algorithms I perform $7*10 = 70$ different learning phases at each simulation; for KNN, the learning phases are $7 * 10 * 6 = 420$. For each learning phase, I recognize the patterns among the $N_{\text{test}}$ learning data, and perform the voting.

Voting

In the voting phase, we attribute a class to the 1000 new and independent voting objects.

Because of the scarce number of available data, a common practice in many real application of SPR is to use the learning dataset both to select the discriminating rules and to estimate the classification error (e.g. Hattori & Takahashi, 1999). The same procedure is used here in order to quantify how much the voting of the same data used for the learning phase leads to more optimistic results, compared to the case in which a new and independent dataset is voted.

According to box 4 of figure 3.1, I evaluate the $\gamma$ for the voting ($\gamma^{(V)}$) and learning ($\gamma^{(L)}$) datasets. For both of these groups of data, the $\gamma$ is also computed separately on the data belonging to Class 1 ($\gamma^{(V1)}$ and $\gamma^{(L1)}$) and Class 2 ($\gamma^{(V2)}$ and $\gamma^{(L2)}$). The latter errors are computed in order to check if the objects belonging to a class are classified more easily than those of the other one.

3.4.3 Classification Results

After selecting the type of dataset, $N_{\text{test}}$ and $N_{\text{fu}}$, I apply the algorithms to 1000 synthetic datasets; then, I compute the mean $\gamma$, named $\overline{\gamma^{(*)}}$, and the standard deviation $\sigma_{\gamma^{(*)}}$, where $\bullet = V, L, V1, L1, V2, L2$.

In figures 3.3 ÷ 3.5 I show the results obtained in the simulations; in particular, for each of the 4 SPR algorithms, I examine:

1. the results obtained in classifying the voting dataset, in the most general case of not binned and uncorrelated data (figure 3.3);
2. the results obtained in classifying the voting dataset, in case of not binned but correlated data, and, vice versa, in case of uncorrelated but binned data;

3. the results obtained by comparing the classification errors in classifying the learning (instead of the voting) dataset, in the most general case of not binned, uncorrelated data (figure 3.4);

4. the difference in the classification results obtained for class 1 ($\gamma^{(V1)}$) and class 2 ($\gamma^{(V2)}$).

Finally I examine the influence of the $K$-value used in the application of KNN algorithm (figure 3.5).

In figures 3.3 ÷ 3.5, unless differently specified, the results displayed for the KNN technique refer to the classification performed by using $K = 1$, because it is the only possible value for $K$, whatever $N_{lear}$ is. For the FIS technique I never display results obtained for $N_{fu} = 1$ because, by definition, this technique works on vectors at least bidimensional.

## 3.5 Discussion of the results

### 3.5.1 Comparison of the 4 SPR techniques in the case of uncorrelated and not binned datasets

In order to compare the SPR algorithms in case of not binned and uncorrelated data, the $\gamma^{(V)}$ is displayed in figure 3.3 as a function of $N_{fu}$ and $N_{lear}$.

Let’s start from the general properties of the results shown in figure 3.3. First of all, plots for cases GEB and EU are similar. In both cases, regardless of $N_{lear}$, all of the techniques display a minimum $\gamma^{(V)}$ when the optimal subset of features is used. When the datasets are small, BDT, FIS and LDA produce the lowest errors (15-20%) independently on $N_{fu}$. With larger datasets, the performance of BDT improves, realizing the absolute lowest errors (around 10%) when $N_{lear}$ is maximum.
Figure 3.3: Average classification error $\gamma^{(V)}$ in case of not binned and uncorrelated synthetic datasets, as a function of $N_{\text{lear}}$. Each curve corresponds to a different $N_{fu}$, whose value is annotated nearby. Left column plots (a for KNN, c for BDT, e for FIS and g for LDA) correspond to GEB case, right column plots (b for KNN, d for BDT, f for FIS and h for LDA) correspond to EU case. The curves relative to the “important” features are plotted in red. The $K$-value used in KNN code is 1.
Looking at the case $N_{fu} = 1$, the features distributed according to $E$ (figure 3.3e-f-g-h, $N_{fu} = 1$) have a better discriminating capability than those governed by $G$ (figure 3.3a-b-c-d, $N_{fu} = 1$), at least with the fixed value of overlapping between the two classes. Indeed, in order to obtain such overlapping, the dispersion in Class 2 for exponentially distributed data is 10 times larger than in Class 1 (for example, see the first two lines of table 3.2), while the dispersion in the two classes for Gaussian data is the same. Because of this, Class 1 exponential data are almost perfectly classified, and the resulting total $\bar{\gamma}$ for $E$ is lower than for $G$.

Another property of these plots is the different relationship between $\bar{\gamma}_{(V)}$ and $N_{fu}$ among the techniques. In particular, the performance of KNN technique is heavily influenced by the presence of the "irrelevant" features, also with a large number of data. FIS and LDA, instead, feel the effects of the "irrelevant" features only with datasets composed by very few data, while the BDT performance is never significantly modified. The dependence of $\bar{\gamma}_{(V)}$ from $N_{fu}$ for BDT, FIS and LDA algorithms suggests that the monotonicity assumption holds for these techniques, at least for dataset with 50 data.

Another important result for BDT, FIS and LDA is that the $\bar{\gamma}$ does not depend considerably on $N_{lear}$: as a matter of fact, when $N_{lear} > 100$ the $\bar{\gamma}$ is comparable with the one obtained for $N_{lear} = 10000$.

Remarkably, the good behavior of BDT algorithm is particularly significant considering that this code was originally designed to classify hierarchically ordered datasets, and thus very different from the data that I simulate. On the other hand, very different considerations apply to KNN results. Its discriminating performance gets worse and worse as "irrelevant" features are added. The $\bar{\gamma}$ strongly depends also on $N_{lear}$, especially when we are far from considering the optimal subset of features. I emphasize that, in this case, the monotonicity assumption does not hold, i.e., the usual feature selection strategies do not work. In fact, when many "irrelevant" features are used, in the small sample-size case the addition of an "irrelevant" feature deteriorates.
the classification ability of the algorithm. Thus, the noise introduced by “irrelevant” features is not negligible compared to the signal of the “important” features.

3.5.2 Correlation and binning of the data

The techniques BDT, FIS and LDA are almost insensitive to correlations in the data. On the other hand, KNN results improve by 6% if the data are many and strongly correlated. In fact, when the data are correlated, the distance (normalized or not) between two points of the same category is generally smaller than in the uncorrelated case, while the distance between two points of different category generally increases.

None of the techniques shows a significant dependence of the $\bar{\gamma}$ on the binning of the data.

3.5.3 Voting of the learning dataset: overfit

For each SPR technique, the difference between the $\bar{\gamma}^{(V)}$ and the $\bar{\gamma}^{(L)}$ is displayed in figure 3.4 as a function of $N_{fu}$ and $N_{lear}$, in case of uncorrelated and not binned data.

The KNN techniques performs exactly in the same way on the voting and learning datasets. This is reasonable, considering that, when voting a learning object, this object is taken out of the learning dataset (i.e. KNN cannot choose it as one of the $K$-nearest neighbors).

For FIS and LDA, the voting of a new and independent dataset obtains a classification error up to 25% higher than the voting of the learning objects, mostly when few data ($N_{lear} \leq 50$) are used and when IRVs are included. This problem, usually named overfit, is well known (Fukunaga, 1991). In these cases, the voting of the learning dataset (a quite common practice in Earth sciences, see above) seems to produce results too much optimistic.

The problem of overfitting is present also in BDT analysis, but here it is significantly less relevant (see figure 3.4b-f). For $N_{lear} \geq 100$, the overfit is negligible for all of the techniques.
Figure 3.4: As in figure 3.3, but relative to the difference $\bar{\gamma}^{(V)} - \bar{\gamma}^{(L)}$ between the average classification error obtained when voting respectively the voting and the learning datasets, in case of not binned and uncorrelated synthetic datasets, as a function of $N_{\text{lear}}$. 

39
3.5.4 KNN: influence of the $K$-value used

In order to investigate on the dependence of the classification results on the $K$-value used in the KNN technique, I examine $\overline{\gamma^{(V)}}$, $\overline{\gamma^{(V2)}}$ and $\overline{\gamma^{(V)}}$ as functions of $N_{fu}$ and $K$. In figure 3.5 these relationships are shown in the case of uncorrelated and not binned data, with $N_{lear} = 10000$. On each panel, the minimum $\overline{\gamma^{(V)}}$ is reported for the corresponding values of $N_{fu}$ and $K$ that realize it.

In order to keep the overlap between the two classes at the fixed value $\eta = 0.9$ (see above), the variance of the data belonging to the second class is larger than that in class 1, except for the G distribution (see tables 2 and 3). Due to the greater dispersion of the objects belonging to class 2, the use of high $K$-values (i.e. $K \approx 0.1 \times N_{lear}$) allows the selection of objects belonging to the wrong class as nearest neighbors. This causes an increase in the $\overline{\gamma^{(V2)}}$ (figure 3.5b-e). As shown in figure 3.5c-f, the dependence of $\overline{\gamma^{(V)}}$ on $K$ is not very strong. However, the best results are obtained (at least for $N_{lear} \geq 1000$) if $0.001 \times N_{lear} \leq K \leq 0.01 \times N_{lear}$.

3.5.5 Case of different numbers of learning objects between the two classes

By means of this simulation, I try to reproduce a situation that is often encountered in real datasets, i.e. a significantly different number of data in the two classes. In real datasets, this is often due to a different frequency of occurrence, in nature, of a type of event with respect to another type. Because of this different frequency, a larger number of data is usually available for the more common of the two events. In order to check if the classification performance deteriorates if a class is less numerous than the other one, I perform a special learning on 20, 50 and 100 data belonging to two classes, in which approximately 25% of the data belongs to class 1, and 75% to class 2.

I do not show the results for KNN, because the performance of classification of this algorithm deteriorates by definition (see appendix A) on the objects belonging to the less numerous class. In fact, for
Figure 3.5: Average classification error $\bar{y}(V)$ realized by KNN, in case of not binned and uncorrelated synthetic datasets, as a function of $K$. Each curve corresponds to a different $N_{fr}$, whose value is annotated nearby. Left column plots (a for Class 1 objects, c for Class 2 objects and e for the totality of the objects) correspond to GEB case, right column plots (b for Class 1 objects, d for Class 2 objects and f for the totality of the objects) correspond to EU case. The curves relative to the “important” features are plotted in red. The number of learning objects is 10000.
any voting object, the probability to be voted in the more numerous class is about 3 times higher than the probability to be voted in the less numerous class.

The difference in the classification error performed by BDT, FIS and LDA in this case, compared to the case of equal population of the two classes, is shown in figure 3.6, relative to the less populated class. While algorithms BDT anf FIS do not loose significantly their classification skill, even on the less populated class, algorithm LDA performs much worse on the less numerous class, than in the case of equal number of data in the two classes. For this reason, LDA cannot be used if the two classes consist of considerably different numbers of objects.

3.6 Final remarks of the chapter

The main purpose of this chapter was to check the performance of some nonparametric statistical pattern recognition algorithms when applied to realistic datasets. In many disciplines, as in Earth sciences, this is a necessary step before applying such algorithms to real measurements, because the available data are often few, correlated, binned and very far from being normally distributed. Moreover, real datasets usually contain measurements relative to many features, but often it is not clear which ones of them play a predominant role in the process studied. A proper algorithm must be able to extract the “optimal” subset of features from the dataset.

In this chapter, I have simulated several realistic synthetic datasets with different properties: kind of statistical distribution governing the data, length of the dataset, degree of correlation between some of the data, binning of the data, inclusion in the datasets of features with no influence. Successively, I have applied 4 nonparametric statistical pattern recognition codes (K-Nearest Neighbors, Binary Decision Tree, Fisher’s Analysis and Linear Discriminant Analysis) to these synthetic datasets.
Figure 3.6: Difference between the average classification error of class 1 objects ($\bar{\gamma}$) realized by BDT, FIS and LDA, in the case in which class 1 consists of only 25% of the total data ($\gamma^{(V1)}_{\text{diff, numb.}}$) and in the case in which class 1 consists of 50% of the data ($\gamma^{(V1)}_{\text{diff, numb.}}$). Again, this is the case of not binned and uncorrelated synthetic datasets, as a function of $N_{\text{lear}}$ (using only 20, 50 and 100 learning objects). Each curve corresponds to a different $N_{fu}$, whose value is annotated nearby. Left column plots (a for BDT, c for FIS and e for LDA) correspond to GEB case, right column plots (b for BDT, d for FIS and f LDA) correspond to EU case. The curves relative to the “important” features are plotted in red. Algorithm KNN is not shown because of obvious bad results.
None of the algorithms has turned out to be sensitively affected by the presence of correlated and/or binned measurements in the dataset. At the same time, the kind of statistical distribution can significantly modify the classification errors.

The performance of BDT, FIS and LDA algorithms is slightly altered if few learning data (20) are used. On the opposite, KNN behavior depends drastically on the number of learning data.

The test on the validity of the "monotonicity" assumption has shown that only BDT, FIS and LDA are potentially able to recognize the features playing an important role in the process, also with a small number of learning data (≥ 20 for BDT, ≥ 50 for FIS and LDA) For KNN this happens only when many data (thousands) are available. Moreover, when using the latter technique, a further complication might arise from the selection of the additional parameter K.

The performed simulations show that, when FIS and LDA are used, the voting of the learning dataset produces classification errors up to 25% smaller than the voting of a new and independent dataset. For BDT, this overfit is less relevant.

I also found that the performance of FIS and BDT on datasets where the two classes consist of considerably different numbers of objects is not significantly altered. On the opposite, LDA performance on the less populated class drastically deteriorates.

According to these results, KNN technique seems the least suitable one, despite its wide popularity, in dealing with few and non-Gaussian data with "irrelevant" features. On the opposite, BDT produces the smallest classification errors and it appears to be potentially suitable in identifying the optimal subset of features, even though the synthetic data are not hierarchically structured. Algorithms FIS and LDA give a quite good classification performance, but they are affected by larger overfitting problems, compared to BDT. However, algorithm LDA can be used only if the number of objects in the classes is comparable.
Chapter 4

Pattern Recognition applied to a dataset of worldwide dataset of seismic swarms recorded in volcanic areas

4.1 Summary of the chapter

The complexity of the process responsible for volcanic eruptions makes a theoretical approach to forecasting the evolution of a volcanic unrest rather unrealistic. A more viable strategy for this purpose appears to be the identification of possible repetitive schemes (patterns) in the pre-eruptive unrest of volcanoes. Nevertheless, the scarce availability and the heterogeneity of pre-eruptive data, and the objective difficulty in quantitatively recognizing complex pre-eruptive patterns, make this task very difficult. In this work I address this issue by using a pattern recognition approach applied to the seismicity recorded during several volcanic unrests around the world. In particular, I use two nonparametric algorithms which have been proved to give satisfactory results also in dealing with few data, even if not normally distributed and/or discrete. The results show evidence of energetic differences in the seismicity between unrests preceding an eruption and isolated unrests. On the opposite, if the unrest is followed by an eruption, it seems that the seismic energy released in the unrest is not indicative
of the magnitude of the impending eruption. I also found that, in general, unrests followed by the largest explosive eruptions have a longer repose time than those related to moderate eruptions. This property supports the fact that the occurrence of a large eruption needs a sufficient amount of time since the last event in order to re-charge the feeding system.

4.2 Introduction

Volcanic eruptions have often devastating effects. A basic step towards the mitigation of their consequences consists of forecasting the time evolution of the pre-eruptive unrest at the volcanoes. The present state of knowledge of the complex physical process responsible for the volcanic eruptions makes a theoretical approach to forecasting rather unrealistic. Hence, the only viable (and realistic) strategy appears to be the empirical identification of possible repetitive schemes (patterns) in the pre-eruptive unrest of volcanoes. These patterns might indicate when, and if, the unrest is evolving into an eruption, as well as they might provide a possible estimation of the energy associated to the volcanic eruption, for instance the Volcanic Explosivity Index (VEI).

The main difficult in reaching this goal is certainly linked to the scarce availability and quality of pre-eruptive data, and to the objective difficulty to recognize quantitatively possible complex pre-eruptive patterns. As regards the first point, almost all the papers dealing with eruption forecasting are based on the experience made on a single volcanic event. Yet, this approach has an intrinsic strong limitation; the analysis, even though detailed, of a single pre-eruptive phase does not allow to discriminate the general pre-eruptive patterns and the peculiarities of the eruption considered. The formers are definitely the most important, because they only contain information useful to improve the knowledge of the physics of the erupting system. At the same time, their identification may furnish quantitative rules that can be profitably used to forecast the time evolution of the unrests in other
volcanoes. In practice, in fact, we often face on the monitoring of very
dangerous volcanoes, for instance the Vesuvius, without any quantita-
tive measurement relative to past pre-eruptive phases. In such cases,
it becomes very difficult to understand when an unusual behavior of
the volcano is really linked to an impending large eruption. Hence,
the common experience acquired in other erupting explosive volcanoes
becomes the most relevant information.

Nevertheless, very few efforts have been dedicated until now to
collect pre-eruptive data coming from different volcanoes in a single
dataset. A remarkable first (and last) attempt is represented by the
catalog compiled by Benoit and McNutt (1996) which provided seis-
mic data in a time interval of 10 years relative to pre-eruptive phases
on more than 100 volcanoes. In spite of the huge effort made by the
authors, the collected data are rather rough, being very often catego-
rical, strongly heterogeneous (for instance, the magnitude measure-
ments) and in some cases semi-qualitative. In my opinion a definite
improvement in this field can be achieved only through an interna-
tional and coordinated effort, such as the WOVOdat project, recently
proposed by Christopher Newhall (www.wovo.org/wovodat.htm).

As regards the second point, the main difficult is predominantly
technical. The volcanic unrests have a complex nature, by involving
different processes strongly interacting among them. An almost com-
plete picture of the phenomena consists of a large variety of different
signals. A robust technique that aims to identify possible pre-eruptive
patterns has to take into account all of these measurements simulta-
neously. As mentioned before, this is almost never done in the past,
because only single precursors are usually considered and analyzed
(e.g. Shibata and Akita, 2001; Londoño and Sudo, 2002; Gottsmann
and Rymer, 2002). A further complication arises by the fact that the
available data are usually few, often categorical, correlated, and their
statistical distributions are seldom Gaussian (see Benoit and McNutt,
1996). This precludes the use of all the parametric multivariate tech-
niques and neural network codes successfully used in many other sci-
cientific fields (e.g. Fukunaga, 1990).

In this chapter I provide a possible strategy of analysis that properly takes into account all of the issues discussed above. In particular, I apply Binary Decision Tree and Fisher’s Analysis codes to seek pre-eruptive patterns in a catalog containing data recorded during several volcanic unrests around the world. The dataset consists of an implementation of the Benoit and McNutt catalog (1996). In particular, I add to the original data all the available information concerning the seismic swarms related to the largest (VEI≥4) explosive eruptions occurred during the last century and to some other volcanic unrests. The main goal is to provide new insights concerning the following questions:

- Do the seismic unrests occurring before a volcanic eruption have common patterns?
- Do these possible common patterns depend on the magnitude of the following eruption?
- Do these possible common patterns depend on the type of the following eruption, or to the initial state of the conduit (close or open)?

I anticipate that, in spite of the catalog used is certainly the greatest available, its intrinsic quality is still too poor to obtain quantitative and useful rules to forecasting volcanic events. At the same time, it is possible to achieve interesting scientific insights to improve our knowledge of the physics of the eruptive processes. In any case, independently from the quality of the scientific results obtained here, an ambitious aim of this study is to introduce a new perspective in approaching the eruption forecasting issue. As soon as a worldwide catalog of volcanic unrest of good quality will be available, the strategy of analysis described here can provide a very powerful tool to recognize quantitative rules to forecast the temporal evolution of the unrests in volcanic areas.
4.3 The Dataset

The dataset that I collect and analyze consists of measurements relative to 217 seismic swarms in volcanic areas. The use of seismic data is mainly due to two reasons: the availability of this type of data for many different volcanic unrests, and the prominent importance of such data in characterizing the period of unrest before a volcanic eruption.

For each swarm, I collect as many measurements as possible that are potentially related to the occurrence of a volcanic unrest and/or to the estimation of the VEI in case an eruption occurs. I mainly refer to Benoit and McNutt (1996), but also to the Bulletin of Volcanic Eruptions, to Volcanoes of the World (Simkin and Siebert, 1994) and to the available literature on the large eruptions (VEI ≥ 4) of the last century (Miller and McGinley, 1998; Gorshkov, 1959; Tokarev, 1985; Gorshkov and Dibik, 1970; Zobin, 1971; Simkin and Howard, 1970; Faberov, 1983; Fedotov et al., 1983; Gorel’chik et al., 1983; Zobin, 1983; Reeder et al., 1977; Swanson and Kienle, 1988; Decker and Decker, 1983; Jensen et al., 1983; Smithsonian Institution’s Global Volcanism Network, 1990; Lee Siebert, personal communication, 2001; Paolo Papale, personal communication, 2002; http://volcano.umd.nodak.edu; http://www.volcano.si.edu; http://vulcan.wr.usgs.gov; http://www.nmnh.si.edu/gvp/volcano). For each swarm, I collect reported measurements of the following variables:

- the duration (DUR) of the swarm (in days);
- the repose time (REP) associated to the swarm, i.e., the time, in years, elapsed between the end of the last eruption and the beginning of the swarm;
- the maximum magnitude (MXM) recorded in the swarm;
- a binary indicator (PRE) of the occurrence of a previous swarm (0=no, 1=yes);
- a binary indicator (TRE) of the occurrence of volcanic tremor (0=no, 1=yes);
the $\phi$ function value (PHI). Considering the $k$-th swarm occurring in a certain volcanic area, PHI ($\phi^{(k)}$) is a perturbation function (Marzocchi, 2002) that mimics the stress induced on this volcanic system by all the large remote earthquakes occurred in the 35 years preceding the $k$-th swarm. In particular:

$$\phi^{(k)} = \sum_{j=1}^{N} M_{0j} \omega(d_{jk})$$  \hspace{1cm} (4.1)

where $N$ is the number of earthquakes occurred in the 35 years preceding the onset time of the $k$-th swarm, $M_{0j}$ is the seismic moment of the $j$-th earthquake and $\omega(d_{jk})$ is a weight function depending on the relative distance between the location of the $k$-th swarm and the epicenter of the $j$-th earthquake (see Marzocchi, 2002). The seismic data are taken from the catalog of Pacheco and Sykes (1992) for the period 1900-1989, and from the CMT Harvard catalog (Dziewonsky et al., 1981; Dziewonsky and Woodhouse, 1983) for the recent years. The earthquakes considered are the events with $M_s \geq 7.0$ and depth $\leq 70$ km.

As reported in figure 4.1, not for all of the swarms in the catalog I have been successful in retrieving all of the measurements. In particular, while DUR, PHI and REP have been retrieved for almost the totality of the catalog, PRE retrieving has been much more difficult.

For TRE and PRE retrieving, it is worth stressing that a 1 value simply means that some information regarding the feature has been reported (Benoit and McNutt, 1996). If a report states that, for example, “TRE measurements have been conducted”, the TRE feature is set to 1 in Benoit and McNutt (1996), and so in my dataset, regardless of the occurrence of TRE. A 0 value means that a negative result on the occurrence of TRE or PRE was reported.

The measurements which could not be retrieved are set to a number standing for missing value.

As final remark, I emphasize that the resulting catalog is certainly the largest one available at present even though its still rough quality recommends further future implementations.
Figure 4.1: Frequencies of the features in the dataset.
4.4 Pattern Recognition Analysis

In the present study, the main goal of the analysis is to recognize the prominent characteristic of the seismic swarms preceding a volcanic eruption and to find possible relationships with the VEI of the impending eruption. Due to the very limited number of data available, in this chapter I will perform only the learning phase, attempting to recognize, as a first step, all the possible patterns in the dataset. In spite of the impossibility to test the results on independent data (voting), I use some empirical strategies to check the presence of possible overfit in the results.

Before performing the learning, I first have to:

1. define the objects to be analyzed and the classes involved in the problem, and

2. select the statistical PR algorithm that is most suitable to the problem I am dealing with.

I shall explain these two steps more accurately in the following.

4.4.1 Definition of the objects and of the Classes

The objects of the analysis are the seismic swarms. Any object is represented by a vector that contains all the measurements (the features) that I can associate to the object. Due to the large differences between the maximum and minimum measurements in the catalog for DUR, REP and PHI, I decide to use the logarithm of these features. Thus, each vector has the following components: ln(DUR), ln(REP), MXM, PRE, TRE and ln(PHI).

Each vector has then a further component: it is the VEI associated to the eruption (if any) following the swarm described by the vector. If the swarm has not been followed by an eruption (I call this an isolated swarm), the VEI associated is assumed equal to -1. The VEI is useful in the definition of the class. In principle, we have 8 different classes represented in the catalog, consisting of 8 different values of
VEI (-1,0,1,...6). Since I want to apply a simple 2-class algorithm, the VEIs of the swarms will be grouped in order to reduce the problem to Class 1 versus Class 2. In order to emphasize the distinction between the two classes I will try to keep at least one unit of VEI between the lowest VEI of the upper class and the highest VEI of the lower class. For example, in order to find patterns that distinguish a swarm preceding an eruption from an isolated swarm, I consider as Class 1 all the swarms with VEI\( \geq 1 \) and as Class 2 all the swarms with VEI\( = -1 \). The complete list of the “matches” (class 1 vs class 2) performed is provided in the following.

4.4.2 Selection of the most suitable statistical PR algorithms

In this application, I will perform statistical PR only on 2-class problems, that is I will try to identify repetitive patterns between two distinct categories of objects. Many statistical PR 2-class algorithms have been successfully used in other scientific fields, such as engineering, biology, economy, medicine. In these disciplines, the available datasets are large, continue, and the variables are normally distributed.

My dataset, as well as most of the datasets in Earth sciences, does not have these ”nice” features. In particular, it is composed by few data, some of the variables (if not all) are not normally distributed (e.g. the duration of the swarm, and the occurrence of previous swarms and tremor), and they might be also correlated (e.g. the duration and the maximum magnitude). Moreover, some of the variables I have collected in the catalog are probably completely irrelevant for the eruptive process. Indeed, I have compiled my catalog by taking the largest possible number of potentially relevant variables available for each seismic swarms, because I do not know which (if any) of these variables are important for the occurrence of a volcanic eruption after the swarm, or for the determination of the VEI of the possible eruption.

As a result, I need to use a statistical PR algorithm that can per-
form satisfactorily on small datasets, characterized by continuous and discrete variables, perhaps correlated. Likely, I am including in the analysis some variables which do not affect the eruption occurrence or the VEI determination, thus it is necessary to make use of a statistical PR algorithm that is able to extract those variables having a predominant influence on the processes related to volcanic unrest. Also, the number of objects in the two classes is probably very different. According to these considerations, in this application I use two statistical PR 2-class algorithms that have proved to recognize patterns satisfactorily on small datasets and to identify the variables having a predominant role on the process (see chapter 3), i.e., Binary Decision Tree (BDT) (Rounds, 1980; Mulargia et al., 1992) and Fisher Discriminant Analysis (FIS) (e.g. Duda and Hart, 1973). The latter is here applied through a so called branch-and-bound technique in order to identify the relevant features of the process.

The use of both algorithms, that are based on very different approaches, allows to check if the results that I will obtain are due to the type of algorithm used, or to possible overfit.

For a more complete definition of the algorithms and of the branch-and-bound technique, see appendices B, C and E.

4.5 Results of the Analysis and Discussion

I performed three different 2-class analyses with different goals. In particular:

Match 1: Class 1 is represented by all the swarms followed by a volcanic eruption (VEI\geq1) and Class 2 by all the isolate swarms (VEI=1). This match is done in order to recognize the general differences between the swarms preceding a volcanic eruption and the isolated swarms.

Match 2: Class 1 is represented by all the swarms followed by a strongly explosive eruption (VEI\geq4) and Class 2 by all the isolate swarms
(VEI=-1). This match is done in order to recognize the differences between the swarms preceding a strongly explosive volcanic eruption and the isolated swarms.

**Match 3:** Class 1 is represented by all the swarms followed by a strongly explosive eruption (VEI≥4) and Class 2 by all the swarms followed by moderate eruptions (0≤VEI≤2). This match is done in order to recognize the differences between the swarms preceding a strongly explosive volcanic eruption and the swarms preceding a moderate eruptions. If the VEI of an eruption is due to random factors, we should not be able to recognize any significant difference between these two subsets of the catalog.

In each “match” I use only complete objects, i.e. the objects having non-missing values of the features considered in the analysis. I start by considering all the 6 features. Due to the missing measurements, the analysis considering all the 6 features is carried out on a lower number of objects (see tables 4.1, 4.2 and 4.3). In order to perform the analysis on a higher number of objects, and to check the stability of the results obtained on different learning datasets, I performed two additional learning phases considering smaller numbers of features. In particular, I repeat the statistical PR analysis considering (i) DUR, REP, MXM and PHI, and (ii) DUR, REP and PHI (see tables 4.1, 4.2 and 4.3). The choice for the features in (i) and (ii) is due to their larger numeroseness (allowing for a larger number of complete objects) and to their importance in the process, as suggested by the analysis carried out on all the 6 features (see below).

Since I am interested in the recognition of possible patterns in the swarms dataset, in each “match” I use all of the available complete objects for the learning phase. This allows to make use of as many data as possible for the definition of the patterns in the data.
Table 4.1: Match 1 - VEI≥1 (class 1) vs VEI=1 (class 2). In the first column the features used in the learning phase are shown; in the second column the number of available objects with non-missing values for all of the features used is shown for each class; in the two last columns, the features recognized as relevant are reported (third column for BDT algorithm, fourth column for FIS algorithm). An empty box means that the algorithm does not recognize any pattern.

<table>
<thead>
<tr>
<th>Features analyzed</th>
<th>Number of complete objects</th>
<th>Relevant feat. identified by BDT</th>
<th>Relevant feat. identified by FIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUR MXM REP PRE TRE PHI</td>
<td>Class 1: 22, Class 2: 13</td>
<td>DUR</td>
<td>DUR, REP</td>
</tr>
<tr>
<td>DUR MXM REP PHI</td>
<td>Class 1: 46, Class 2: 85</td>
<td>DUR</td>
<td>DUR</td>
</tr>
<tr>
<td>DUR REP PHI</td>
<td>Class 1: 66, Class 2: 121</td>
<td>DUR</td>
<td></td>
</tr>
</tbody>
</table>

MATCH 1

![Decision Tree Diagram]

Figure 4.2: BDT results - Pattern recognition results relative to “Match 1”, when all the features are considered. The classification is performed on the basis of the DUR and of the threshold value indicated. In total, 30 swarms are correctly classified and 5 are not.
Figure 4.3: FIS results - Pattern recognition results relative to “Match 1”, when all the features are considered. The classification is performed on the basis of DUR and REP. The standardized learning objects, projected along the line satisfying Fisher’s maximization (see x-axis title), are shown as “+” if they belong to class 1 or as “x” if they belong to class 2. For reasons of clarity, the two classes have been drawn on two different levels. The arrows mark the mean of the standardized and projected objects in the two different classes. Each object is assigned to the class whose mean is closest. In total, 31 swarms are correctly classified and 4 are not. Data shown are standardized.
<table>
<thead>
<tr>
<th>Features analyzed</th>
<th>Number of complete objects</th>
<th>Relevant feat. identified by BDT</th>
<th>Relevant feat. identified by FIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUR MXM REP PHI</td>
<td>Class 1: 5</td>
<td>DUR</td>
<td>DUR</td>
</tr>
<tr>
<td></td>
<td>Class 2: 13</td>
<td></td>
<td>PHI</td>
</tr>
<tr>
<td>DUR MXM REP PHI</td>
<td>Class 1: 17</td>
<td>MXM</td>
<td>DUR</td>
</tr>
<tr>
<td>DUR REP PHI</td>
<td>Class 1: 18</td>
<td>REP</td>
<td>REP</td>
</tr>
<tr>
<td></td>
<td>Class 2: 85</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Class 2: 121</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Match 2 - VEI≥4 (class 1) vs VEI=−1 (class 2). In the first column the features used in the learning phase are shown; in the second column the number of available objects with non-missing values for all of the features used is shown for each class; in the two last columns, the features recognized as relevant are reported (third column for BDT algorithm, fourth column for FIS algorithm).

### 4.5.1 Match 1: VEI≥1 vs. VEI=−1

As shown in table 4.1, both algorithms recognize the DUR as the predominant variable for the discrimination between Class 1 and Class 2. In particular, swarms preceding a volcanic eruptions are generally longer than isolated swarms. As an example, figures 4.2 and 4.3 show the case in which all the 6 features are considered in the analysis. Even though FIS recognizes two relevant features in this particular case (the second one is REP), we see that the largest part of the discriminating capability in figure 4.3 is given by DUR (see the coefficient relative to Log(DUR) of the line satisfying Fisher’s criterion, much larger than that relative to Log(REP)). The importance of DUR in discriminating between these two families of swarms might be an indication of a higher energy release during pre-eruptive unrests, although the magnitude of the maximum event in the seismic swarms does not seem to be different in the two types of unrests.
Figure 4.4: BDT results - Same as in figure 4.2, but relative to “Match 2”, when all the features are considered. The classification is performed on the basis of the DUR and of the threshold value indicated. In total, 16 swarms are correctly classified and 2 are not.

4.5.2 Match 2: VEI ≥4 vs. VEI = -1

As shown in table 4.2, it seems that different energies are involved also in these two types of seismic swarms (Newhall and Hoblitt, 2002). In particular, the predominant variable is DUR for both algorithms (table 4.2). A swarm preceding a large explosive eruption is generally longer than an isolated swarms. Figures 4.4 and 4.5 are shown as an example for the case in which all the 6 features are considered in the analysis. Also in this case, FIS recognizes two relevant features in this particular case (the second one is PHI), but, again, the largest part of the discriminating capability in figure 4.5 is given by DUR (see the coefficient relative to Log(DUR) of the line satisfying Fisher’s criterion, much larger than that relative to Log(PHI)).
Figure 4.5: FIS results - Same as in figure 4.3, but relative to “Match 2”, when all the features are considered. The classification is performed on the basis of DUR and PHI. In total, 17 swarms are correctly classified and 1 is not. Data shown are standardized.
<table>
<thead>
<tr>
<th>Features analyzed</th>
<th>Number of complete objects</th>
<th>Relevant feat. identified by BDT</th>
<th>Relevant feat. identified by FIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUR MXM REP PRE TRE PHI</td>
<td>Class 1: 5 Class 2: 12</td>
<td>REP</td>
<td></td>
</tr>
<tr>
<td>DUR MXM REP PHI</td>
<td>Class 1: 17 Class 2: 17</td>
<td>REP</td>
<td></td>
</tr>
<tr>
<td>DUR REP PHI</td>
<td>Class 1: 18 Class 2: 34</td>
<td>REP REP</td>
<td>PHI</td>
</tr>
</tbody>
</table>

Table 4.3: Match 3 - VEI≥4 (class 1) vs 0≤VEI≤2 (class 2). In the first column the features used in the learning phase are shown; in the second column the number of available objects with non-missing values for all of the features used is shown for each class; in the two last columns, the features recognized as relevant are reported (third column for BDT algorithm, fourth column for FIS algorithm). An empty box means that the algorithm does not recognize any pattern.

4.5.3 Match 3: VEI≥4 vs. 0≤VEI≤2

As shown in table 4.3, there is no evidence of energetic difference in these two seismic swarm types, suggesting that the energy released in the seismic swarm is not indicative of the eruption magnitude (see e.g., Newhall and Hoblitt, 2002). The only (or the most) relevant feature, identified by both algorithms, is REP (see table 4.3). Generally, the swarms corresponding to the most explosive eruptions have a longer repose time than those related to moderate eruptions (see also Newhall and Hoblitt, 2002), as shown in figures 4.6 and 4.7 for the case in which DUR, REP and PHI are considered in the analysis. A long repose time might indicate that the volcano system had a sufficient time, since the last eruption, to close the conduit and to re-charge the system.

4.6 Final remarks of the chapter

The main goal of this study was to identify common pre-eruptive patterns in worldwide volcanic unrests. At this purpose I applied non-parametric pattern recognition codes to a catalog of seismic data rel-
The swarm is classified as an eruption (Class 1)

The swarm is classified as an eruption (Class 2)

SWARMS CORRECTLY CLASSIFIED IN CLASS 1: 14
SWARMS MISCLASSIFIED IN CLASS 1: 7

SWARMS CORRECTLY CLASSIFIED IN CLASS 2: 27
SWARMS MISCLASSIFIED IN CLASS 2: 4

Figure 4.6: BDT results - Same as in figure 4.2, but relative to “Match 3”, when the features DUR, REP and PHI are considered. The classification is performed on the basis of REP and of the threshold value indicated. In total, 41 swarms are correctly classified and 11 are not.
Figure 4.7: FIS results - Same as in figure 4.3, but relative to “Match 3”, when the features DUR, REP and PHI are considered. The classification is performed on the basis of REP and PHI. In total, 41 swarms are correctly classified and 11 are not. Data shown are standardized.
ative to seismic swarms recorded in volcanic areas. The use of two algorithms based on very different "philosophies" allows the check of the stability of the results. I mainly used seismic data because they were the easiest to be retrieved and because seismic information is of prominent importance in characterizing the unrest in volcanic areas.

Although the roughness of my dataset does not allow to issue quantitative rules able to discriminate between pre-eruptive and isolated volcanic unrests, the results obtained provide useful information on the physical process governing the occurrence of a volcanic eruption after a period of unrest. In particular, there are evidence of an energetic difference between pre-eruptive and isolated periods of unrests, both considering only large explosive eruptions (VEI $\geq 4$) and all the eruptions with VEI $\geq 1$. On the contrary, no significant energetic difference is found between unrests preceding large explosive eruptions (VEI $\geq 4$) and moderate eruptions ($0 \leq$ VEI $\leq 2$). Here, although less evident, the only pattern found is based on a longer time of repose preceding the largest eruptions compared to the moderate ones. The necessity of a longer repose for a large eruption to occur might be linked to the time needed to close up the conduit, and to re-charge the feeding system.

As a final consideration, I want to stress that the quality and the practical usefulness (eruption forecasting) of the results can be dramatically improved by using this kind of techniques on large worldwide datasets of volcanic unrests, such as the one proposed in the WOVO-DAT project.
Chapter 5

Recognition of possible recurrent patterns in seismicity preceding recent flank eruptions at Mount Etna volcano

5.1 Summary of the chapter

Mount Etna flank eruptions represent the most prominent source of volcanic damage in the Etnean area. Past works suggested that the occurrence of flank eruptions in Mount Etna is linked to the regional state of tectonic stress. In this chapter, I make use of two nonparametric statistical pattern recognition algorithms and a multiple regression analysis to analyze seismic clusters occurring around Mount Etna. The goal is to study whether there is a statistically significant link between the regional “quasi-elastic” tectonic stress regime and the occurrence of flank eruptions, by looking for recurrent patterns in the seismic clusters preceding a flank eruption on Mount Etna. From the analysis, I find that the discrimination between clusters preceding flank eruptions and clusters occurring “far away” in time from flank activity is linked to the time elapsed from the end of the previous flank eruption and to its volume output, and not to the seismicity signature of the cluster itself. I apply the patterns found to a new and
independent dataset to test the validity of our findings. These results do not confirm the existence of a significant link between the regional “quasi-elastic” state of stress and the occurrence of flank eruptions on Mount Etna volcano. On the contrary, the patterns found might indicate that a prominent role in the flank eruption occurrence is played by the re-charging of the feeding system.

5.2 Introduction

Mount Etna volcano is one of the most extensively monitored volcanic systems of the world. Besides an almost continuous activity in the summit craters, every few years Mount Etna shows episodes of flank activity. Due to its densely populated slopes, the flank eruptions represent the most prominent source of volcanic damage and panic in that area.

Previous studies suggested that the occurrence of flank eruptions in Mount Etna is linked to the “quasi-elastic” regional state of tectonic stress, even though the results show different relationships, or patterns (e.g. Mulargia et al., 1991; 1992; Vinciguerra et al., 2001). I use the term “quasi-elastic” because of the time and spatial scales used. In fact here, as well as in previous studies, the time lags considered between the seismic and volcanic activity are typical of “quasi-elastic” stress perturbation, being too long for the dynamic perturbation (instantaneous, e.g. Hill et al, 1993) and too short for the post-seismic one (from years to decades, e.g. Piersanti et al, 1995; 1997; Pollitz et al, 1998).

Mulargia et al. (1991) found a correlation between the seismicity in the Gulf of Patti (see figure 5.1) and subsequent flank eruptions. Mulargia et al. (1992) found that the occurrence of flank eruptions is characterized by a large seismic activity in a broad region. Vinciguerra et al. (2001), instead, found that there is a seismicity increase some days before a flank eruption in some sectors of the area analyzed. Nevertheless their differences, all of these papers notably suggest a
tectonic control on the flank eruptions of Mt. Etna volcano. I argue that the differences among these papers can be ascribed to different factors:

1. the use of a strategy of analysis whose performance was not previously tested on datasets consisting of few data, maybe not normally distributed, and including variables that do not play a significant role in the flank eruption process;

2. the use of different seismic and volcanic datasets. In the past ten years, in fact, a definite improvement has been achieved in both fields;

3. the use of different “objects” of the analysis. For instance, Mulargia et al. (1992) and Vinciguerra et al. (2001) analyzed time intervals, while in Mulargia et al. (1991) the objects of the analysis are seismic clusters.

In this work, I re-address the issue by means of the two nonparametric statistical pattern recognition algorithms previously tested on synthetic data, and of a multivariate regression analysis, to analyze seismic clusters occurring around Mount Etna. The original seismic and volcanic catalogs available for this study are much more reliable than those used in past works (cf. Mulargia et al., 1991; 1992). The goal is to check whether there is a statistically significant link between the regional tectonic regime and the occurrence of flank eruptions on Mount Etna volcano, in order to throw light on the ambiguous results obtained in past studies. This is accomplished by searching for recurrent patterns among the seismic clusters preceding a flank eruption on Mount Etna. I also check whether these possible recurrent patterns might be used to forecast efficiently its flank eruptions.

The validity of the patterns found has been verified by applying them to an independent dataset.

67
From 1974, Mc=3.0

From 1983, Mc=2.2

Figure 5.1: Map of the Etnan area and seismic clusters extracted from the seismic catalog from (i) 1974 ($M_c = 3.0$), (ii) 1983 ($M_c = 2.2$). In both panels the circle of radius of 120Km, centered over Mount Etna volcano (star), is shown.
5.3 The Data

I analyze two different catalogs: one for the seismic events and one for the flank eruptions of Mount Etna.

Regarding the seismic data, similarly to Mulargia et al. (1991), I take into account seismic events that occurred in a circle of radius 120Km, centered over Mount Etna. The catalog from which I extract the seismic events consists of the merging of three different seismic catalogs. In fact, for the earthquakes occurred from 1960 to 1980, I refer to the PFG catalog (Postischl, 1985) with an empirical correction to the magnitude in order to make the catalog homogeneous (Lolli and Gasperini, 2003). For the events occurred from 1981 to 1996 I make use of the CSTI catalog. Finally, for the earthquakes occurred from 1997, I refer to INGV Seismic Bulletin, with a correction on the magnitude to make the catalog homogeneous.

I analyze the completeness of the catalog obtained by merging PFG, CSTI and Bollettino Sismico catalogs. In figure 5.2 the cumulative for two different threshold magnitudes are shown. I decide to extract two different datasets of seismic events: the first one starting from 1974 and having a completeness magnitude $M_c = 3.0$ (figure 5.2-i); the second one starting from 1983 and having a completeness magnitude $M_c = 2.2$ (figure 5.2-ii). The seismic events extracted for the two datasets are then grouped in order to obtain seismic clusters lasting at most one week. The first cluster consists of the first seismic event extracted and of all the seismic events occurred on the same day and in the following 6 days. The other clusters are formed subsequently in the same way. In this way I obtain 282 seismic clusters for the first dataset (from 1974, $M_c = 3.0$, see figure 5.1-i) and 638 for the second one (from 1983, $M_c = 2.2$, see figure 5.1-ii).

Concerning the flank eruption data, I refer to the list personally communicated by Marco Neri (see also Azzaro & Neri, 1992; Calvari et al., 1994; Neri & Villari, 1995; Neri & Tomarchio, 2000; Calvari et al., 2002; Harris & Neri, 2002; Behncke & Neri, 2003) and reported in
Figure 5.2: Cumulative curve of the number of seismic events listed in the original seismic catalog available, from 1960 to 2001. In panel (i) the magnitude threshold used is 3.0, while in panel (ii) it is 2.2. The arrow shows the year from which the catalog is complete (for the threshold magnitude considered).
Table 5.1: Catalog of flank eruptions at Mount Etna volcano used in this work

<table>
<thead>
<tr>
<th>Onset (yyyy mm dd)</th>
<th>End (yyyy mm dd)</th>
<th>Volume Output ($10^6 m^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1971 04 05</td>
<td>1971 06 12</td>
<td>75.00</td>
</tr>
<tr>
<td>1975 02 24</td>
<td>1975 08 29</td>
<td>11.76</td>
</tr>
<tr>
<td>1975 11 29</td>
<td>1977 01 08</td>
<td>29.40</td>
</tr>
<tr>
<td>1978 04 29</td>
<td>1978 06 05</td>
<td>27.50</td>
</tr>
<tr>
<td>1978 08 25</td>
<td>1978 08 30</td>
<td>4.00</td>
</tr>
<tr>
<td>1978 11 23</td>
<td>1978 11 30</td>
<td>11.00</td>
</tr>
<tr>
<td>1979 08 03</td>
<td>1979 08 09</td>
<td>7.50</td>
</tr>
<tr>
<td>1981 03 17</td>
<td>1981 03 23</td>
<td>30.00</td>
</tr>
<tr>
<td>1983 03 28</td>
<td>1983 08 06</td>
<td>100.00</td>
</tr>
<tr>
<td>1985 03 10</td>
<td>1985 07 13</td>
<td>30.00</td>
</tr>
<tr>
<td>1986 10 30</td>
<td>1987 03 01</td>
<td>60.00</td>
</tr>
<tr>
<td>1989 09 11</td>
<td>1989 10 09</td>
<td>38.40</td>
</tr>
<tr>
<td>1991 12 14</td>
<td>1993 03 31</td>
<td>250.00</td>
</tr>
<tr>
<td>2001 07 13</td>
<td>2001 08 08</td>
<td>35.00</td>
</tr>
</tbody>
</table>

The time series of the seismic clusters and flank eruptions in the Etnean area for both cases available (from 1974, $M_c = 3.0$; from 1983, $M_c = 2.2$) is displayed in figure 5.3. All but one the flank eruptions are preceded by at least a seismic cluster. The only flank eruption occurring without precursory seismicity is the one starting on November, 29, 1975.

5.4 The Analysis

I perform two multivariate statistical analyses by means of two different methods. The first is the Pattern Recognition method discussed in chapters 1, 2 and 3.

In the present study, I use again the two statistical PR algorithms, Binary Decision Tree and Fisher Analysis, because the simulations (see chapter 3) have shown that both of them provide reliable results when applied to datasets that mimic the basic features of our seismic cluster.
Figure 5.3: Time series of the seismic clusters (pulses) extracted from the seismic catalog (left axis, the number of events in each cluster is given), and of the flank eruptions (shaded bars, covering the period of eruption) at Mount Etna volcano (right axis, the volume output of each eruptive episode is given). Panel (i) is from 1974 and for $M_c = 3.0$, panel (ii) from 1983 and for $M_c = 2.2$. 

72
dataset, i.e. few objects, each represented by many features, often not normally distributed, and with considerably different numbers of objects in the two classes.

In this chapter I will perform the learning phase both on the totality of available data and on one of its randomly chosen subset (consisting of 80% of the available data). In the former case I attempt to recognize, as a first step, all the possible patterns in our dataset. In the latter case I keep a part of the available data as an independent data to be used for the voting phase.

Finally I perform some control experiments to check the stability of the results obtained.

Before performing the learning phase, I first have to define the objects to be analyzed and the classes involved in the problem. In this study, the objects are the seismic clusters. Any object is represented by a vector that contains all the measurements (the features) that I can associate to the object. For each seismic cluster I retrieved the following measurements:

1. maximum magnitude recorded in the cluster (MXM),

2. number of events in the cluster (NEV),

3. trimmed mean of the latitudes of the events in the cluster (LAT),

4. trimmed mean of the longitudes of the events in the cluster (LON),

5. season of occurrence of the cluster (SEA),

6. time elapsed from the previous cluster (also called repose time from the previous cluster, REC),

7. time elapsed from the end of the previous flank eruption (also called repose time from the previous flank eruption, REE),

8. maximum magnitude recorded in the previous cluster (MXP),
9. volume output of the previous flank eruption in millions of $m^3$ (VOP).

The former five measurements (MXM, NEV, LAT, LON and SEA) are relative to the cluster itself, while the latter four (REC, REE, MXP and VOP) are relative to the recent volcanic and seismic history of the system. Due to the large differences between the maximum and minimum measurements in the catalog for the repose times (REC and REE) and for the volume output of the previous eruption (VOP), I decide to use the logarithm of these features. Thus, each object has the following components: MXM, NEV, LAT, LON, SEA, Log(REC), Log(REE), MXP and Log(VOP).

Regarding the definition of the class of each object, I mainly identify 3 types of clusters:

Type "A" clusters $\iff t_{\text{succ.erupt}} - t_{\text{cluster}} \leq \tau$ (5.1)

Type "B" clusters $\iff t_{\text{cluster}} - t_{\text{prev.erupt}} \leq \tau$ (5.2)

Type "C" clusters $\iff$ otherwise (5.3)

where $t_{\text{cluster}}$, $t_{\text{succ.erupt}}$ and $t_{\text{prev.erupt}}$ are, respectively, the time of the beginning of the cluster, the time of the beginning of the flank eruption following the cluster, and the time of the end of the last flank eruption before the beginning of the cluster. If both inequalities (5.1) and (5.2) are verified, then:

Type "A" clusters $\iff t_{\text{succ.erupt}} - t_{\text{cluster}} \leq t_{\text{cluster}} - t_{\text{prev.erupt}}$

Type "B" clusters $\iff t_{\text{succ.erupt}} - t_{\text{cluster}} > t_{\text{cluster}} - t_{\text{prev.erupt}}$

Type “A” clusters are precursors to flank eruptions, while type “B” occur either during or right after the end of a flank eruption. Type “C” clusters are registered “away” from flank eruptions (figure 5.4).

Clearly, the attribution of an object to one of these types depends on the parameter $\tau$. On the basis of the frequency distribution of
Flank Eruption

Seismic Cluster

Figure 5.4: Attribution of seismic clusters (circles) to either type "A" (precursory), or "B" (following), or "C" ("away"), depending on their proximity to flank eruptions (asterisks).

the time elapsed between a cluster and its successive flank eruption, observed in our dataset, I define that a cluster has to be classified as “A” if it occurs at most about 3 months before the following flank eruption, i.e. $\tau = 100$ days. With this value for $\tau$, about 12% of the clusters contained in the dataset are of type “A”. In the control experiments, I will check the stability of our results to changes of the value of $\tau$.

By means of such definitions (equations 5.1, 5.2, 5.3 and 5.4) we have, in principle, a 3-class problem. Since I am mainly interested to a possible regional tectonic stress pattern favoring flank activity at Mount Etna, I take into account only the precursory clusters (type “A”, from now on representing class 1) and the clusters occurring “away” from flank eruptions (type “C”, from now on representing class 2). If there is a regional stress regime favoring the flank eruption mechanism, type “A” and “C” clusters should have significant differences.
5.4.1 Multivariate Regression fit

I also analyze the seismic cluster dataset with a multivariate regression analysis (Draper and Smith, 1981). Before going through the details of the analysis, it is worth to make some cautionary remarks on this kind of analysis. In particular, when we do regression calculations on unplanned data (that is, data arising from continuing operations and not from a designed experiment), some potentially dangerous possibilities can arise. For example, a false effect (a bias) on a visible variable may be caused by an unmeasured latent variable. Another undesired effect is linked to the case in which the most effective variable is kept within quite a small range, that might lead to interpret the variable as no significant. A third problem is that the use of unplanned data often causes large correlations between predictors; this makes it impossible to attribute a causal effect to one specific predictor. Actually, this technical problem can be avoided by using orthogonal dummy variables obtained, for example, through the Principal Component Analysis procedure (Draper and Smith, 1981). Unfortunately, these dummy variables do not have any real meaning, therefore their use has little to recommend it. A good detailed description of these effects can be found in Draper and Smith (1981).

From a qualitative graphical point of view, the analysis I perform builds a linear model in which the dependent variable ($y$) is the logarithm of the time from the beginning of the cluster to the following flank eruption. The independent variables are the 9 variables observed for each cluster. The first step of the analysis consists of selecting the relevant variables to explain the variability of the dependent variable in the cluster dataset. This is a very important and delicate operation in regression. Here, I select the most relevant variables by using the procedure called “Best subset search” described by Garside (1971). The procedure first requires the fitting of every possible regression equation which involves any combination of independent variables, and then selecting the case which is the best respect one criterion. The criterion adopted here is based on the $R^2$ coefficient. In few words, I report
the case with the highest $R^2$ coefficient obtained by using a single independent variable, two independent variables, three independent variables, and so on (e.g. see table 5.4). The “best” regression is the one for which the further gain in $R^2$ neglactable.

5.5 Results of the Analysis and Discussion

I have assumed that seismicity is an indicator of the tectonic stress, and that the seismicity variables considered are exhaustive in describing the seismicity. If there is a link between the regional tectonic stress and the occurrence of flank eruptions, the most important patterns in the data should be based mainly on the features characterizing the seismicity of the cluster itself. In particular, I should observe patterns involving either the number of seismic events in the clusters, or the maximum magnitude, or the coordinates of the cluster centroid, or the time elapsed from the previous cluster, or a combination of these. Instead, I do not observe any of these features as important in discriminating seismic clusters occurring just before or away from a flank eruption.

The results of the statistical PR analysis, when all the data are used in the learning phase, are shown in figures 5.5-5.6 (BDT) and 5.7 (FIS). In both cases (from 1974, $M_c = 3.0$; from 1983, $M_c = 2.2$) and for both algorithms (BDT and FIS) there is a clear pattern found. This pattern considers as discriminant features the repose time from (Log(REE)) and the volume output (Log(VOP)) of the previous eruption. In particular, for a seismic cluster occurring around Mount Etna, the larger the VOP is, the longer the REE is needed in order to be considered as a precursory cluster (class 1 or type ”A”). No other significant pattern is found.

The clusters of type “A” correctly classified by BDT are 91% and 100% respectively for the two cases (from 1974 $M_c = 3.0$, and from 1983 $M_c = 2.2$), while those of type “C” are 74% and 81% (see the
Figure 5.5: Results of the Binary Decision Tree analysis in the case from 1974 and for $M_e = 3.0$
The cluster is classified as "C" (Class 2) 

The cluster is classified as "A" (Class 1) 

The cluster is classified as "C" (Class 2) 

CLUSTERS CORRECTLY CLASSIFIED "A": 47 
CLUSTERS MISCLASSIFIED as "A": 83 
CLUSTERS CORRECTLY CLASSIFIED as "C": 359 
CLUSTERS MISCLASSIFIED as "C": 0 

Figure 5.6: Results of the Binary Decision Tree analysis in the case from 1983 and for $M_{c} = 2.2$. 

79
Figure 5.7: Results of the Fisher analysis. Panel (i) refers to the case from 1974 and for $M_c = 3.0$, panel (ii) for the case from 1983 and for $M_c = 2.2$. In each panel the standardized learning clusters, projected along the line satisfying Fisher’s maximization (see x-axis title), are shown as “+” if they belong to class 1 (type ”A”) or as “x” if they belong to class 2 (type ”C”). For reasons of clarity, the two classes have been drawn on two different levels. The arrows mark the mean of the standardized and projected objects in the two different classes. Each object is assigned to the class whose mean is closest.
<table>
<thead>
<tr>
<th>Case 1</th>
<th>Class1 (&quot;A&quot;): 26 (9)</th>
<th>Class2 (&quot;C&quot;): 139 (32)</th>
<th>Log(VOP) (23%, 32%)</th>
<th>Log(REE) Log(VOP) (34%, 29%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 2</td>
<td>Class1 (&quot;A&quot;): 34 (13)</td>
<td>Class2 (&quot;C&quot;): 357 (85)</td>
<td>Log(REE) Log(VOP) SEA (0%, 1%)</td>
<td>Log(REE) Log(VOP) (24%, 22%)</td>
</tr>
</tbody>
</table>

Table 5.2: Results of the pattern recognition analysis when only a random selected subset consisting of 80% of the available data is used in the learning phase. For both cases, the number of learning (voting) objects are given, together with the variables identified as discriminating and the classification error produced when voting the learning (voting) data.

boxes in figures 5.5 and 5.6). The results obtained by FIS algorithms are a bit worse, respectively 86%, 79%, 71%, 76% (see the boxes in figures 5.7i and 5.7ii).

By using a randomly selected subset consisting of 80% of the available data for the learning phase, and voting the remaining 20% of data as new and independent objects, I obtain the results displayed in table 5.2. In particular, I notice that:

- the pattern found is the same as in the case when all the data are used for the learning step;

- the fraction of learning objects correctly classified is comparable to the one obtained when classifying the voting data, i.e., the risk of having overfitted the data can be excluded;

- as before, FIS performance is a bit worse than BDT performance, and the classification error in the case from 1983 ($M_c = 2.2$) is much lower than in the other case.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_e = 3.0$</td>
<td>$M_e = 2.2$</td>
<td>$M_e = 3.0$</td>
<td>$M_e = 2.2$</td>
<td>$M_e = 3.0$</td>
<td>$M_e = 2.2$</td>
</tr>
<tr>
<td>$\tau = 50$ d</td>
<td>Log(VOP)</td>
<td>Log(REE)</td>
<td>Log(VOP)</td>
<td>Log(REE)</td>
<td>Log(REE)</td>
<td>Log(REE)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Log(REE)</td>
<td></td>
<td>Log(REE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clusters of one</td>
<td></td>
<td></td>
<td>Log(VOP)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>month</td>
<td></td>
<td></td>
<td></td>
<td>Log(REE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Log(REE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Excluding eruption</td>
<td></td>
<td></td>
<td>Log(VOP)</td>
<td>Log(VOP)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>Log(REE)</td>
<td>Log(REE)</td>
<td></td>
<td></td>
<td>Log(REE)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Log(VOP)</td>
<td>Log(VOP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3: Results of the control experiments for the three algorithms and the two cases.

The control experiments have confirmed the presence of the same pattern in the data, and excluded other patterns. In particular, I have repeated three times the PR analysis, each time considering one of the following variants:

- a different value of the parameter $\tau$, by using $\tau = 50$ days;

- a different cluster duration, by re-extracting clusters of one month (instead of one week), from the original seismic catalogs;

- a different flank eruption catalog, by excluding the last eruption listed (the one starting in 1991, see table 5.1).

The results obtained in these three cases are similar to those just described (see table 5.3).

As regards the multivariate regression analysis, table 5.4 reports the results obtained by using all the available data and, as dependent variable, the logarithm of the time to the next flank eruption.

These results show that the “best” regression is the one obtained when using two independent variable, in particular Log(REE) and
<table>
<thead>
<tr>
<th>N</th>
<th>1974, $M_c = 3.0$ Variables</th>
<th>$\sigma^2_{exp}$</th>
<th>1983, $M_c = 2.2$ Variables</th>
<th>$\sigma^2_{exp}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>36.7</td>
<td>9</td>
<td>41.6</td>
</tr>
<tr>
<td>2</td>
<td>7,9</td>
<td>44.8</td>
<td>7,9</td>
<td>61.8</td>
</tr>
<tr>
<td>3</td>
<td>4,7,9</td>
<td>46.1</td>
<td>6,7,9</td>
<td>63.4</td>
</tr>
<tr>
<td>4</td>
<td>3,4,7,9</td>
<td>46.6</td>
<td>3,6,7,9</td>
<td>64.5</td>
</tr>
<tr>
<td>5</td>
<td>2,3,4,7,9</td>
<td>47.1</td>
<td>1,3,6,7,9</td>
<td>65.0</td>
</tr>
<tr>
<td>6</td>
<td>2,3,4,5,7,9</td>
<td>47.3</td>
<td>1,3,6,7,8,9</td>
<td>65.3</td>
</tr>
<tr>
<td>7</td>
<td>2,3,4,5,6,7,9</td>
<td>47.4</td>
<td>1,3,5,6,7,8,9</td>
<td>65.4</td>
</tr>
<tr>
<td>8</td>
<td>2,3,4,5,6,7,8,9</td>
<td>47.4</td>
<td>1,3,4,5,6,7,8,9</td>
<td>65.5</td>
</tr>
<tr>
<td>9</td>
<td>1,2,3,4,5,6,7,8,9</td>
<td>47.4</td>
<td>1,2,3,4,5,6,7,8,9</td>
<td>65.5</td>
</tr>
</tbody>
</table>

Table 5.4: Results of the multivariate regression analysis when all the data are used in the learning phase. In each row, by using only $N$ ($N \leq 9$) independent variables, the best regression is obtained if the variables indicated are used, and the correspondent amount of variability in the data explained by this regression is given ($\sigma^2_{exp}$).

Log(VOP), with coefficients:

\[
\begin{align*}
\text{Logy} &= 2.01 + 1.25 \text{Log}(VOP) - 0.666 \text{Log}(REE) & \text{case } 1974, \ M_c = 3.0 \\
\text{Logy} &= 2.17 + 1.51 \text{Log}(VOP) - 0.897 \text{Log}(REE) & \text{case } 1983, \ M_c = 2.2
\end{align*}
\]

The addition of any other feature improves very little the $R^2$ coefficient. The results obtained on a randomly chosen subset of the dataset available, consisting of only 80% of the total amount of data, are shown in table 5.5. Again, the best subset of variables consists of Log(REE) and Log(VOP).

In the last two columns of table 5.3 the results of the multivariate regression analysis obtained in the control experiments are displayed. They do not differ significantly from those obtained by the pattern recognition algorithms. Note, however, that the errors affecting the coefficients of the multiple regression analysis are too large to be profitably used as a forecasting rule.

In the light of these considerations, it is clear that our results are quite different from what was previously obtained by Mulargia et al. (1991,1992) and Vinciguerra et al. (2001). All of these studies, in
<table>
<thead>
<tr>
<th>N</th>
<th>1974, $M_c = 3.0$</th>
<th>$\sigma^2_{exp}$</th>
<th>1983, $M_c = 2.2$</th>
<th>$\sigma^2_{exp}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>27.4</td>
<td>9</td>
<td>34.0</td>
</tr>
<tr>
<td>2</td>
<td>7,9</td>
<td>31.9</td>
<td>7,9</td>
<td>55.0</td>
</tr>
<tr>
<td>3</td>
<td>5,7,9</td>
<td>33.6</td>
<td>3,7,9</td>
<td>55.6</td>
</tr>
<tr>
<td>4</td>
<td>5,7,8,9</td>
<td>35.0</td>
<td>1,3,7,9</td>
<td>55.9</td>
</tr>
<tr>
<td>5</td>
<td>4,5,7,8,9</td>
<td>35.6</td>
<td>1,3,6,7,9</td>
<td>56.2</td>
</tr>
<tr>
<td>6</td>
<td>2,4,5,7,8,9</td>
<td>35.9</td>
<td>1,3,6,7,8,9</td>
<td>56.5</td>
</tr>
<tr>
<td>7</td>
<td>2,3,4,5,7,8,9</td>
<td>36.1</td>
<td>1,3,4,6,7,8,9</td>
<td>56.6</td>
</tr>
<tr>
<td>8</td>
<td>2,3,4,5,6,7,8,9</td>
<td>36.1</td>
<td>1,3,4,5,6,7,8,9</td>
<td>56.7</td>
</tr>
<tr>
<td>9</td>
<td>1,2,3,4,5,6,7,8,9</td>
<td>36.1</td>
<td>1,2,3,4,5,6,7,8,9</td>
<td>56.7</td>
</tr>
</tbody>
</table>

Table 5.5: Results of the multivariate regression analysis when only a random selected subset consisting of 80% of the available data is used in the learning phase. In each row, by using only $N$ ($N \leq 9$) independent variables, the best regression is obtained if the variables indicated are used, and the correspondent amount of variability in the data explained by this regression is given ($\sigma^2_{exp}$).

In fact, recognized particular seismicity patterns preceding flank activity, linked, in some way, to the tectonic regime.

### 5.6 Final remarks of the chapter

In this chapter I have applied statistical multivariate analysis, pattern recognition and multivariate regression, to a dataset of seismic clusters registered around Mount Etna, in order to check whether there is a statistically significant link between the regional tectonic stress at a "quasi-elastic" scale that might favor the occurrence of flank eruptions on this volcano. All but one the flank eruptions are preceded by at least a seismic cluster. The only flank eruption occurring without precursory seismicity is the one starting on November, 29, 1975.

The results of this analysis, confirmed by the control experiments, show that the only pattern found is linked to the time elapsed from the end of the previous flank eruption and to its volume output. In particular, for a seismic cluster occurring around Mount Etna, the larger the volume output of the last eruption is, the longer the repose
time is needed in order to be considered as a precursory cluster.

Concerning the main goal of our study, these results stand for a not significant link between the regional state of stress and the occurrence of flank eruptions on Mount Etna volcano.

At the same time, I note that the pattern found in this work might indicate the need of the system to re-charge, before the next flank eruption can occur.
Chapter 6

On the validation of earthquake forecasting models: the case of Pattern Recognition algorithms

6.1 Summary of the chapter

Earthquake forecasting is one of the geophysical issues with a potentially large social and political impact. Besides the purely scientific interest, the loss of lives and the huge damage caused by seismic events in many regions of the world have led many research groups to work in this field. Until now, however, the results obtained are not convincing and often they are a matter of intense debates. In part, these debates are due to the ambiguous definition of key concepts, such as precursor and forecast/prediction, as well as the lack of a clear strategy to set up and check an earthquake forecasting model. In this chapter, I provide insights that might contribute to better define formally the earthquake forecasting problem, both in setting up and in testing the validity of the forecasting model. As an illustration, I apply these insights to the forecasting models M8 and CN based on a Pattern Recognition approach. I find that the forecasting capability of these algorithms is very likely significantly overestimated.
6.2 Introduction

The forecasting of large earthquakes is one of the main goals in Earth sciences. Besides the scientific interest, its primary importance is mainly linked to the social, economic, and political impact of the devastating effects caused by these natural events.

In spite of many efforts devoted to forecasting earthquakes, the results obtained so far are not particularly encouraging. Studies of a variety of possible precursors has not shown a statistically significant capability to predict seismic events. In particular, most of the precursors suggested have been observed only in retrospective studies, but forward analysis has shown that they cannot be considered characteristic for the processes that are the prelude to a large earthquake. Nowadays, it is generally accepted that retrospective analyses are necessary to set up an earthquake forecasting model (EFM), but they cannot be used to test its performance. Thus, an a posteriori claim of an earthquake forecast should not be taken as a definitive test. This is important, because a posteriori warnings carelessly amplified by mass media can increase the frustration of the people struck by the calamity.

From a scientific point of view, the lack of convincing evidence of consistent precursory phenomena, together with theoretical considerations, led to a widespread skepticism about the real possibility to forecast earthquakes (cf. Geller et al., 1997). Some researchers, for instance, described earthquake prediction as the alchemy of our century (Geller, Nature Debate, 1999). Others disagree, believing that we will be able to improve significantly our ability to forecast great seismic events (e.g., Wyss, 1997). Apart from these interesting philosophical speculations, it is unquestionable that a careful reading of the large body of literature on this issue might leave the reader with a hazy picture, typical of an undefined problem. For instance, the terms prediction and forecasting are usually, but not always, taken as synonymous. The meaning of the term precursor appears strongly sub-
jective, embracing a vast number of diverse definitions (e.g., Kagan, 1997). Moreover, unambiguous and generally accepted procedures to set up and test EFMs are not yet available. Also, randomness and unpredictability (of the earthquakes) are usually (and erroneously) considered as synonymous concepts.

The purpose of this chapter is to provide insights to better define some of the issues mentioned above. In particular, I focus the attention on defining an unambiguous and correct statistical procedure to evaluate the performance of any EFM. Finally, I apply it to critically evaluate two of the most interesting forecasting models reported in the literature in the last years, i.e., the M8 and the CN algorithms originally proposed by Keilis-Borok and Kossobokov (1990a), and Keilis-Borok et al. (1988).

6.3 Earthquake Forecasting Model

As mentioned before, many researchers believe that earthquake prediction is an unachievable scientific goal. At the same time, it is obvious that we can successfully forecast some seismicity attributes. For example, at least one earthquake of magnitude $M \geq 7.0$ will occur somewhere on the Earth surface in the next ten years. However, such a forecast is not particularly useful. Therefore, here and in all related papers, it is more or less tacitly assumed that the spatio-temporal window considered for the forecasting is small enough to have some kind of practical utility. For instance, we can range from a space-time window small enough to develop rational strategies for earthquake mitigation, such as evacuations, to larger dimensions that are characteristic of the seismic hazard domain.

An EFM can be theoretical and/or empirical. Our present lack of knowledge of the earthquake source process limits drastically the use of theoretical models (a first attempt has been made recently by Parsons et al. (2000)). Consequently, most EFMs are empirical. It is important to distinguish between two kinds of empirical information:
(a) the past distribution of earthquakes of the same magnitude as the ones we want to forecast (from now on target earthquakes);

(b) the observation of precursors.

Here, the term precursor is taken to mean all of the signals, seismic or not, that precede a target earthquake, except other previous target earthquakes (defined in (a)). In the following, for the sake of conciseness and in order to distinguish models based on different information, I use the acronyms REFM (Reference Earthquake Forecasting Model) for an EFM based only on the distribution of the past target earthquakes (case (a) above; see, for example, Kagan and Jackson, 2000), and PEFM (Precursor Earthquake Forecasting Model) if the EFM is also based on observation of precursors (case (b) above) that might improve the REFM to a greater or lower extent.

In general, the setting up of an EFM consists of defining the quantitative formulation of the model and estimation of its parameters by using a retrospective analysis. I call this step the learning phase, borrowing the term from the literature on Pattern Recognition (Duda and Hart, 1973; Fukunaga, 1990; Gelfand et al., 1976). Regardless of the nature of the EFM, it has to be clearly defined to be used correctly in any circumstances by any researcher. Although this is an obvious requirement, past experience shows that it is not always fulfilled. The VAN method is a good example (see, e.g., Mulargia and Gasperini, 1992, and references therein; Geophys. Res. Lett. Debate, 1996).

6.4 EFM validation

The evaluation of the forecasting capability of an EFM is a critical step to judge its scientific and practical relevance. There is not yet a consensus on a well defined procedure to follow. As a consequence, many of the discussions in recent years concern this issue. A general strategy consists of the application of the EFM to a new and independent data set. I call this step the voting phase. It is crucial that all
the parameters of the model estimated in the learning phase are kept fixed in this phase (e.g. Mulargia, 2001; Console and Murr, 2001).

Even though in many cases this step is performed during a retrospective analysis (for example, by dividing the original data set in two parts, one for the learning phase, and the other for the voting phase), it is advisable to use real new data that can be obtained only by forward analysis (e.g. Rhoades and Evison, 1989; Mulargia, 2001; Console and Murr, 2001). Retrospective analysis, in fact, allows one to tune the parameters after the event to produce apparently high statistical significance (Geller, *Nature Debate*, 1999; Mulargia, 2001). Forward analysis prevents the parameters of the models from being chosen (consciously or not) in such a way that the model performs well in both the learning and voting phases. Sub-conscious tuning of a model to fit the voting phase can result in an overconfidence in the capability of the EFM, achieved only as a good faith by the researchers.

The validation of the REFM and PEFMs is different. For the REFM, we can use the earthquakes of the voting phase to perform a goodness-of-fit test of the model (e.g. Kalbfleisch, 1985). We can compare, for instance, the expected and the observed number of events in a chosen spatio-temporal window. In general, the PEFMs can be grouped in two broad classes: PEFMs consisting of probability maps (similar to the REFM) that do not explicitly declare earthquake "predictions", and PEFMs providing explicit proclamations of "prediction" of events. In the first case, the comparison between REFM and PEFM can be done by matching the probabilities of the two models associated to the earthquakes of the voting phase. In the second case, the best procedure is to compare the forecasting ability of the PEFM with the forecasting ability of a REFM (cf. Stark, 1997; Evison and Rhoades, 1999; Console and Murr, 2001). In both cases, to accomplish this goal correctly, we have to pay attention to two crucial issues: the definition of an appropriate REFM model, and the statistical procedure to compare the results obtained by the PEFM and the REFM.
In this case, the REFM plays the same role as the null hypothesis in statistics, since it represents our state of knowledge of the earthquake occurrence process. If the REFM is inappropriate, a PEFM based on a set of well defined precursors might appear statistically significant beyond its real merits (see Stark, 1997).

### 6.5 PEFM testing

In the case the PEFM consists of a probability map, the comparison with the REFM can be accomplished in a statistically standard way. The probabilities associated to the earthquakes of the voting phase estimated through the PEFM and REFM can be considered as two random samples of two apparently different parent distributions. In this case, we can use one standard statistical test, for instance the likelihood ratio test (e.g., Kalbfleisch, 1985), to evaluate the type and amplitude of the differences between the two samples. If the PEFM consists of a model providing an explicit declaration of earthquake “prediction”, the validation is certainly less standard. Here, I focus the attention on this particular issue for different reasons; first, most of the PEFMs proposed are of this type (e.g., the VAN method, and the M8 algorithm). Second, these PEFMs have a larger impact from a social point of view because of their explicit declaration of earthquake “prediction”. Finally, as mentioned above, the validation of these models is still poorly defined.

The simplest way to compare the forecasting ability of a PEFM with a REFM is to examine the expected and observed number of successful forecasts. Assuming that each success or failure to forecast is independent from the others, we can use a Bernoulli trial scheme (see, e.g., Kossobokov et al., 1999a). The statistical significance $\alpha$ of the PEFM, i.e., the probability to observe PEFM ”successful forecasts” by chance, is estimated by

$$\alpha = \sum_{i=n}^{N} \binom{N}{i} \tilde{\theta}^i (1 - \tilde{\theta})^{N-i},$$

(6.1)
where \( n \) is the number of successes, \( N \) is the total number of target earthquakes, and \( \hat{\theta} \) is the probability to have a success by chance.

Here, all the information concerning the REFM is contained in the parameter \( \hat{\theta} \) of the binomial distribution. The parameter \( \hat{\theta} \) is calculated by relating the spatio-temporal coverage of the alarms with the probability provided by the REFM

\[
\hat{\theta} = \int \int_\Omega f(\vec{x}, t) \, d\vec{x} \, dt,
\]

where \( \Omega \) is the total spatio-temporal coverage of the alarms, \( \vec{x} \) are the spatial coordinates, \( t \) is the time, and \( f(\vec{x}, t) \) is the probability density function (pdf) that describes the spatio-temporal distribution of the past seismic events with magnitude comparable to the ones we want to forecast. In other words, \( f(\vec{x}, t) \) is the REFM.

An implicit assumption in calculating \( \alpha \) through equation (6.1) is that the parameter \( \hat{\theta} \) is exactly known, i.e., we know exactly the form and the parameters of the REFM, \( f(\vec{x}, t) \). On the other hand, if the REFM is not appropriate or it is estimated from empirical data, equation (6.1) provides incorrect significance levels that usually lead to overestimation of the PEFM capability. In statistical terms, if the REFM is not appropriate, the significance level is relative to a wrong null hypothesis (cf. Stark, 1997). Let us consider, for example, a REFM formed by earthquakes that follow a spatio-temporal Poisson process over a sphere. A PEFM that is capable of capturing the tendency of past earthquakes to cluster along the plate margins will have a statistically significant success, compared to the REFM, independently of the set of precursors used.

To summarize, in order to estimate the real forecasting capability of a PEFM, it is necessary:

1. to define a proper null hypothesis, i.e., to set up the REFM by using past events of the same size as the ones we want to forecast, and

2. to account for possible uncertainties in the estimation of \( \hat{\theta} \).
6.5.1 Effects of "incorrect" REFM

An "incorrect" REFM is characterized by a pdf \( f^*(\bar{x}, t) \) (the symbol * indicates an "incorrect" REFM) that is different from the "correct" pdf \( f(\bar{x}, t) \). Independently of the kind of differences in the statistical distributions, I remark that the effect of the incorrect choice of REFM is almost always to overestimate the PEFM forecasting capability, by underestimating the parameter \( \bar{\theta} \). The few cases in which wrong REFM\s overestimate the parameter \( \bar{\theta} \) are linked to unrealistic statistical distributions with sharp peaks of probability. These cases can be easily recognized because the derived REFM\s forecast a number of earthquakes that is inconsistent with the parameter \( \bar{\theta} \).

Let us consider, for example, a "correct" REFM consisting of a filtered map of shallow (depth \( \leq 70 \) km) seismicity with \( M_s \geq 7.5 \) in the period 1900-1984 (see figure 6.1).

The earthquakes are taken from the Pacheco and Sykes' (1992) catalog. This REFM has been chosen because it will be useful also in testing the M8 algorithm (see below). The filtered map has been built in two steps: in the first step I divide the earth surface into cells having approximately a square shape and equal area \( (A) \), then I count the number of earthquakes with \( M \geq 7.5 \) and depth \( \leq 70 \) km for each cell. In the second step the value for each cell has been spatially filtered by using a Gaussian bell with a standard deviation approximately equal to the linear dimensions of the cell. Finally, the values are normalized in such a way that I have a probability map; that is I have imposed the condition that the sum over the Earth surface is 1. The resulting probability map represents a binned estimation of \( f(\bar{x}, t) \), where a probability \( p_k \) is assigned to the \( k \)-th cell. In figure 6.1, I show a probability map obtained by using cells with \( 300 \times 300 \) km. I check the stability of the results for different dimension of the cells.

A simple way to forecast earthquakes with this REFM is to select the cells with the highest probabilities and with a cumulative probability equal to the percentage of spatio-temporal coverage of alarms.
Figure 6.1: Probability map for the occurrence of earthquakes with $M \geq 7.5$ and depth $\leq 70$ km. The map has been obtained by fitting the earthquakes reported by the Pacheco and Sykes' (1992) seismic catalog (see text). Different colors mean different probabilities (see the legend). The black and red dots are the target earthquakes in the time interval 1985-1997 with $7.5 \leq M < 8.0$ and $M \geq 8.0$, respectively.
<table>
<thead>
<tr>
<th>$\theta$ (%)</th>
<th>$\theta^*$ (%)</th>
<th>$\theta^*$ (%)</th>
<th>$\theta^*$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“correct” REFM</td>
<td>$M \geq 7.5,$ 1900-1984&lt;sup&gt;a&lt;/sup&gt;</td>
<td>$M \geq 7.0,$ 1900-1984&lt;sup&gt;a&lt;/sup&gt;</td>
<td>$M \geq 6.0,$ 1950-1984&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>0.50</td>
<td>0.40</td>
<td>0.32</td>
<td>0.31</td>
</tr>
</tbody>
</table>

<sup>a</sup> Pacheco and Sykes (1992)
<sup>b</sup> Perez (1999)
<sup>c</sup> CMT catalog (see Dziewonski et al., 1981; Dziewonski and Woodhouse, 1983)

Table 6.1: Estimation of the "correct" coverage of alarm $\tilde{\theta}$, made through the "correct" and "incorrect" REFMs. I also show the threshold magnitudes and the time intervals in which the catalogs are complete for these thresholds.

If we call this set of cells $\Omega'$, we have

$$\tilde{\theta} = \sum_{k \in \Omega'} p_k \quad (6.3)$$

In this example I consider $\tilde{\theta} = 0.50$. These cells are considered in an "alarm" state for the whole time period (cf. Kossobokov et al., 1999a). In order to emphasize the bias in the $\tilde{\theta}$ estimation performed by means of an "incorrect" REFM, I build other REFMs based on different information, i.e., on seismicity with lower thresholds. This guarantees having REFMs with a larger dispersion of the "correct" REFM. Then, for each of these "incorrect" REFMs, I calculate the spatio-temporal coverage of alarms, i.e., $\tilde{\theta}^*$, associated with the set of cells $\Omega'$. In other words,

$$\tilde{\theta}^* = \sum_{k \in \Omega'} p_k^* \quad (6.4)$$

where $p_k^*$ is the probability of the $k$-th cell in the "incorrect" REFM.

In table 6.1 I report the values of $\tilde{\theta}$ calculated for the "correct" REFM and $\tilde{\theta}^*$ calculated for the "incorrect" REFM considered. Note that in all cases $\tilde{\theta}^* < \tilde{\theta}$. The bias $\delta$, defined as

$$\delta = \tilde{\theta} - \tilde{\theta}^* \quad (6.5)$$
ranges from 0.1 to 0.2. This means that an “incorrect” REFM underestimates the real spatio-temporal coverage of alarms. The effect is similar to what was observed by Kossobokov et al. (1999a) comparing the spatio-temporal coverage of alarms for a uniform space-time distribution with a spatio-temporal distribution weighted with the $M \geq 4.0$ seismicity. The final effect of the bias $\delta$ is to introduce a related bias in the estimation of $\alpha$ through equation (6.1), leading to an overestimation of the forecasting ability of the PEFM. The amplitude of this bias depends on $\delta$, $\tilde{\theta}$, $N$, and $n$ (see equation (6.1)).

### 6.5.2 REFM uncertainties

In reality, we never know exactly the real REFM, but we estimate it from data. In this case, each value calculated from this estimated REFM has an uncertainty attached that depends on the quality of our estimation of the REFM. In this section, I suggest a possible strategy to properly take into account the uncertainty in the $\tilde{\theta}$ estimation. To accomplish this goal, it is helpful to use a Bayesian approach (cf. Geller et al., 1997; Gelman et al., 1995). In this view, the probability to have $y$ successes is

$$ p(y) = \frac{1}{\theta} \int p(y, \theta) \, d\theta = \frac{1}{\theta} \int p(\theta) \, p(y | \theta) \, d\theta, \quad (6.6) $$

where $p(\theta)$ is the pdf of the parameter $\theta$ that accounts for its uncertainty, and $p(y | \theta)$ is the binomial distribution (cf. equation (6.1)).

The form of $p(\theta)$ describes our knowledge of the REFM. For instance, if we have a perfect knowledge of the REFM (of the form of $f(\bar{x}, t)$) and its parameters, $p(\theta)$ is a Dirac delta distribution:

$$ p(\theta) = \delta(\theta - \tilde{\theta}) \quad (6.7) $$

In this case equation (6.6) becomes

$$ p(y) = p(y | \tilde{\theta}) \quad (6.8) $$

where the significance level $\alpha$ is given exactly by equation (6.1). Alternatively, if we do not know anything about the REFM, a reasonable
choice for $p(\theta)$ is a $[0,1]$ uniform distribution (see e.g., Feller, 1968; Gelman et al., 1995).

These two limiting cases represent two unrealistic extremes, i.e., perfect knowledge and perfect ignorance of the REFM. In case we have an estimation of the parameter $\hat{\theta}$, it is more appropriate to use an unimodal pdf centered around $\hat{\theta}$. It is worth pointing out that the choice of pdf is rather subjective. This is the main ”weakness” associated with the Bayesian approach, that has been a matter of hot debates in the statistical community. It is not a goal of this chapter to deepen this issue, but I stress that any possible bias introduced by a subjective choice of the pdf is certainly smaller than the bias obtained by using a Dirac delta distribution. The latter is, actually, a strongly subjective choice too. In fact, a Dirac delta distribution can be obtained by a beta distribution with a dispersion that tends towards 0. As a rule of thumb, we can overcome this problem by using different reasonable pdfs and verifying the stability of the results. Here, I suggest two possible choices. The first one is to transform the random variable $\theta$ through a logit transformation

$$ u = \log \frac{\theta}{1 - \theta} $$

This transformation allows one to work with a continuous random variable in the domain $(-\infty, \infty)$ rather than with a random variable, $\theta$, defined in $[0,1]$. This allows one to use a Gaussian distribution for $p(u)$ in the transformed integral (see equation (6.6))

$$ p(y) = \int_{-\infty}^{\infty} p(u) p(y|u) \, du $$

Another possibility is to use a beta distribution for $p(\theta)$ (Gelman et al., 1995)

$$ p(\theta) = \frac{\Gamma(a + b)}{\Gamma(a) \cdot \Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1}, $$

where $a$ and $b$ are the parameters of the beta distribution and $\Gamma$ is the gamma function. The parameters $a$ and $b$ are functions of the mean
and variance of the distribution

\[
a = E(\theta) \left[ \frac{E(\theta)(1 - E(\theta))}{\text{var}(\theta)} - 1 \right] \quad (6.12)
\]

\[
b = (1 - E(\theta)) \left[ \frac{E(\theta)(1 - E(\theta))}{\text{var}(\theta)} - 1 \right] \quad (6.13)
\]

where \( E(\theta) \) and \( \text{var}(\theta) \) are, respectively, the mean and variance of \( \theta \). Since the beta distribution is conjugate to the binomial distribution (see Gelman et al., 1995), the integral of equation (6.6) can be solved analytically. By using equation (6.11), equation (6.6) becomes

\[
p(y) = \binom{N}{y} \frac{\Gamma(a + b)}{\Gamma(a) \cdot \Gamma(b)} \int_0^1 \theta^{y+a-1} (1 - \theta)^{N+b-y-1} \, d\theta \quad (6.14)
\]

\[
p(y) = \binom{N}{y} \frac{\Gamma(a + b) \cdot \Gamma(a + y) \Gamma(N + b - y)}{\Gamma(a) \cdot \Gamma(b) \cdot \Gamma(N + a + b)} \quad (6.15)
\]

Finally, equation (6.6) can be rewritten as

\[
\alpha = \frac{\Gamma(a + b)}{\Gamma(a) \cdot \Gamma(b)} \sum_{i=n}^{N} \left[ \binom{N}{i} \frac{\Gamma(a + i) \Gamma(N + b - i)}{\Gamma(N + a + b)} \right] \quad (6.16)
\]

where \( a \) and \( b \) are estimated by equations (6.12) and (6.13).

Both strategies require an estimation of the central value and dispersion of \( \hat{\theta} \). The central value is directly given by the \( \hat{\theta} \) estimation. The dispersion can be attributed heuristically, according to our degree of confidence about the quality of the REFM. In practice, it is convenient to check the forecasting ability of the PEFM as a function of possible (realistic) values of the dispersion of \( \hat{\theta} \). In general, it is easy to demonstrate that the higher the standard deviations of \( p(\theta) \) and \( p(u) \) in equations (6.6) and (6.10), the higher the \( \alpha \) values (i.e., less statistically significant). For example, in case we follow the latter of the two strategies above mentioned, the first two moments (mean and variance) of the beta distribution (see equations (6.11), (6.12) and (6.13)) are

\[
E(\theta) = \frac{a}{a + b} \quad (6.17)
\]
Figure 6.2: Trend of $\alpha$ (see equation (6.16)) as a function of $x$. The variable $x$ is defined in the Appendix. Note that the variance of the $beta$ distribution decreases as $x$ increases.

and

$$
\sigma^2 = \text{var}(\theta) = \frac{ab}{(a + b)^2(a + b + 1)} = E(\theta)(1 - E(\theta)) \frac{1}{a + b + 1} \quad (6.18)
$$

In order to take into account the effects of the variance of the $beta$ distribution on $\alpha$ given by equation (6.16), I define $a = x \cdot a'$ and $b = x \cdot b'$, where $x$ is a real variable $\geq 1$ and $a'$ and $b'$ are the lowest values for $a$ and $b$ that correspond to the maximum variance. In this way, I can decrease the variance of the distribution leaving unaffected the average, simply by increasing $x$.

I investigate numerically on this relationship. In particular, I carefully study the parameter space defined by the conditions

$$
a, b \geq 1 \quad , \quad 0 < E(\theta) < 1 \quad , \quad p - \sigma_p > E(\theta) \quad (6.19)
$$

where $p = n/N$, $\sigma_p = \sqrt{p(1-p)}$, and $n$ and $N$ have the same meaning than in equation (6.16). The first condition is necessary to have a finite pdf for the $beta$ distribution (see Gelman et al., 1995). The third
condition means that the percentage of the “success” observed minus its standard deviation has to be larger than the expected average. Note that this does not represent a real limitation because it is the only case of practical interest.

In figure 6.2, I report the value of $\alpha$ given by equation (6.16) as a function of $x$. In this figure I set $N = n = 7$, and $a' = b' = 1$ that correspond to a uniform distribution (a beta distribution with $E(\theta) = 0.5$ and maximum variance). The intensive exploration of the parameter space given by the conditions reported above shows that the trend is always monotonically decreasing as the one reported in figure 6.2. This implies that, in the parameter space investigated, $\alpha$ increases (i.e., it becomes less statistically significant) when $x$ decreases, i.e., when the variance of the beta distribution increases.

To summarize, I emphasize two effects that, if not properly taken into account, tend to produce underestimation of $\alpha$, that is overestimation of the forecasting capability of a PEFM:

1. the use of an ”incorrect” REFM usually produces a bias in the estimation of $\hat{\theta}$, i.e., an underestimation of the percentage of the spatio-temporal coverage of alarms;

2. the failure to account for unavoidable uncertainty in the estimation of $\hat{\theta}$.

6.6 Pattern Recognition applied to earthquake forecasting: M8 and CN algorithms

Pattern Recognition (PR) is a very powerful multivariate analysis technique allowing, in principle, the identification of possible repetitive schemes among the objects belonging to distinct categories. Whilst usual data analysis takes into account only one variable of the process at a time, PR is able to extract information from any possible
combination (linear or not) of variables that are supposed to have an influence on the process. Moreover, PR does not need the construction of a theoretical model, but it is usually based on a basic and sole hypothesis, i.e., the assumption that the phenomenon under study is governed by a finite number of complex, but repetitive patterns of the variables.

These appealing features led PR to be applied with success in several and diverse disciplines which share the study of complex systems. It has been applied successfully in many fields, such as waveform analyses in engineering, brain modeling in biology and psychology, stock market behavior in economy, medical records in medicine, to mention just a few (Duda and Hart, 1973; Fukunaga, 1990). Earth sciences, as well as most of the natural disciplines, are characterized by the ubiquitous presence of complex systems, governed by a very high number of degrees of freedom. In general, classical physical/mathematical modeling does not appear suitable to describe satisfactorily the nature of these processes, while PR may provide useful theoretical and empirical insights by identifying repetitive patterns that characterize the system under study.

The most common application of PR in geophysics is to forecasting geophysical events, such as strong earthquakes (e.g. Keilis-Borok et al., 1988; Keilis-Borok and Kossobokov, 1990a) and volcanic eruptions (Mulargia et al., 1990; Mulargia et al., 1992). In particular, the most relevant applications in earthquake forecasting are the CN and M8 codes. These algorithms have been applied systematically to forecasting worldwide seismicity since the early 90’s (e.g., Keilis-Borok and Kossobokov, 1990a; Keilis-Borok and Rotwain, 1990). The same codes have also been applied to the local seismicity in many areas of the world (e.g., Keilis-Borok et al., 1988; Keilis-Borok et al., 1990a,b; Keilis-Borok and Kossobokov, 1990b; Gahalaut et al., 1992a,b; Kossobokov et al., 1999b; Costa et al., 1995; Costa et al., 1996; Costa et al., 1997; Romachkova et al., 1998; Peresan et al., 1999). From the outset, the models and their forecasting capability led to many
debates, that are partly due to a lack of agreement on how to check the forecasting capability of a model.

Since M8 and CN codes are based on the observation of seismic activity of magnitude lower than the target events (the earthquakes to be forecast), they can be classified as PEFMs (see above). Before giving a detailed description of the two codes, it is necessary to recall a few basic points. As mentioned above, the correct set up of a PEFM needs a clear definition of the rules of the model, that should be obtained in a learning phase. In spite of the obviousness of such a statement, many ambiguities come out in defining what is a rule of the model. As far as a PEFM is concerned, the rules to be determined in the learning phase are not only the parameters of the algorithm, but also external parameters such as, for example, the definition of the area under study, and the choice of the threshold magnitude \((M_0)\) that defines a strong earthquake in that area.

Another important point is the way in which the learning phase has been conducted. Even though not strictly necessary (the only proper way to test the reliability of a model is through the voting phase), an objective and quantitative strategy used in the learning phase can provide further support to the reliability of the results of the model, especially when the voting dataset is not very large. In the two following paragraphs I will give a brief description of CN and M8, referring the interested reader for full formalization to Keilis-Borok et al. (1988), and Keilis-Borok and Rotwain (1990) regarding CN, and Keilis-Borok and Kossobokov (1990a) concerning M8.

6.6.1 Algorithm CN

CN is a code which attempts to characterize the time interval before a large earthquake. For each time interval, several variables that describe the seismicity in the area under study are measured, such as the seismic activity, quiescence, temporal rates of change of seismic activity, size of earthquakes, spatial and temporal clustering and long-range interactions between earthquakes (Keilis-Borok et al., 1988). The
acronym CN derives from the California and Nevada states, where the learning phase of the code was performed (Keilis-Borok et al., 1988). In the voting phase, the rules determined in the learning phase are used in the same area, as well as in many other regions of the world, to declare TIPs (Time of Increase Probability).

The CN algorithm calls for a large number of parameters to be fixed in the learning phase. Although this is not necessarily a problem, very few efforts were devoted, or at least reported, to verify the stability of the results as a function of the subjective choices made. The arbitrariness of the learning phase is further increased by ad hoc implementations. In particular, during the formalization of the CN algorithm, it was noticed that the code produced too many false alarms. In order to lower their number an additional rule was introduced for the declaration of TIPs (Keilis-Borok and Rotwain, 1990). Remarkably this new rule was not identified by the initial learning phase of the algorithm, and, surprisingly, it seemed somehow opposite to what the learning stage suggested.

More important, some parameters of the model are not determined in the learning phase, and therefore they are not kept fixed in the voting phase. These free parameters inevitably produce an overconfidence in the results of the voting phase. The choice of the area size is probably the most questionable in most of the applications of the CN code. In CN application to the Italian region, for instance, there is a series of works in which the area of forecast was progressively revised retrospectively, in order to improve the performance of CN (Keilis-Borok et al., 1990a; Costa et al., 1995; Costa et al., 1996; Costa et al., 1997; Peresan et al., 1999). Another free parameter was the number of features of a category that an object must possess in order to be classified as belonging to that category. Many works, in fact, display quite different results in the voting phase, depending on the different values used for this parameter (e.g. Keilis-Borok et al., 1988; Keilis-Borok et al., 1990a).

To summarize, in addition to a large arbitrariness in defining the
rules of the PEFM, I note that in past applications of the CN, the tuning of the free parameters was done by looking for a good retrospective forecasting system, i.e., the voting was performed, at least partially, on the same data as the learning (e.g., Gelfand et al., 1976; Caputo, 1987; Keilis-Borok et al., 1988; Keilis-Borok and Rotwain, 1990; Keilis-Borok et al., 1990a; Keilis-Borok and Kossobokov 1990a,b; Kossobokov et al., 1990; Gahalaut et al., 1992a,b; Costa et al., 1995; Costa et al., 1996; Costa et al., 1997; Romachkova et al., 1998; Peresan et al., 1999; Kossobokov et al., 1999b) and not on a new and independent data set as previously recommended (see above). This aspect precludes the possibility to check statistically the real forecasting capability of the algorithm and therefore it will not be further considered here.

6.6.2 Algorithm M8

The acronym M8 derives from the fact that this algorithm was originally designed to declare TIPs of occurrence of earthquakes with \( M_0 \geq 8 \). M8 considers spatio-temporal windows as objects of the analysis. In particular, it scans the area of forecast by several overlapping circles whose radius depends, through an empirical function, on \( M_0 \) (Keilis-Borok and Kossobokov, 1990a). Successively, for each circle, it takes into account seven variables (functions of time), describing the number of earthquakes above a certain magnitude that occurred inside the circle, the deviation of seismic activity from a long time trend, the concentration of main shocks, and the clustering of earthquakes.

The definition of these functions requires the choice of several arbitrary parameters, such as the magnitude threshold \( M_0 \), the coefficients governing the relationship between \( M_0 \) and the circle radius, the duration of the time intervals, and so on. In order to define the rules for the M8 code, no objective learning has been performed, or at least reported. Based on the experience of the researchers, it was established that when a sufficient, and arbitrarily chosen, number of functions
were above a fixed quantile (Keilis-Borok and Kossobokov, 1990b), a TIP had to be declared, for a period of 5 years. A further evidence of the subjectivity of the model is that most of the alarms identified by M8 are strongly dependent on the catalog used (Habermann and Creamer, 1994).

Besides a large amount of subjectivity in the learning phase, I note that an effort has been made to provide a more objective application of the code in the voting phase. However, it is also interesting to note that further variability has been introduced into the analysis by setting up a new algorithm called Mendocino Scenario (MSc) (Kossobokov et al., 1990). It requires the further scanning of each M8 circle subject to a TIP, into 265 smaller square cells. Inside each cell, the number of earthquakes with a magnitude above a certain threshold (often chosen as $M = 3$, see Kossobokov et al., 1990) is counted, in order to describe the seismic activity. It is then postulated that the strong earthquake forecast by the TIP is most likely to occur in the region where a sufficiently large cluster of squares experienced quiescence. In this respect, MSc acts as a further rule for M8 forecasts, and, as in CN, it is worth stressing that this accessory rule does not come out from a learning process as would be required by a standard PR procedure.

In any case, the forward application of the M8 code guarantee the possibility to check the forecasting capability of this PEFM.

### 6.6.3 Test of PEFM based on M8

M8 has been applied to many areas of the world. In recent years, it has been used in a forward application in the Circum-Pacific area. The results for the time period 1985-1997 were recently published (Kossobokov et al., 1997; Kossobokov et al., 1999a). Here I keep separated the time intervals 1985-1991 and 1992-1997 because a "real" forward test was made only for the latter period (cf. Kossobokov et al., 1997). The complete list of target earthquakes is reported in table 2 of the Kossobokov et al.'s (1999a) paper.

The test of the M8 forecasts proposed by Kossobokov et al. (1999a)
uses a Bernoulli trial scheme (equation (6.1)) in order to estimate the statistical significance of such forecasts. The authors estimated $\hat{\theta}$ through equation (6.2), by using the Poisson process in time, and, for the spatial domain, a nonparametric distribution built through a fit of twenty years of worldwide seismicity with $M \geq 4.0$. This procedure implicitly assumes that the occurrence in times of the seismic events is completely random, that two decades of seismicity (1963-1984) with $M \geq 4.0$ is a reliable estimation of the spatial distribution of the earthquakes with $M \geq 7.5$ (the REFM), and that the parameters of the spatio-temporal distribution are exactly known.

Actually, all these assumptions are strongly questionable, if not unacceptable. Here, since I do not have either the original REFM used by Kossobokov et al. (1999a), or the spatial and temporal distribution of the alarms, I provide an evaluation of the M8 algorithm based on the results published in Kossobokov et al. (1997, 1999a).

First, I note that the REFM used by the authors is clearly incorrect. In addition to being set up without consideration of past earthquakes of magnitudes comparable to the target events, the forecasting capability of the REFM proposed by Kossobokov et al. (1999a) is significantly lower than what is expected from a "correct" REFM. In the time interval 1985-1997, the REFM proposed by Kossobokov et al. (1999a) forecasts 8 out of 32 earthquakes with $M \geq 7.5$ and 2 out of 7 with $M \geq 8.0$, using a 42% spatio-temporal coverage of alarms ($\hat{\theta} = 0.42$). If this coverage of alarms is correct, there is a less than 5% probability of having $\leq 8$ successful forecast out of 32 target earthquakes, while a "correct" REFM should forecast, on average, about 13-14 earthquakes with $\hat{\theta} = 0.42$.

To illustrate this point, I apply the REFM described earlier consisting of a filtered map of the spatial occurrence of the earthquakes with $M_s \geq 7.5$ and depth $\leq 70$ km (see above). By taking the cells with the highest probabilities, whose cumulative probability is about 0.42 (see equation (6.3)), I forecast, for the same time interval, 15 out of 32 earthquakes with $M \geq 7.5$ and 6 out of 7 earthquakes with $M \geq 8.0$
\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\sigma & \delta & 0.00 & 0.05 & 0.10 & 0.15 & 0.20 \\
\hline
0.00 & 0.19 & 0.33 & 0.50 & 0.67 & 0.81 \\
     & (0.09) & (0.25) & (0.45) & (0.65) & (0.80) \\
\hline
0.05 & 0.21 & 0.34 & 0.50 & 0.66 & 0.80 \\
     & (0.13) & (0.27) & (0.48) & (0.63) & (0.78) \\
\hline
0.10 & 0.25 & 0.37 & 0.50 & 0.63 & 0.75 \\
     & (0.18) & (0.29) & (0.44) & (0.59) & (0.72) \\
\hline
0.15 & 0.29 & 0.39 & 0.50 & 0.61 & 0.71 \\
     & (0.20) & (0.30) & (0.42) & (0.54) & (0.66) \\
\hline
0.20 & 0.33 & 0.41 & 0.50 & 0.59 & 0.67 \\
     & (0.20) & (0.30) & (0.40) & (0.50) & (0.60) \\
\hline
\end{array}
\]

\[\theta\text{ estimated by Kossobokov et al. (1999a)} = 0.40 (0.13)\]

\[N = 19\]

\[n = 10(5)\]

\[M \geq 7.5\]

\[\text{Time interval: 1992-1997}\]

Table 6.2: Statistical significance \(\alpha\) of the PEFM based on the M8 algorithm. The target earthquakes are in the time interval 1992-1997 and have \(M \geq 7.5\). The parameter \(\delta\) is the bias in the estimation of \(\hat{\theta}\) (see equation (6.5)), \(\sigma\) is the uncertainty of this estimate, \(N\) is the number of target earthquakes, and \(n\) is the number of events forecast by the PEFM. The results for the Mendocino Scenario are reported inside the brackets.

(see figure 6.1). These numbers match the expected value (13-14) much better than the ones reported by Kossobokov et al. (1999a).

For the time period 1992-1997, by using \(\hat{\theta} = 0.40\) I correctly forecast 12 out of 19 earthquakes with \(M \geq 7.5\), and 4 out of 5 earthquakes with \(M \geq 8.0\). Note that these values are almost the same reported for the PEFM based on the M8 algorithm.

In order to evaluate the possible effects of the "incorrect" REFM and of its unavoidable uncertainties, I estimate the significance level \(\alpha\) of the PEFM based on the M8 algorithm by taking into account both the bias \(\delta\), and the uncertainty \(\sigma\) \((\sigma^2 = \text{var}(\theta))\) of \(\hat{\theta}\). For this purpose I use different reasonable values of \(\delta\) (see equation (6.5)) and \(\sigma\). The results are reported in tables 6.2 and 6.3. I consider only the time
<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>0.00</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>(0.01)</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>0.05</td>
<td>(0.01)</td>
<td>0.01</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>0.10</td>
<td>(0.01)</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>0.15</td>
<td>(0.03)</td>
<td>0.03</td>
<td>0.05</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td>0.20</td>
<td>(0.05)</td>
<td>0.05</td>
<td>0.07</td>
<td>0.10</td>
<td>0.13</td>
</tr>
</tbody>
</table>

$\theta$ estimated by Kossobokov et al. (1999a) = 0.36 (0.18) 
$N = 5$
$n = 5(4)$
$M \geq 8.0$

Table 6.3: The same as for table 6.2, but relative to target earthquakes with $M \geq 8.0$.

interval 1992-1997 because it is the only period for which a real forward analysis has been performed (see Kossobokov et al., 1997). The results show that the capability of M8 to forecast $M \geq 7.5$ earthquakes is not statistically significant. This means that the M8 algorithm does not work better than the REFM. The same is valid if I take into account the Mendocino scenario (see table 6.2). If I consider the results reported by Kossobokov et al. (1999a) for the time interval 1985-1997, the forecasting ability of M8 (with and without the Mendocino scenario) is not statistically significant at the 0.99 confidence level, by using $\delta \geq 0.05$ and/or $\sigma \geq 0.05$. The same conclusions can be reached heuristically by looking at the forecasting capability of the REFM proposed above.

As regards the earthquakes with $M \geq 8.0$ (see table 6.3), the forecasting ability of M8 (with and without the Mendocino scenario) is not statistically significant at the 0.99 confidence level (i.e., $\alpha \leq 0.01$)
in either of three cases: (1) $\delta \geq 0.10$; (2) $\sigma \geq 0.15$; (3) $\delta \geq 0.05$ and $\sigma \geq 0.10$. Note that the bias $\delta$ is smaller than what I have found before (see table 6.1) by comparing the "correct" REFM proposed here, and an "incorrect" REFM based on seismicity with a lower threshold, like the one suggested by Kossobokov et al. (1999a). As far as $\sigma$ is concerned, I am not able to provide a realistic estimation from the data, because I do not have the REFM used by Kossobokov et al. (1999a). I think, however, that $\sigma \approx 0.10$ might be considered as a reasonable choice to properly account for uncertainties and possible weak nonstationarity of the process. The results reported in tables 6.2 and 6.3 are stable if I use a beta distribution for $\theta$ in equation (6.6), or a Gaussian distribution for $u$ in equation (6.10).

In order to facilitate the comparison of the results in tables 6.2-6.3 with those provided by Kossobokov et al. (1999a), I note that in the latter (see also Keilis-Borok, 1999; Rotwain and Novikova, 1999), there is an improper use of the terms confidence level (C.L.) and significance level ($\alpha$). Specifically, they were switched in meaning while they should be related as

$$C.L. = 1 - \alpha.$$  \hspace{1cm} (6.20)

Their meaning is only related to a rejection of a null hypothesis (e.g., Kalbfleisch, 1985). Here, I have used the critical value of $99\%$ of confidence level because it is a convenient choice for two main reasons: first, the higher the scientific and practical relevance in rejecting a null hypothesis is, the higher the confidence level has to be. Second, if the test is performed with few data, the inclusion or removal of each single datum may drastically change the test statistics. In this case, a conservative choice is to select a high confidence level to reject the null hypothesis.

The M8 performance displayed in tables 6.2 and 6.3 agree with the results obtained by Minster and Williams (1998) through an independent and systematic test of M8. According to their study, in which the worldwide seismicity is used to compute the Receiver Operating
Characteristic (ROC) curve, there is no difference, at a 95% confidence level, between the performance of M8 and of a random forecast based on a uniform pdf.

### 6.7 Final remarks of the chapter

The main purpose of this chapter was to discuss the set up and test of an earthquake forecasting model. I provided some new insights that can help to evaluate quantitatively the performance and reliability of any forecasting model. I have also applied this strategy to check the forecasting ability of two Pattern Recognition based algorithms, M8 and CN.

In particular, I have emphasized that many debates on earthquake prediction result from lack of clear definitions of key concepts (such as the term precursor) and the strategies needed to set up a forecasting model and to evaluate its reliability. For this purpose, I have proposed a classification of earthquake forecasting models based on the kind of information used to set them up. I have noted that the forecasting capability of any model based on the observation of supposed precursors (PEFM) must be compared to the forecasting ability of a "correct" reference model (REFM) that describes the knowledge of the spatio-temporal occurrence of the target events. I have shown that the use of an "incorrect" REFM very often leads to an overestimation of the forecasting capability of the PEFM. Overestimation also results from not properly taking into account random uncertainties in the REFM estimation.

Finally, I have applied these concepts to two forecasting models based on Pattern Recognition. I have suggested that the forecasting capability of the CN code cannot be objectively tested. In the case of M8 algorithm, I have shown that its forecasting capability as estimated by previous authors is much too optimistic, because the method used to check its performance was based on unrealistic assumptions, and the use of an "incorrect" REFM. Taking into account possible biases and
uncertainties introduced in the REFM as used by the M8 authors, I find that the forecasting capability of M8 is not statistically significant. Indeed, the successful forecast of M8 are comparable to those obtained by using a more suitable REFM.
Chapter 7

Final remarks

In this Ph.D. study I have used a multivariate statistical approach, based on *Pattern Recognition* algorithms, to characterize some relevant Geophysical processes. First of all, I have implemented four non-parametric Pattern Recognition techniques, named *K-Nearest Neighbors, Binary Decision Tree, Fisher’s Analysis* and *Linear Discriminant Analysis*. I have evaluated quantitatively their performance on synthetic data that I have generated. The synthetic data mimic some of the basic properties of Geophysical datasets, i.e. they could consist of few data, often correlated and/or discretized, often not normally distributed, and containing measurements relative to variables that have no influence on the process. The evaluation of the performance of the algorithms, before applying them to real data, represents a fundamental step in order to correctly apply a statistical method of analysis. By means of the results of the simulated applications, I have been able to select 2 algorithms, *Binary Decision Tree* and *Fisher’s Analysis*, as the most suitable to be applied to Geophysical datasets similar to those I have.

Successively, I have applied these two algorithms to two volcanological datasets. First, I have analyzed a dataset of seismic swarms recorded during unrests in different volcanic areas of the world, aiming to find possible repetitive and common patterns among pre-eruptive seismicity compared to isolated seismicity. The results of the analysis show that there is, indeed, a different seismic energy release between
pre-eruptive swarms and isolated ones, because the former last longer. On the opposite, there is no such difference between seismicity preceding eruptions of different magnitude. This means that the energy released in the seismic swarm is not indicative of the magnitude of the impending eruption. My analysis has also highlighted that, before large explosive eruptions, the repose time elapsed from the end of the previous eruption is longer, compared to smaller eruptive episodes. This might give us an indication of the need of the system to re-charge and close up the conduit, before being able to give a large explosive eruption.

Secondly, I have analyzed a dataset of seismic clusters recorded in the Etna area in recent decades (from 1974 to 2001). The goal of this second application was to identify a possible link, suggested by previous works on this area, between the tectonic stress regime and the occurrence of flank eruptions at Mount Etna volcano. I have also analyzed the same dataset by means of another multivariate statistical technique, i.e., a multivariate regression fit. The results of both analyses do not highlight any significant link between seismicity, assumed as a tectonic stress indicator, and the occurrence of flank eruptions at Mount Etna volcano. I have found a link between the occurrence of flank eruptions and the repose time and volume output of the previous flank eruption. Again, this might represent an evidence of the need of Mount Etna feeding system to re-charge, before giving another flank eruptive episode.

Finally, I have critically reviewed the two most used Pattern Recognition algorithms applied in Geophysics, called M8 and CN. I have proposed a statistical formal procedure to validate a generic prediction algorithm, and have applied it to specifically estimate the statistical significance of the only forward prediction experiment currently conducted to forecast earthquakes with magnitude \( M \geq 7.5 \), based on M8 algorithm. The results show that the forecast by M8 are not statistically significant, because they perform similarly, or even worse, to a model based on the past occurrence of seismic events of similar
magnitude.
Chapter 8

References


Dziewonsky AM, Woodhouse JH (1983): An experiment in systematic


Gahalaut VK, Kuznetsov IV, Kossobokov VG, Gabrielov AM, Keilis-Borok VI (1992b): Application of pattern recognition algorithm in the


Gottsmann J, Rymer H (2002): Deflation during caldera unrest: constraints on subsurface processes and hazard prediction from gravity-


Minster JB, Williams N (1998): Systematic global testing of intermediate-
term earthquake prediction algorithms. *1st ACES Workshop Proceedings.*


*Nature Debate* (1999): Is the reliable prediction of individual earthquakes a realistic scientific goal?,

[http://helix.nature.com/debates/earthquake/](http://helix.nature.com/debates/earthquake/)


Newhall CG, Hoblitt RP (2002): Constructing event trees for volcanic


www.wovo.org/wovodat.htm

http://volcano.und.nodak.edu

http://www.volcano.si.edu

http://vulcan.wr.usgs.gov

http://www.nmnh.si.edu/gvp/volcano
Appendix A

K-Nearest Neighbors

In the space of the features, suppose to have a cell surrounding the object \( \mathbf{x} \) that has to be classified, and containing \( k \) objects. Suppose that, for each class \( i \), \( k_i \) objects of class \( i \) are contained in the cell. The k-nearest neighbors method assigns the objects \( \mathbf{x} \) to the class most largely represented in the cell, i.e. \( \mathbf{x} \) is classified in class \( j \) where \( j \) is such that

\[
  k_j > k_i \quad \forall \text{ class } i \neq j
\]
Appendix B

Binary Decision Tree

This method has been developed by Rounds (1980) and, slightly modified, successfully applied to volcanic data by Mulargia et al. (1992). It can be used only in the 2-class problem and it was originally designed for hierarchically ordered datasets, even though tests on synthetic data have shown a very good behavior also on different types of datasets.

Once the data have been collected, and objects and classes have been defined, BDT integrates feature selection and binary decision tree, according to the following steps:

(1) Fix a level $\alpha$ for the decision rule. This level represents the risk we accept of a wrong attribution at each step. We use $\alpha = 0.1$.

(2) Compute the cumulative distribution in both classes, for each feature, taken one at a time, and identify the feature, and the relative threshold value, for which the statistical difference between the cumulatives of the two classes is the largest. This means that the significance level of this statistical difference must be (a) lower than the level $\alpha$ and (b) minimum. The feature (if any) for which both (a) and (b) conditions are satisfied is the first-order feature, often called the root of the pattern. On the basis of the root feature and its threshold value, each object is assigned to either one of two subsets, formed respectively by data with a value of the root feature lower/higher than the threshold.

(3) Identify the second-order features, and their thresholds, for which
the statistical difference satisfies again (a) and (b) conditions. These features are found by reanalyzing all the features as in (2) separately in the two subsets, and are at most two (i.e., one for each subset).

(4) Repeat step (3) from each second-order feature, in order to identify progressively higher orders, as long as it is possible to find a feature for which the cumulatives in the two classes are statistically different at a significance level lower than \( \alpha \). The progressive branching of the tree gives all the possible patterns. The procedure automatically terminates when no further branching is possible at the given level \( \alpha \).

Steps (2)-(4) are performed by means of a nonparametric Kolmogorov-Smirnov two sample statistics (Hollander and Wolfe, 1973). Note that the use of an \( a \) priori fixed level \( \alpha \) reduces the possibility to obtain overfitting patterns.
Appendix C

Fisher’s Discriminant Analysis

This method (see e.g. Duda and Hart, 1973) is based on the reduction of the $n$-dimensional space of the objects (where $n$ is the number of variables describing the objects, i.e. the dimension of the vectors) to an $L - 1$ dimensional space (where $L$ is the number of classes). In our two-class problem ($L = 2$), Fisher’s method simply projects the objects onto a line. The basic idea, called Fisher criterion, is to project the objects onto the direction that maximizes the ratio of the dispersion between the classes to the dispersion within the classes. More rigorously, let us suppose to have $N$ objects $\mathbf{x}$, each represented by a vector consisting of $n$ components $x_k$ ($k = 1,...n$). Of these, $N_i$ belong to the subspace $H_i$, i.e. they belong to Class $i$. We linearly combine the components of $\mathbf{x}$, i.e. the $x_k$ ($k = 1,...n$), in order to obtain a $(L-1)$-dimensional vector $\mathbf{y}$, consisting of $(L-1)$ components $y_j$:

$$y_j = W_{kj}^T x_k \quad j = 1,...L - 1 \quad (C.1)$$

where $W_{kj}$ are the elements of an $n \times (L - 1)$ matrix that projects $\mathbf{x}$ onto $\mathbf{y}$. In this way, we obtain $N$ objects $\mathbf{y}$, each represented by a vector of dimension $L - 1$, spread over $L$ subspaces $F_i$.

The unknown in equation C.1 is the projector, i.e. the matrix $W$. As mentioned above, we would like to choose the projection for whom the ratio of the dispersion between the classes to the dispersion within
the classes is maximum. In order to do this, first we need to define some quantities.

We define \( \mathbf{m}_i \) as the average vector of Class \( i \):

\[
\mathbf{m}_i = \frac{1}{N_i} \sum_{x \in H_i} \mathbf{x}
\]

We also define \( \mathbf{m} \) as the average of all the \( \mathbf{x} \):

\[
\mathbf{m} = \frac{1}{N} \sum_{x} \mathbf{x}
\]

Thus, the dispersion matrix within the subspace \( H_i \) is given by:

\[
S_i = \sum_{x \in H_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T
\]

and the dispersion within all of the subspaces is

\[
S_w = \sum_{i=1}^{L} S_i
\]

The total dispersion matrix is given by

\[
S_T = \sum_{\mathbf{x}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T
\]

It follows that

\[
S_T = \sum_{i=1}^{L} \sum_{x \in H_i} (\mathbf{x} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m})(\mathbf{x} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m})^T
\]

\[
= \sum_{i=1}^{L} \sum_{x \in H_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T + \sum_{i=1}^{L} \sum_{x \in H_i} (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T
\]

\[
= S_w + \sum_{i=1}^{L} N_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T
\]

The second addendum of the right side term of equation C.7 is a dispersion matrix \( S_b \) that gives an idea of the dispersion between the partial means \( \mathbf{m}_i \) over the subspaces \( H_i \) and the total mean \( \mathbf{m} \):

\[
S_b = \sum_{i=1}^{L} N_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T
\]
In order to achieve the matrix $W^*$ that maximizes the ratio of the $S_b$ to the $S_w$, we need to project these matrixes onto the $y$ space and compute the $W^*$ such that:

$$\frac{|W^{*T}S_bW^*|}{|W^{*T}S_wW^*|} = \max_{w} \frac{|W^{T}S_bW|}{|W^{T}S_wW|}$$  \hspace{1cm} (C.9)

Once the maximization has been carried out, Fisher’s analysis projects the $x$ vectors onto the $y$ space. Then, each object $y$ is assigned to the class $i$ whose mean $m_i$ is closest to $y$. 

135
Appendix D

Linear Discriminant Analysis

In this type of analysis it is assumed that the decisional boundaries separating the different classes are of a known shape, i.e. they are hyperplanes. It is still a nonparametric method because, although an assumption is made on the boundaries delimiting the classes, no assumption is made on the distribution of the objects. It is a particularly efficient method to classify objects governed by unimodal distributions.

The unknown parameters of the classification problem are the hyperplanes. Their coefficients should be determined by minimizing some kind of criterion. The most obvious one is the minimization of the classification error, but this is usually too difficult. Because of this, simpler criteria are usually taken. For reasons of clarity, we will now concentrate on the 2-class problem. In this case, the boundary between the two classes can be written as

$$g(x) = w^T x + w_0$$  \hspace{1cm} (D.1)

where $w$ is the vector of the unknown coefficients (weights) of the hyperplane and $w_0$ is the threshold weight. The classification is immediate: a given object $x_0$ is assigned to class 1 if $g(x_0) > 0$, to class 2 otherwise. The hyperplane $g(x) = 0$ defines the boundary between the two classes.

We will concentrate only on the linear boundary (from this, the name of the technique), although more sophisticated boundaries are possible.
In order to determine the unknown parameters, we can rewrite equation D.1 as
\[ g(x) = a^t y \] (D.2)
where
\[ a = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \] (D.3)
and
\[ y = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \] (D.4)

There is a number of different ways to find the unknown vector \( a \). We will focus on the least square solution, i.e. a solution to the system
\[ Y a = b \] (D.5)
where \( Y \) is a matrix consisting of the row vectors \( y_i^t \) and \( b \) is an arbitrary vector. There are several reasonable ways to choose \( b \); we will here focus on the so-called “stochastic approximation” method (Duda and Hart, 1973). According to this approximation, the vector \( b \) is a binary vector in which its \( i \)-th component is equal to 1 if the object \( x_i \) belongs to class 1, while it is equal to -1 if \( x_i \) belongs to class 2. In this way, from equation D.5, we have that
\[ Y^t Y a = Y^t b \] (D.6)
so that the sought unknown vector is given by
\[ a = (Y^t Y)^{-1} Y^t b \] (D.7)
Any object \( x_i \) is classified in class 1 if \( g_1(x_i) > g_2(x_i) \), in class 2 if \( g_1(x_i) < g_2(x_i) \).
Appendix E

Branch-And-Bound Technique

This technique (see e.g. Fukunaga, 1990) allows to select the subset of relevant features among those available. In fact, given $n$ features for each object, apart from few statistical PR algorithms (e.g. BDT) that automatically provide the subset of features by which the classification is carried out (named optimal subset), most of the statistical PR algorithms just perform the pattern recognition and the classification of the objects, but do not explicitly provide the optimal subset. The basic concept in the selection of the optimal subset of features is to find, among all possible subset of the $n$ features, the one leading to the lowest classification error and consisting of the smallest number of features. In such situation, we are confident that we are considering all of the important variables (otherwise the classification error would not be the lowest) and we are excluding the irrelevant ones (otherwise the number of features in the optimal subset would not be the smallest).

A simple, but very time consuming, way to find such an optimal subset consists of exploring the performance of the chosen statistical PR algorithm on all the possible subsets of the $n$ features. This becomes prohibitive as $n$ increases, since we have to explore $\sum_{k=1}^{n} \binom{n}{k}$ subsets. In order to avoid the application of the chosen statistical PR algorithm to all the possible subsets of features, the branch-and-bound
technique has been developed. This technique is applied iteratively $n$ times; at each iteration $k$ ($k = 1, \ldots, n$), it allows the identification of the sub-optimal subset consisting of $k$ features, by applying the statistical PR algorithm only on the “most promising” subsets of $k$ features. The sub-optimal subset is then the one consisting of $k$ features and leading to the lowest classification error.

The branch-and-bound method relies on a basic assumption, i.e., it assumes that the noise introduced by irrelevant features does not deteriorate the signal given by the relevant features. In a previous study we have tested the validity of this assumption for algorithms BDT and FIS. Based on this assumption, when a certain subset of $k$ features does not produce a good discrimination rule, the branch-and-bound method assumes that any other subset of $k + l$ ($l = 1, n - k$) features containing those $k$ features will not be the optimal one. In this way, a considerable portion of all the possible subsets is discarded \textit{a priori}, thus saving computation time and effort.