Università degli Studi di Bologna

Facoltà di Scienze Matematiche Fisiche e Naturali
Dipartimento di Fisica

Dottorato di Ricerca in Geofisica

XVII Ciclo

Analysis of the spatio-temporal distribution of large earthquakes

Tesi di: **Licia Faenza**

Tutore: **Prof. Warner Marzocchi**

Referente: **Prof. Paolo Gasperini**

Coordinatore: **Prof. Maurizio Bonafede**

Bologna, Marzo 2005

Contents

Chapter 1

Introduction

The investigation on the spatio-temporal distribution of large earthquakes is still a controversial issue in geophysics and many works in scientific literature have been devoted to this topic. The importance of understanding the statistical distribution of large events is aimed not only to extract information on the physics of the earthquakes occurrence process, but also to make reliable earthquake forecasting. As far as theoretical aspects are concerned, a satisfactory modelling may allow, at least in principle, to test a variety of hypotheses, such as the presence of any regularity in time, and the influence of different tectonic/physical factors that regulate the spatial occurrence of earthquakes. At the same time, a reliable earthquake forecasting has undoubtedly a huge social impact because it may mitigate the seismic risk.

Even if many papers in the past have been focused on studying the distribution of large earthquakes (e.g. Vere-Jones, 1970; Shimazaki & Nakata, 1980; Nishenko, 1985; Boschi *et al.*, 1995; Ellsworth *et al.*, 1998; Ogata, 1998; Kagan & Jackson, 2000; Stock & Smith, 2002; Posadas *et al.*, 2002, and many other references therein), a statistical distribution of general consensus could not be reached so far, and the results obtained seem to be sometimes contradictory. On the contrary, for events of small magnitude, the Epidemic Type Aftershocks-Sequences (ETAS) model (e.g. Ogata, 1988) is widely accepted from the scientific community as a toll to model the spatio-temporal distribution of events.

The factors that may contribute to such difference are the magnitude

threshold and the investigated spatial scale. Concerning the distribution of large earthquakes, it happens frequently that only the temporal domain is investigated, by selecting small seismic areas where the hypothesis of spatially homogeneous sampling probably holds. Unfortunately, these small areas do not usually contain a sufficient number of earthquakes to adequately test any hypothesis and model (cf. Jackson & Kagan, 1993). For this reason, quite different distributions, such as Poisson (e.g. Kagan & Jackson, 1994), Poisson generalised (e.g. Kagan, 1991), Brownian Passage Time (Ellsworth *et al.*, 1998), Weibull (Nishenko, 1985), lognormal (e.g. Nishenko & Buland, 1987; Michael & Jones, 1998) distributions, or competitive models, such as general seismic gap hypothesis (McCann *et al.*, 1979), time-predictable model (Shimazaki & Nakata, 1980; Papazachos, 1992), clustering of earthquakes (e.g. Kagan & Jackson, 2000) are commonly used in hazard studies.

The current scarce knowledge about this topic is well highlighted by the fact that, in some cases, these antithetical models are simultaneously used (e.g. Working Group on California earthquake probabilities, 1999). Remarkably, the differences between these models are not only of statistical nature, but imply opposite physical mechanisms for earthquake occurrence. To sum, the rationale under the choice of any model is very often based only on theoretical assumption or on a personal subjective belief.

The problem in defining the distribution of earthquakes is further complicated by the ambiguous definition of terms such as mainshock-aftershock, and "large" earthquake. The latter, in particular, seems strongly related to the tectonic region considered. In Italy, for instance, earthquakes with $M \geq 5.5$ are usually defined "large" and they are considered independent one from each other (e.g. Boschi *et al.*, 1995; see also below), while in the Pacific Ring many of such events are clear aftershocks of larger events (e.g. Kagan & Jackson, 2000).

The aim of this thesis is to provide an analysis method to characterise the spatio-temporal distribution of large earthquakes. We use a multivariate nonparametric hazard model that presents several and im-

portant theoretical and technical advantages with respect to more traditional approaches. This model is a representation of the time-behaviour of earthquake occurrence, and it is convenient, flexible and yet entirely empirical.

The subject of this work is the definition of a model that can relate the hazard function and some explanatory variables. Typically such explanatory variables describe the sample under study or its domain. The hazard function is a statistical function that is very useful in survival analysis. It specifies the instantaneous rate of occurrence at a time $T = t$, conditional upon survival to time t. Since the hazard function describes unequivocally the statistical distribution of the events (see Appendix A), the knowledge of the hazard function means the knowledge of the distribution of the point process, but its empirical trend as a function of time provides an immediate tool to rule out some of the aforementioned distributions.

Moreover, the hazard function gives an immediate answer to the famous Davis et al.'s question (1989):

"Can it be that the longer it has been since the last earthquake, the longer time till the next?"

The answer to this question essentially depends on the statistical distribution of the inter-event times (i.e. the time between two consecutive events), as also shown by Sornette & Knopoff (1997). An increasing trend of the hazard function would provide an answer "no" to such question. On the other hand, a negative trend would lead to an answer "yes". In geophysics, an increasing trend would stand for an increasing in the probability of occurrence with the time elapsed since the most recent event. This case would support, in general, the seismic gap hypothesis (McCann *et al.*, 1979; Nishenko, 1985) at least for a time interval comparable to the mean recurrence time. On the other hand, a decreasing trend of the hazard function would stand for a "clustered" sequence, where the probability of occurrence decreases with the time elapsed since the most recent event. This is the case of the ETAS model (Ogata, 1998). A constant hazard function characterises the Poisson process; in such case the

conditional probability to have an earthquake in a fixed time interval is "time independent", i.e. it is independent from the time elapsed since the most recent event (i.e., Cornell, 1968).

In addition to these three simple models, there are distributions with more complicated trend of the hazard function. For instance, a statistical distribution often used in earthquake forecasting is the lognormal distribution; it presents a hazard function increasing until a certain maximum, representing the mean of the distribution; after that the hazard rate is gradually decreasing up to go to zero. This suggests that the mean inter-event time is the most probable interval between the most recent event and the next one. The conditional probability thus increases before approaching the mean value till it reaches the maximum. If the earthquakes have not occurred near the mean inter-event time, however, the hazard rate becomes decreasing. This represents the behavior of the "characteristic earthquakes" or "seismic cycle". (i.e., Musson *et al.*, 2002; Schwartz & Coppersmith, 1984). In Appendix B a list of the most common and used statistical distributions in earthquake forecasting is listed, underlining the characteristics that make them appealing in seismology.

The model we propose is called Proportional Hazard Model (PHM hereinafter) and it was first introduced in literature by Cox, 1972. The basic idea is that, on each individual under test, values of one or more explanatory variables are available. The hazard function is taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time (Kalbfleisch & Prentice, 1980):

$$\lambda(t, \mathbf{z}) = \lambda_0(t) \exp(\mathbf{z}\boldsymbol{\beta}) \tag{1.1}$$

where $\lambda_0(t)$ is an arbitrary and unspecified base-line hazard function, $\mathbf{z}$ is the vector of the explanatory variables and $\boldsymbol{\beta}$ is a vector of coefficients that represents the statistical weigh of each component of $\mathbf{z}$.

Several are the advantages of this technique compared to a more traditional one. First, the investigation of the temporal domain is completely nonparametric. This means that we do not impose any *a priori* statistical distribution of the events, but we leave the data "speak" by

themselves. The study of $\lambda_0(t)$ is therefore completely empirical and represents the temporal dependence of the model. Second, it is possible to consider information of different nature that may influence the occurrence of further events because of the explanatory variables, or vector of covariates, that bear physical, geological or tectonic factors related to the events or the spatial area where those are sampled. In particular, the Maximum Likelihood Estimation of the regression coefficient, or vector of parameters, allows a direct check of the statistical importance of each covariate. This means that, in the study of the spatio-temporal distribution of events, it is possible to consider different factors and therefore to merge, in a formally correct manner, different pieces of information. Lastly, the technique is robust because it simultaneously considers all the spatially inhomogeneous data to build the statistical model: the analysis is hence performed using a consistent number of data, and this allows a forward check of the model.

The random variables of the system are the time interval between two earthquakes and the censoring time, i.e., the time interval between the most recent event in the catalogue and the end of the catalogue itself. The censoring times carry out a certain amount of information on the process of earthquake occurrence and are not usually considered in more traditional approaches. These kinds of data usually greatly complicate the distribution theory for the estimation, but they are very important in time-dependent analysis, where the probability of the next event depends on the time elapsed since the previous one.

At the same time, the technique is very flexible, being based on a small number of mild assumptions, two of which are fundamental. First, the time-dependence of $\lambda(t; \mathbf{z})$ is the same for all the areas, only the parameters can vary. The vectors $\mathbf{z}$ and $\boldsymbol{\beta}$ act multiplicatively on the base-line hazard function $\lambda_0(t)$ and they can not modify the time behaviour of the hazard function. Second, for the spatial domain we are assuming that the earthquakes will occur in the same areas as in the past, and that the spatial coverage of the earthquakes reported in the analysed catalogue is representative of the spatial distribution of the events. The strategy of

the PHM and detailed description of its characteristics are reported in chapter 2.

The PHM has been applied to the seismic catalogue of strong earthquakes in Italy in the last centuries and to the world seismicity. For the investigation of the spatial domain of the Italian territory, two different kinds of grid have been used: first a regular, geometrical grid (see chapter 3) and, in a second step, a grid based on a seismo-tectonic data (see chapter 4). We have also investigated the properties of the temporal dependence of the hazard function found for the Italian seismicity (see chapter 5). The results obtained applying the PHM to a regular grid in the world have been reported in chapter 6.

In the application of PHM to the Italian catalogue, the seismicity of the last four centuries is investigated. In order to have a complete set of events we consider the earthquakes with M$\geq$ 5.5 since 1600, according to Mulargia & Gasperini, 1995, that adopts a new statistical strategy to evaluate the catalogue completeness proposed by Mulargia & Tinti, 1985. The threshold magnitude has also been chosen taking into account practical aspects. In fact, M= 5.5 generally represents a threshold value for which the earthquakes in the Italian territory are considered "dangerous" due to the high number of historical buildings. Previous papers assume that similar events are almost independent one from each other (e.g. Boschi *et al.*, 1995). The first important aspect that focuses our attention is the trend of the hazard function as a function of time (see figure 3.4 in chapter 3 and figure 4.6 in chapter 4). For both the two grids, the hazard function has a decreasing trend versus time. This means that the probability of having an earthquake is higher immediately after an event and decreases with time increasing since the previous one. The cluster length is a few years, a longer time window than the one for the aftershocks sequences, usually taken of three months for Italy (see CPTI Working Group 1999, 2004). Then the trend becomes almost flat as expected for a Poisson process. The presence of a significant departure from the Poisson model implies that the statistical distribution is time-dependent, i.e. the probabilities of occurrence depend on the time

elapsed since the most recent event. Note that there is no significant evidence of a positive trend. A possible explanation of this time behaviour can be found in a viscoelastic coupling between faults (Piersanti *et al.*, 1997; Pollitz *et al.*, 1998; Freed & Lin, 2001; Marzocchi *et al.*, 2003B).

The vector of the covariates is, for the regular grid, a two dimensional vector. Only its first component, the rate of occurrence of the area, is statistical significant, while its second component, the magnitude of the previous event, is not. In other words, the magnitude of the last event does not seem to significantly modify the probability to have another large event. From a physical point of view, this means that there is no evidence in favor of some kind of "time predictable" model (e.g., Shimazaki & Nakata, 1980; Papazachos & Papadimitriou, 1997). This result is also in contradiction with models characterised by a time behaviour that somehow depends from the magnitude of the earthquakes (e.g. Ogata, 1988; Kagan & Jackson, 2000). Remarkably, this may lead to suggest that the Gutenberg-Richter law does not always hold, at least for large events occurring in small spatio-temporal windows.

The usage of a grid based on seismo-tectonic data allows testing a variety of factors. The setting of the grid and the definition of the geological and tectonic parameters of the model have been done in collaboration with Dr. Francesca Cinti and Dr. Paola Montone of Istituto Nazionale di Geofisica e Vulcanologia (INGV), Rome. The grid, or "regionalization", has been designated with the purposes to define areas that are homogeneous with regard to seismic behaviour and to kinematics and orientation of the stress field. The vector of the covariates has therefore seven components. Besides the rate of occurrence and the magnitude, the prevalent stress regime within each zone, the homogeneity of the stress orientation, the number of seismogenetic faults, the homogeneity of the topography and the extent of the zone have been tested. Likewise for the regular grid, the rate of occurrence seems to be the only important covariate (among the ones taken into consideration) in modelling the spatio-temporal distribution of large earthquakes. As a further check on the importance of each covariate, we also the program by using in-

dividual covariate one at a time, and, in this case, also the number of active faults appears statistically significant. The fact that the number of active faults is important only when considered alone indicates that the rate of occurrence may be linked not only to the tectonic rate, but also to the average number of active faults in the area.

One of the principal goals in applying PHM to the Italian territory has been providing a time-dependent probability map of occurrence for the next moderate to large earthquakes in Italy. In both cases, the most likely regions where the next large earthquakes may occur are Friuli, Umbria -Marche, and part of Southern Apennines and Calabrian arc (see figures 3.6 and 3.7 in chapter 3 and figure 4.8 in chapter 4). A validation test on independent data has been done, showing that the model fits well the data, see figures 3.4 in chapter 3 and 4.7 in chapter 4.

In chapter 5 we have performed an analysis of the time clustering property of large earthquakes in Italy. In particular, a comparison between the hazard function found for a specific ETAS model for aftershock sequence for the Italian seismicity and the one obtained applying PHM to the real Italian seismicity has been done using a two-sample Kolmogorov-Smirnov test. This part of the work has been made in collaboration with Dr. Anna Maria Lombardi and Dr. Rodolfo Console of INGV, Rome. The results show that the cluster imposed by the ETAS model is weaker than the real one of large events, and it may be due to the fact that the interaction between earthquakes lasts longer with respect to what imposed by the specific ETAS model used.

In chapter 6, PHM is applied to the global seismology, considering the events with M$\geq$ 7.0 occurred from 1900 until 2004. For the spatial domain, a regular grid has been applied. All the areas with at least one event have been taken into account in the analysis. The vector of the covariates is a three dimensional vector, compounded by the rate of occurrence of the area, the magnitude of the previous event and the average tectonic regime of the area. As in previous cases only the rate of occurrence of the area is statistically significant. The magnitude of the previous event seems to be a factor that does not modify the prob-

ability of occurrence of a next event, even in case of a large one. This is very important for understanding the role played by the magnitude in earthquakes occurrence, and it neglects any kind of "time predictable" model for the spatio-temporal distribution of large events M$\geq$ 7.0 (e.g. Papazachos, 1989; Papazachos & Papaioannou, 1993).

Figure 6.4 in chapter 6 shows that the spatio-temporal distribution of large earthquakes is a cluster for a time window that reaches few years after the event. This means that the probability of having a strong event is higher immediately after a previous strong event. This result is in contradiction to the seismic gap hypothesis, which assumes that recent strong earthquakes inhibit further one. A time-dependent probability map is reported in figures 6.6 and 6.7 for the next one and ten years, respectively, showing that the most dangerous areas are the Kurile Island, the Solom Islands and the western cost of Mexico.

The works presented in this thesis have yielded to the edition on international journals of four articles, three puplished and one in preparetion

- Faenza L., Marzocchi, W. & Boschi, E., 2003. A non-parametric hazard model to characterize the spatio-temporal occurrence of large earthquakes; an application to the Italian catalogue, *Geophys. J. Int.*, **155** (2), 521-531.

- Cinti F.R., Faenza, L., Marzocchi, W. & Montone, P., 2004. Probability map of the next M$\geq$ 5.5 earthquakes in Italy, *Geochem. Geophys. Geosyst*, **5**, Q11003, doi:10.1029/2004GC000724.

- Faenza, L., Marzocchi, W., Lombardi, A.M. & Console, R., 2004, Some insights into the time clustering of large earthquakes in Italy, *Annals of Geophys.*,**47**, 1635-1640.

- Faenza, L., Marzocchi, W., Serretti, P. & Boschi E. On the spatio-temporal distribution of large earthquakes M$\geq$ 7.0, in preparetion.

Chapter 2

Proportional Hazard Model

Here we model the spatio-temporal occurrence of earthquakes through a nonparametric multidimensional fit of the hazard function. In particular, we use the Proportional Hazard Model (hereinafter PHM) described by Cox (1972) and Kalbfleisch & Prentice (1980).

As there is a biunivocal correspondence between the hazard function and either the probability density function or the cumulative and survivor functions (e.g., Kalbfleisch, 1985), the hazard function defines unambiguously the statistical distribution of the inter-event times (see details in Appendix A). Nevertheless, the hazard function is often preferable because a simple analysis of its trend can provide useful insights on the physics of the process. For example, some of the most used statistical distributions (i.e., the Weibull distribution) were built by imposing a parametric form (a power law) for the hazard function.

Basically, the model deals with two kinds of random variable (from now on RV), namely the inter-event time between two consecutive earthquakes (from now on IET), and the censoring time (from now on CT), i.e., the time interval between the present time and the last earthquake occurred. At each one of these RVs a vector of explanatory variables (from now on covariates) that carries out any kind of information relative to the IET or CT considered has been attached. In particular, in the analysis of a seismic catalogue, the vector of covariates can contain any spatial/tectonic information on the subregion where the IETs and/or CT are sampled.

The PHM (Cox, 1972; Kalbfleisch & Prentice, 1980) for a generic time $t^\star$ since the most recent event can be written as

$$\lambda(t^\star; \mathbf{z}) = \lambda_0(t^\star) \exp(\mathbf{z}\boldsymbol{\beta}) \tag{2.1}$$

where $\mathbf{z}$ is the vector of covariates, $\boldsymbol{\beta}$ is a vector of coefficients,and $\lambda_0(t^\star)$ is an arbitrary and unspecified base-line hazard function. The dimension of the vectors $\mathbf{z}$ and $\boldsymbol{\beta}$ is M, which represents the number of (quantitative and qualitative) covariates considered in the analysis. The survivor function $S(t^\star; \mathbf{z})$ and pdf $f(t^\star; \mathbf{z})$ for such model are

$$S(t^\star; \mathbf{z}) = \exp\left[- \int_0^{t^\star} \lambda_0(u) \exp(\mathbf{z}\boldsymbol{\beta}) du \right] = S_0(t^\star)^{\exp(\mathbf{z}\boldsymbol{\beta})} \tag{2.2}$$

and

$$f(t^\star; \mathbf{z}) = \lambda(t^\star, \mathbf{z}) S(t^\star, \mathbf{z}) \tag{2.3}$$

Note that the covariates act multiplicatively on the hazard function and that $\mathbf{z}$ does not depend on time (equation 2.1). This means that the "form" of the base-line $\lambda_0(\cdot)$ is always the same apart from a multiplicative factor that depends on the covariates. From a physical point of view, this means that the mechanism of earthquake occurrence (described by the function $\lambda_0(\cdot)$) is the same for different areas; only the parameters of the system can vary (i.e., $\exp(\mathbf{z}\boldsymbol{\beta})$). The model is nonparametric because it does not assume any specific form for the base-line hazard function, therefore it is more flexible compared to the traditional parametric approaches. In practice, it means that we do not impose any *a priori* assumption on the mechanism of the earthquake occurrence process, but the most likely probabilistic model is suggested by the empirical data.

The main problem here is to estimate $\boldsymbol{\beta}$ and the nonparametric form of $\lambda_0(\cdot)$ in equation 2.1. These quantities have a great importance. The coefficients contained in vector $\boldsymbol{\beta}$ give the relative importance of each considered covariate. The form of $\lambda_0(\cdot)$ gives important insights on the physics of the process.

2.1 Estimation of $\boldsymbol{\beta}$

Let us suppose to have N inter-event times t_i $(i = 1, ..., N)$ and the corresponding covariates $\mathbf{z}_i$. For the sake of simplicity, at this first stage we do not consider the CTs and possible ties in t_i. Let us define two additional vectors, $\mathbf{o}(t_i)$ and $\mathbf{r}(t_i)$

$$\mathbf{o}(t_i) = [t_{(1)}, t_{(2)}, ..., t_{(N)}] \tag{2.4}$$

$$\mathbf{r}(t_i) = [(1), (2), ..., (N)] \tag{2.5}$$

where $t_{(i)}$ are the IETs ordered in increasing order.

We can image a number of transformations of the variable t through strictly increasing a differentiable transformation of $(0, \infty)$ onto $(0, \infty)$. Since $\lambda_0(\cdot)$ is completely unknown, the inference problem about $\boldsymbol{\beta}$ is only based on the vector $\mathbf{r}(\cdot)$, because it is the only one that is always invariant under the transformations done on t. In technical terms we can say that $\mathbf{r}(\cdot)$ is sufficient for the estimation of $\boldsymbol{\beta}$ in the absence of knowledge of $\lambda_0(\cdot)$. In such case, the likelihood is proportional to (see Kalbfleisch & Prentice, 1980)

$$\begin{aligned} P(\mathbf{r}; \boldsymbol{\beta}) &= P\{\mathbf{r} = [(1), ..., (N)]; \boldsymbol{\beta}\} \\ &= \int_0^\infty \int_{t_{(1)}}^\infty ... \int_{t_{(N-1)}}^\infty \prod_{i=1}^N f(t_{(i)}; \mathbf{z}_{(i)}) dt_{(N)} ... dt_{(1)} \\ &= \prod_{i=1}^N \frac{\exp(\mathbf{z}_{(i)}\boldsymbol{\beta})}{\sum_{j \in \Omega(t_{(i)})} \exp(\mathbf{z}_j \boldsymbol{\beta})} \end{aligned} \tag{2.6}$$

where $f(\cdot)$ is the pdf (see equation 2.3), and $\Omega(t_{(i)})$ is the set of labels attached to the IETs with length $\geq t_{(i)}$, that is $\Omega(t_{(i)}) = \{(i), (i+1), ..., (N)\}$. A demonstration of this relation can be found in Appendix C.

In most practical cases, such as in the hazard studies, it is very important to take also into account CTs which often represent a significant part of the available data. For instance, it is very common to have very few earthquakes inside each seismic area, therefore the only CT available

for that area is usually a relevant percentage of the data. To properly handle CTs we need to introduce some modifications to equation 2.6. The greatest problem in dealing with CTs is that their statistical distribution is almost certainly different from the distribution of the IETs (e.g. Kalbfleisch & Prentice, 1980) and it is very difficult to figure out. Here, as well as in most of the scientific literature on this issue, it is assumed a random mechanism for the censoring. In practice, this implies that the end of the seismic catalogue is independent of the time of occurrence of the last earthquakes in each zone. From a statistical point of view, this also means that the CTs are RVs stochastically independent of each other and of the IETs. Note that this hypothesis seems very reasonable for earthquake catalogues.

Let us consider N_1 IET, $t_{(1)}, ..., t_{(N_1)}$, with corresponding covariates $\mathbf{z}_{(1)}, ..., \mathbf{z}_{(N_1)}$, and N_2 CTs $\tilde{t}$, of which n_i are censored in the i-th interval $[t_{(i)}, t_{(i+1)})$, i.e., $t_{(i)} \leq \tilde{t}_{i1}, ..., \tilde{t}_{in_i}$ $(i = 0, ..., N_1)$, and with covariates $\mathbf{z}_{i1}, ..., \mathbf{z}_{in_i}$. The CTs contain only partial information on the rank vector $\mathbf{r}(\cdot)$, because the ordering vector including the CTs is not necessarily the "true" ordering that we would have if we waited for the occurrence of the earthquakes. In this case, the contribution of the CTs in the time interval $[t_{(i)}, t_{(i+1)})$ to the marginal likelihood is

$$g(t_{(i)}) = \exp\left[-\sum_{j=1}^{n_i} \exp(\mathbf{z}_{ij}\boldsymbol{\beta}) \int_0^{t_{(i)}} \lambda_0(u)du\right]. \tag{2.7}$$

Therefore equation 2.6 can be rewritten as

$$\begin{aligned} P(\mathbf{r};\boldsymbol{\beta}) &= \int_0^{\infty} \int_{t_{(1)}}^{\infty} ... \int_{t_{(N_1-1)}}^{\infty} \prod_{i=1}^{N_1} f(t_{(i)};\mathbf{z}_{(i)}) g(t_{(i)}) dt_{(N_1)} ... dt_{(1)} \\ &= \prod_{i=1}^{N_1} \frac{\exp(\mathbf{z}_{(i)}\boldsymbol{\beta})}{\sum_{j \in \Omega'(t_{(i)})} \exp(\mathbf{z}_j\boldsymbol{\beta})} \end{aligned} \tag{2.8}$$

where $\Omega'(t_{(i)})$ is the set of labels attached to the IETs and CTs with length $\geq t_{(i)}$, i.e., $\Omega'(t_{(i)}) = \{[(j), j1, ..., jn_j], j = i, ..., N_1\}$. (See for detail Appendix C).

In both cases, equations 2.6 and 2.8 are the product of single terms, each one representing the rate between the hazard function of the i-th IET, and the sum of the hazard functions relative to the IETs and CTs $\geq t_{(i)}$. In practice, we can consider each IET and CT as labelled objects with a weight $\exp(\mathbf{z}\boldsymbol{\beta})$. In this picture equation 2.8 calculates the probability of observing a particular sequence of labels relative only to the IETs. Note that equation 2.8 is not the probability to get the observed IETs and CTs in the censored experiment. This probability would depend on the (unknown) censoring mechanism and also on the form of $\lambda_0(\cdot)$. Equation 2.8 is the probability that, in the underlying uncensored version of the experiment, the event consisting of the IETs and CTs observed would occur.

Finally, let us consider the possibility to have some ties in the data. This can occur, for instance, when the measurement errors lead to some kind of binning of the IETs (e.g., for paleoseismic data). For this purpose we use the same strategy used to incorporate the CTs. Let N be the number of individuals under test. Suppose to have d_i IETs with length $t_{(i)}$ $(i = 1, ..., N_1)$, where $t_{(1)}, t_{(2)}, ..., t_{(N_1)}$, and $\sum d_i = N$ (no censoring). In this case we still have partial information on the rank vector, as for the CTs above. Although the ranks of IETs with length $t_{(i)}$ are known to be less than those with length $t_{(j)}$ $(i < j)$, the arrangement of the ranks of the d_i IETs is otherwise unknown. As before, it is hardly conceivable to take into account all the possible dispositions of the ranks. Peto (1972) and Breslow (1974) have shown that if the number of ties d_i is small compared to the number of objects contained in the set Ω (see equation 2.6), the likelihood can be written as

$$L = \prod_{i=1}^{N_1} \frac{\exp(\mathbf{s}_i\boldsymbol{\beta})}{\left[\sum_{j\in\Omega'(t_{(i)})} \exp(\mathbf{z}_j\boldsymbol{\beta})\right]^{d_i}} \tag{2.9}$$

where $\mathbf{s}_i$ is the sum of the covariates of the IETs with length $t_{(i)}$ (see Appendix C).

The Maximum Likelihood Estimation (MLE) $\hat{\boldsymbol{\beta}}$ can be obtained as a

solution of the system of equations

$$U_j(\hat{\boldsymbol{\beta}}) = \frac{\partial \ln(L)}{\partial \beta_j} = 0 \tag{2.10}$$

where U_j is the score vector (e.g. Kalbfleisch, 1985). Even though only partial checks have been performed on this field until now, it has been suggested that only mild conditions on the covariates and censoring are required to ensure the asymptotic normality of $\hat{\boldsymbol{\beta}}$. Some numerical investigations, for instance, have shown that the asymptotic normality is a good approximation for data sets with no ties and less than 10 data (Kalbfleisch & Prentice, 1980). In this case $\hat{\boldsymbol{\beta}}$ has a pdf $N(\boldsymbol{\beta}, I(\hat{\beta})^{-1})$, where I_{jk} is the information matrix (Kalbfleisch, 1985).

$$I_{jk} = -\frac{\partial^2 \ln(L)}{\partial \beta_j \partial \beta_k} \tag{2.11}$$

This allows a direct test to check the relative importance of the factors included in the covariate vector. As a matter of fact, the question of the adequacy with which the asymptotic form of the distribution approximates the actual sampling distribution must be kept in mind in any practical applications. In cases where equivocal results arise, it might be wise to perform some numerical check for the sampling distribution of $\hat{\boldsymbol{\beta}}$. In Appendix D an explanation of properties of the score function and the information matrix is provided.

2.2 Estimation of the base-line function

Now the problem is to estimate the shape of the base-line function $\lambda_0(\cdot)$. As a first step we focus our attention on equation 2.2 and estimate $S_0(\cdot)$. As before, let $t_{(1)}, t_{(2)}, ..., t_{(N_1)}$ be the IETs, let P_i be the set of labels associated with the IETs with length $t_{(i)}$, and Q_i the set of labels associated with the CTs censored in $[t_{(i)}, t_{(i+1)})$ $(i = 0, ..., N_1)$, where $t_{(0)} = 0$ and $t_{(N_1+1)} = \infty$. The CTs in the interval $[t_{(i)}, t_{(i+1)})$ are $\tilde{t}_j$ where j ranges over Q_i. The contribution to the likelihood of an IET with length $t_{(i)}$ and covariates $\mathbf{z}$ is, under "independent" censorship,

$$S_0(t_{(i)})^{\exp(\mathbf{z}\boldsymbol{\beta})} - S_0(t^+_{(i)})^{\exp(\mathbf{z}\boldsymbol{\beta})} \tag{2.12}$$

where $S_0(t^+_{(i)}) = \lim_{\delta \to 0^+} S_0(t_{(i)} + \delta)$. The contribution of a CT $\tilde{t}$ is

$$S_0(\tilde{t}^+)^{\exp(\mathbf{z}\boldsymbol{\beta})} \tag{2.13}$$

Therefore, the likelihood can be written as

$$L = \prod_{i=0}^{N_1} \left\{ \prod_{j \in P_i} \left[S_0(t_{(i)})^{\exp(\mathbf{z}_j \boldsymbol{\beta})} - S_0(t^+_{(i)})^{\exp(\mathbf{z}_j \boldsymbol{\beta})} \right] \prod_{k \in Q_i} \left[S_0(\tilde{t}k^+)^{\exp(\mathbf{z}_k \boldsymbol{\beta})} \right] \right\} \tag{2.14}$$

where P_0 is empty. It is clear that L is maximized by taking $S_0(\tilde{t}) = S_0(t^+_{(i)})$ for $t_{(i)} < \tilde{t} \leq t_{(i+1)}$ and allowing probability mass to fall only at the observed IETs $t_{(i)}$.

The form of the base-line function is completely empirical, therefore it has a discrete form. The survivor function at a generic time $t^\star$ since the last event for a discrete process with CTs is (Kaplan & Meier, 1958)

$$S(t^\star) = \prod_{j | t_j < t^\star} (1 - \lambda_j) \tag{2.15}$$

where $\lambda_j = P(T = t_j | T \geq t_j)$. Note that equation 2.15 is the discrete version of equation 2.2. These observations lead to the consideration of a discrete model with hazard contribution λ_j at $t_{(j)}$ $(j = 1, ..., N_1)$. Thus we take

$$S_0(t_{(i)}) = S_0(t^+_{(i-1)}) = \prod_{j=0}^{i-1} \alpha_j \tag{2.16}$$

where $\alpha_0 = 1$. Substituting equations 2.2 and 2.15 in equation 2.14, and rearranging the terms we obtain

$$L = \prod_{i=1}^{N_1} \left\{ \prod_{j \in P_i} \left[1 - \alpha_i^{\exp(\mathbf{z}_j \boldsymbol{\beta})} \right] \prod_{k \in \Omega'(t_{(i)}) - P_i} \left[\alpha_i^{\exp(\mathbf{z}_k \boldsymbol{\beta})} \right] \right\} \tag{2.17}$$

The estimation of the survivor function can be carried out by joint estimation of the α_i and $\boldsymbol{\beta}$ in equation 2.17. More simply, we can take $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ as estimated previously and then maximize the logarithm of equation 2.17 with respect to α_i. In this case we obtain

$$\sum_{j \in P_i} \frac{\exp(\mathbf{z}_j \hat{\boldsymbol{\beta}})}{1 - \hat{\alpha}_i^{\exp(\mathbf{z}_j \hat{\boldsymbol{\beta}})}} = \sum_{k \in \Omega'(t_{(i)})} \exp(\mathbf{z}_k \hat{\boldsymbol{\beta}}) \tag{2.18}$$

If we do not have ties, equation 2.18 can be directly solved for $\hat{\alpha}_i$, otherwise an iterative solution is required. A suitable initial value for such iteration is

$$\ln(\alpha_{i0}) = \frac{-d_i}{\sum_{k \in \Omega'(t_{(i)})} \exp(\mathbf{z}_k \hat{\boldsymbol{\beta}})} \tag{2.19}$$

obtained by approximating $\hat{\alpha}_i^{\exp(\mathbf{z}_j \hat{\boldsymbol{\beta}})} \simeq 1 + \exp(\mathbf{z}_j \hat{\boldsymbol{\beta}}) \ln(\hat{\alpha}_i)$. A demonstration from equation 2.18 to 2.19 can be found in Appendix E. The MLE of the base-line survivor function for a generic time $t^\star$ since the last earthquake is therefore

$$\hat{S}_0(t^\star) = \prod_{i | t_{(i)} < t^\star} \hat{\alpha}_i \tag{2.20}$$

which is a step function. The equation of the proportional model for a generic time $t^\star$ and covariates $\mathbf{z}^\star$ is

$$\hat{S}(t^\star, \mathbf{z}^\star) = \prod_{i | t_{(i)} < t^\star} \hat{\alpha}_i^{\exp(\mathbf{z}^\star \hat{\boldsymbol{\beta}})} \tag{2.21}$$

As mentioned before, the trend of the function $\lambda_0(\cdot)$ has an important theoretical and practical meaning. A simple way to show such a trend is through the comparison of the empirical survivor function $\hat{S}_0(\cdot)$ and the survivor function for the Poisson process. Let us consider the transformation

$$u(t) = \ln\{-\ln[\hat{S}_0(t)]\} \tag{2.22}$$

that makes the RV $u(t)$ asymptotically normally distributed (e.g. Kalbfleisch & Prentice, 1980). The transformation 2.22 applied to the survivor function of the Poisson process gives $u_P(t) = \ln\lambda + \ln(t)$. Therefore, a plot of the residuals

$$\varepsilon(t) = u(t) - u_P(t) \tag{2.23}$$

versus t shows the departures of the empirical survivor function from a theoretical Poisson distribution. By looking at equation 2.22 and to the relation between survivor and hazard function, it is easy to show that the trend of $\varepsilon(t)$ has a behaviour comparable to the trend of $\lambda_0(t)$.

2.3 Checking the model

A basic step of each kind of modeling is the statistical validation of the model. It is well known by now that the check of any model, regardless of its nature, should be performed on an independent data set, i.e., on data that have not been considered at any step of the modeling. For time sequences, a certain independent data set is given by future events. In any case, a preliminary check can be done by dividing the available data set in two, one used to set up the model (the *learning* phase), and the other to check the model (the *validation* phase). This approach provides a necessary condition for the validity of the model. In other words, if the model does not fit the available data we can disregard it. On the contrary, if the model fits well we can not rule out the possibility of an overfitting due to unconscious suitable choices. This chance may be drastically reduced by using a model with a number of parameters much lower than the independent observations used. In any case, as mentioned before, the real capability of the model to describe the process can be obtained only in real forward analysis.

As regards the model described in the previous section, the goodness-of-fit can be empirically checked by using independent data of the *validation* phase. Each IET (t) of this data set is transformed through the equation

$$\hat{e}_i = \hat{\Lambda}_0(t_i) \exp\{\mathbf{z}_i \hat{\boldsymbol{\beta}}\} \tag{2.24}$$

where

$$\hat{\Lambda}_0(t) = \sum_{i|t_{(i)}<t} \Big(\sum_{k \in \Omega'(t_{(i)})} \exp\{\mathbf{z}_k \hat{\boldsymbol{\beta}}\} \Big)^{-1} \tag{2.25}$$

(see equations 2.2, 2.19 and2.20) and $\hat{\boldsymbol{\beta}}$ are estimated by the data of the *learning* phase. In Appendix F a brief explanation of this technique is provided. If the model is appropriate, the residuals $\hat{e}_i$ should be similar to a sample drawn by an exponential distribution with $\lambda = 1$ (Kalbfleisch &Prentice, 1980; see also Ogata, 1988). Therefore, a comparison of the cumulative of the residuals $\hat{e}_i$ with a theoretical exponential curve provides a goodness-of-fit test of the model. The comparison can be statisti-

cally checked through a one-sample Kolmogorov-Smirnov test (Appendix G).

Chapter 3

Application to the Italian territory: regular grid

In this chapter the non-parametric multivariate model described previously is applied to the large Italian seismicity. PHM presents several advantages compared to other more traditional approaches. In particular, it allows testing straightforwardly a variety of hypothesis, such as any kind of time dependence (i.e., seismic gap, cluster, and Poisson hypothesis). Moreover, it may account for tectonics/physics parameters that can potentially influence the spatio-temporal variability, and to test their relative importance. In the application to the Italian seismicity of the last four centuries, interesting results are found. In particular, they show that large earthquakes in Italy tend to cluster; the instantaneous probability of occurrence in each area is higher immediately after an event and decreases until it reaches, in few years, a constant value representing the average rate of occurrence for that zone. The results also indicate that the clustering is independent of the magnitude of the earthquakes. Finally, a map of the probability of occurrence for the next large earthquakes in Italy is provided.

3.1 Results

For applying PHM to the Italian seismicity, we use the parametric seismic catalogue published by the Working Group on CPTI (1999) integrated with more recent data (since 1980) coming from the Working Group on

CSTI (2001) and from the Bulletin of the Istituto Nazionale di Geofisica e Vulcanologia (INGV). In order to have a complete set of events we consider the earthquakes with $M \geq 5.5$ since 1600 (see figure 3.1). The

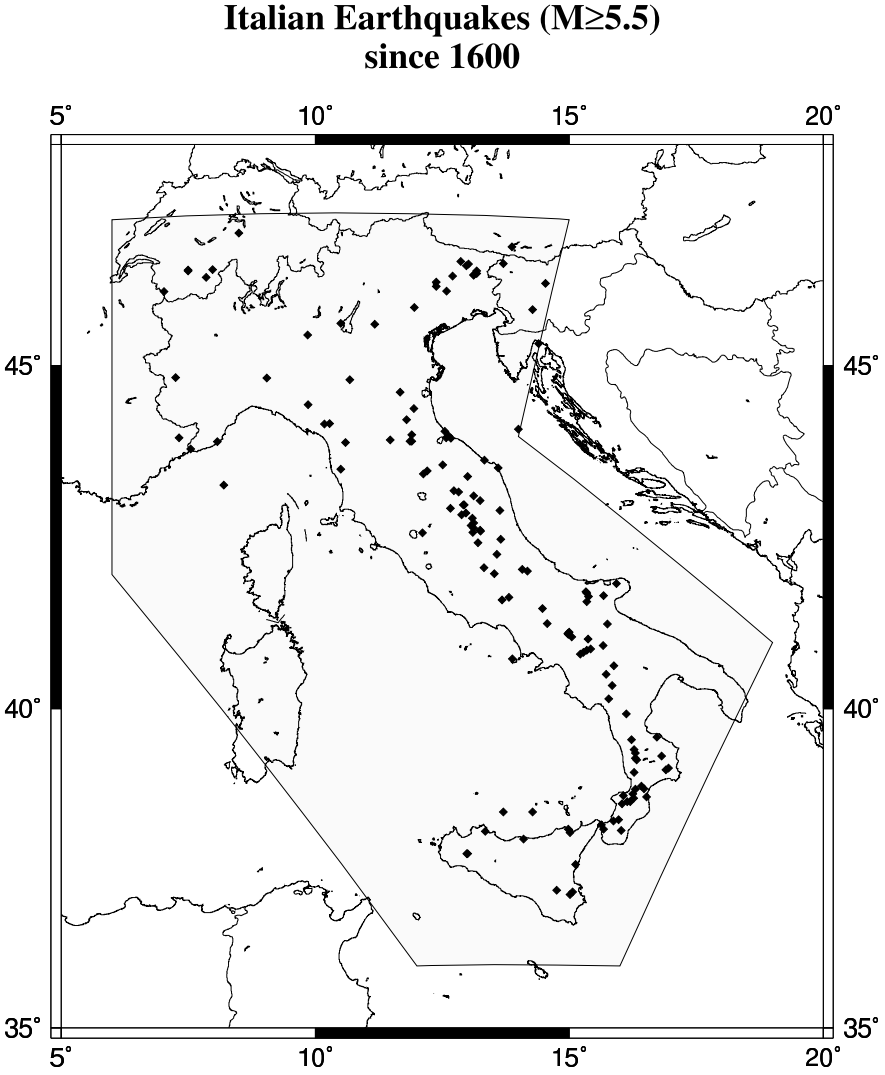

Figure 3.1: Map of the earthquakes with $M \geq 5.5$ since 1600 occurred in Italy.

threshold magnitude has been chosen taking into account practical aspects. In fact, $M = 5.5$ generally represents a threshold value for which the earthquakes in the Italian territory are considered "dangerous" due to the high number of historical buildings. Previous papers assume that similar events are almost independent one from each others (e.g. Boschi *et al.*, 1995). This assumption would seem corroborated by the analysis

of the autocorrelation and of the goodness-of-fit with a Poisson distribution of the IETs calculated for the whole Italian catalogue (see figure 3.2). In particular, the two null hypotheses, of no autocorrelation and no departures from a Poisson process, are not rejected at a 0.05 significance level.

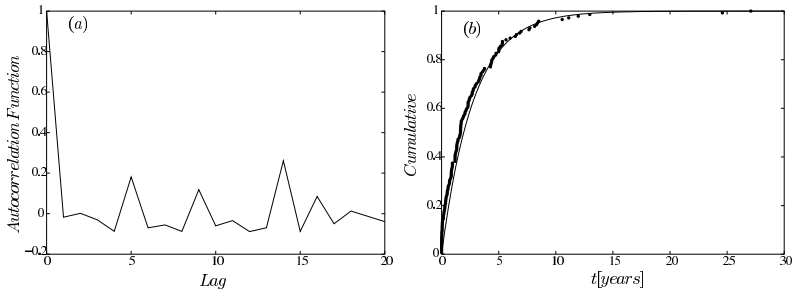

Figure 3.2: (a) Plot of the empirical autocorrelation function of the IETs calculated for the whole Italian catalogue. (b) Empirical cumulative distribution of the IETs calculated for the whole Italian catalogue.

The results obtained from the analysis of the whole catalogue may be strongly biased because the events can be hardly considered as an homogeneous sample. The Italian territory presents a strong and complex spatial variability due to the presence of different tectonics domains. In order to account for the inhomogeneities in the spatial distribution of the events, we consider a 13×13 grid between the latitudes 36-48 N, and longitudes 5-20 E (see figure 3.1). Each node of the grid is the center of a circle. In order to cover the whole area the radius R of the circle is set about $D/\sqrt{2}$, where D is the distance between the nodes. For such grid, $D \approx 100$ km and $R = 70$ km.

In a following step, we select the circles that contain at least 3 earthquakes; for each one of these circles we calculate the IETs among the earthquakes inside the circle, one CT relative to the time elapsed since the most recent event, and a vector of covariates $\mathbf{z}$ attached to each RV that may accommodate the characteristics of the earthquakes as well as the spatial/tectonic characteristics of the circle considered. Here, for each RV we consider a two-dimensional vector $\mathbf{z}$ composed by the rate of occurrence, i.e., the total number of clusters occurred inside the circle divided by 403 years (1600-2002), and by the magnitude of the event from which

we calculate the IET and the CT. The latter might be relevant in case the earthquakes occur following any general model where the probability to have another large earthquake in a given time interval depends on the magnitude of the last event. Some examples are the "time predictable" model (Shimazaki & Nakata, 1980; Papazachos & Papadimitriou, 1997), and the stochastic models (mainly based on the Gutenberg-Richter and a general Omori law) that describe the number of triggered events as a function of the magnitude of the triggering predecessor (e.g. Ogata, 1988, Kagan & Jackson, 2000).

By using the data of the Italian catalogue (see figure 3.1) and equations 2.10-2.11 of chapter 2, we obtain $\beta_1 = 55 \pm 8$ and $\beta_2 = 0.2 \pm 0.2$. The first coefficient, β_1, is the first component of the vector $\hat{\boldsymbol{\beta}}$, i.e., the coefficient that is linked to the rate of occurrence. The coefficient β_2, instead, is linked to the magnitude of the event from which the IET is calculated. As it appears clear, β_2 is not significantly different from zero, thus the magnitude of the events is not relevant for the distribution of the IETs. On the contrary, β_1 is positive and significantly different from zero.

In figure 3.3 we report the empirical survivor function $\hat{S}_0(\cdot)$ of equation 2.20 in chapter 2. Compared to the Poisson distribution, such empirical survival function has a greater number of small IETs. This departure is shown in figure 3.4, where the plot of the residuals $\varepsilon(t)$ (see equation 2.23 in chapter 2) is reported. As mentioned above, the trend of $\varepsilon(t)$ is comparable to the trend of $\lambda_0(t)$. The decreasing trend means that the earthquakes tend to be more clustered that in a simple Poisson process. The negative trend lasts for few years after the event, then the trend becomes almost flat as expected for a Poisson process. The presence of a significant departure from the Poisson model implies that the statistical distribution is time dependent, i.e., the probabilities of occurrence depend on the time elapsed since the most recent event. Note that there is no significant evidence of a positive trend.

In figure 3.5 we report the graphs representing the goodness-of-fit for the *learning* and *validation* data sets (see equation 2.24 in chapter 2).

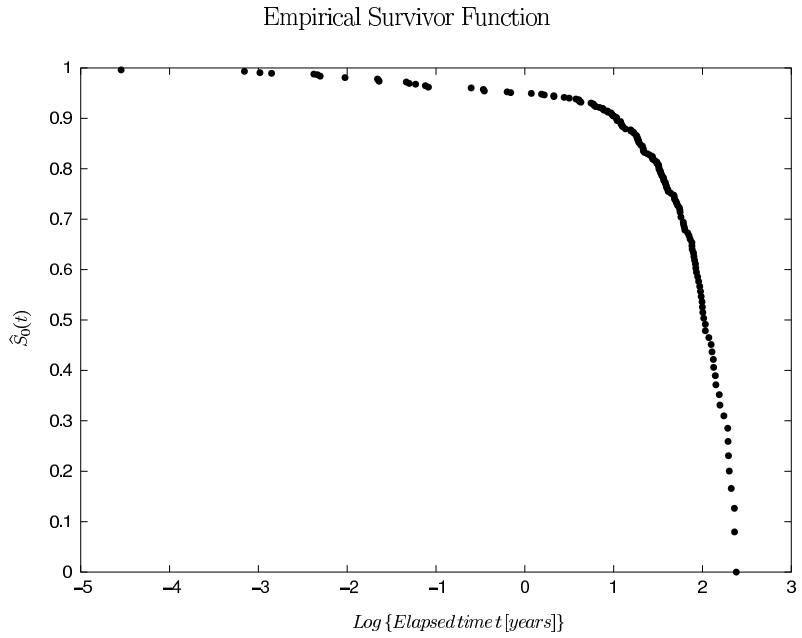

Figure 3.3: Empirical base-line survival function (equation 2.20 in chapter 2) for the Italian seismicity.

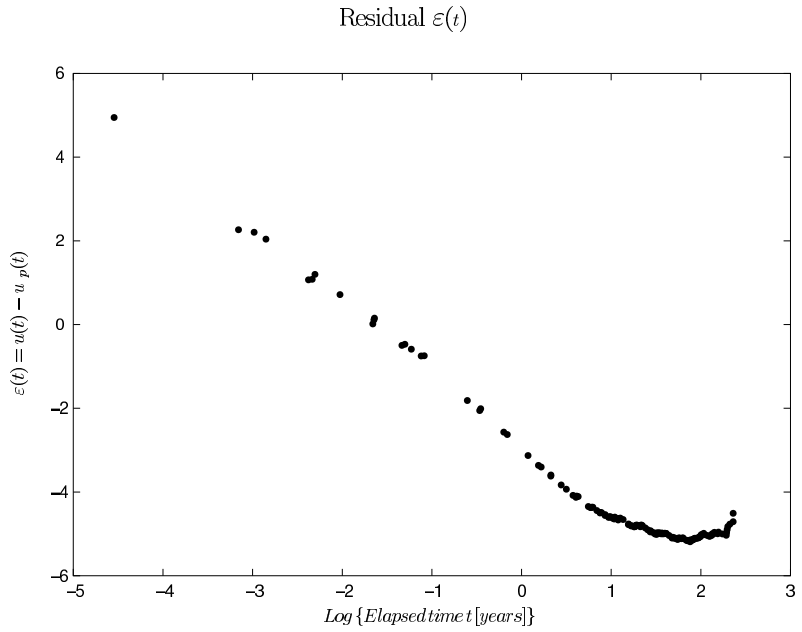

Figure 3.4: Plot of the residuals $\varepsilon(t)$ (equation 2.23 in chapter 2) as a function of the time elapsed since the last event.

In particular, we use the time interval 1600-1950 for the *learning* phase (124 earthquakes), i.e., to set up the model. The time interval 1951-2002, instead, is used for the *validation* phase (21 earthquakes), i.e., to verify the model with an independent data set. The figures clearly show that

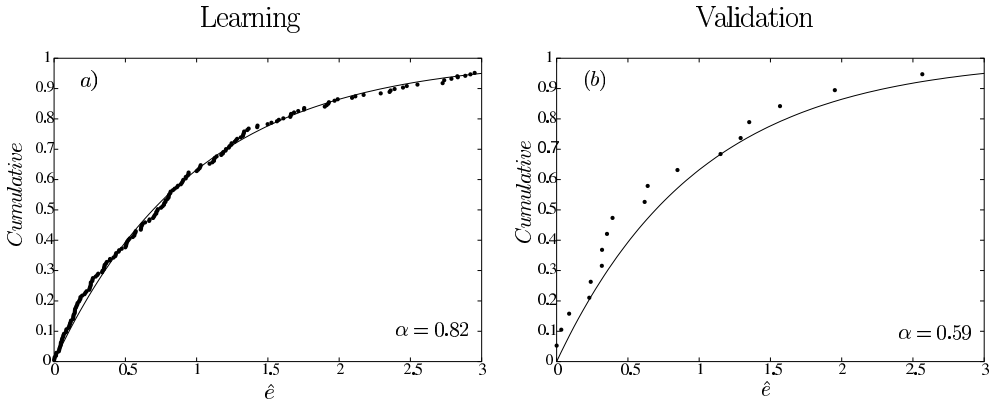

Figure 3.5: (a) Empirical (dotted line) and theoretical (solid line) cumulative functions for the *learning* data set. The parameter α is the significance level at which the null hypothesis (Poisson hypothesis) can be rejected (see text for more details). (b) The same as for (a), but relative to the *validation* data set.

the model fits the data very well, both in the *learning* and *validation* data sets. The goodness-of-fit is evaluated quantitatively through a one-sample Kolmogorov-Smirnov test (see Appendix G). The significance levels at which the null hypothesis of equal distributions is rejected are reported in the figures. In both cases, they are > 0.10.

In the next step, we calculate, from the empirical survival function of equation 2.21 in chapter 2, a map of the probability of occurrence for the future large earthquakes. The conditional probability to have an earthquake in the next τ years, given a time t since the most recent event, can be approximated by

$$P(t, \tau; \mathbf{z}) = \frac{S(t; \mathbf{z}) - S(t + \tau; \mathbf{z})}{S(t; \mathbf{z})} \tag{3.1}$$

where $\mathbf{z}$ is the vector of covariates relative to the IET $(t, t+\tau]$ of the circle considered. In figures 3.6 and 3.7, we report the map of the probability of occurrence for an earthquake of $M \geq 5.5$ in the next one ($\tau = 1$) and ten ($\tau = 10$) years in Italy. We identify as the most dangerous regions,

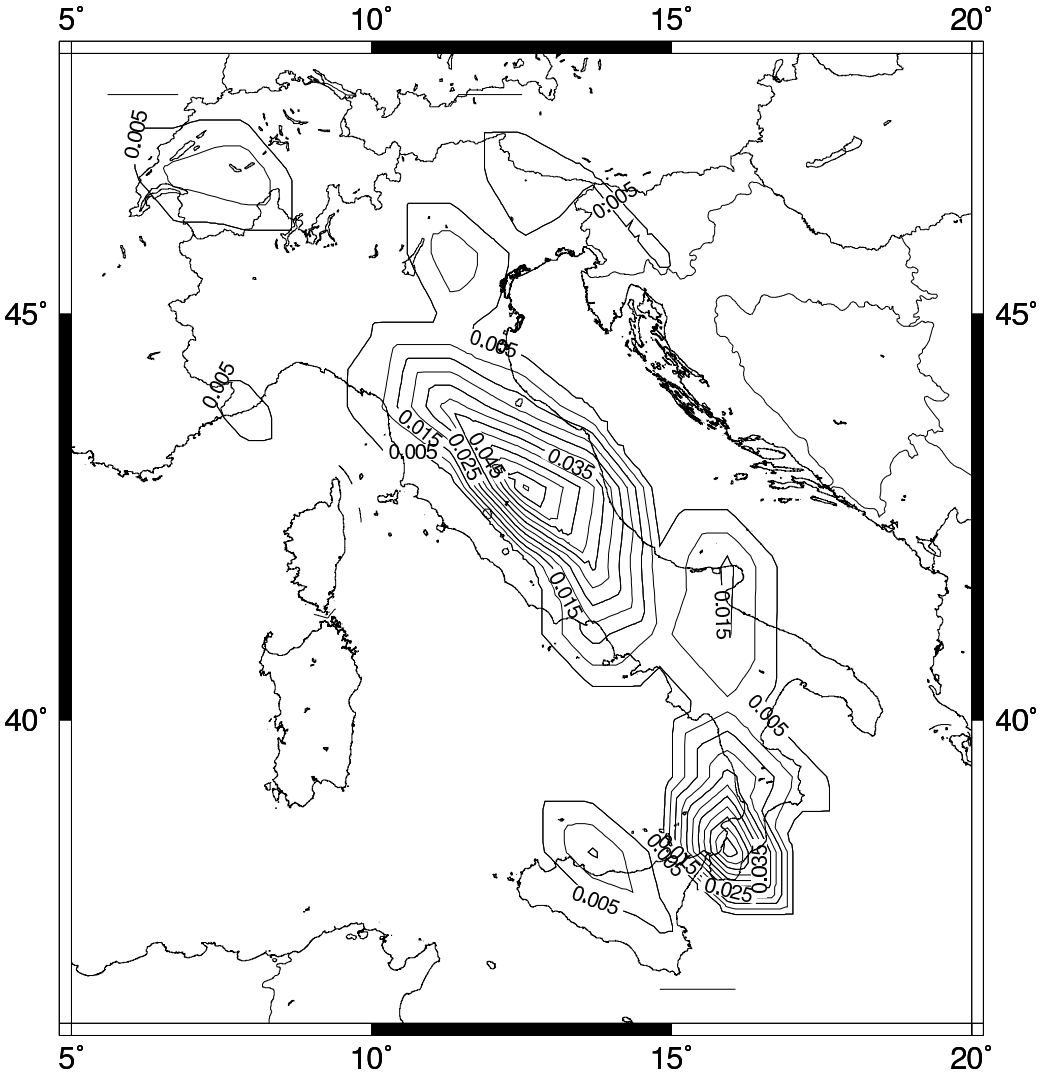

Figure 3.6: Map of the probability of occurrence for the next large earthquake in Italy calculated for $\tau = 1$ year (equation 3.1).

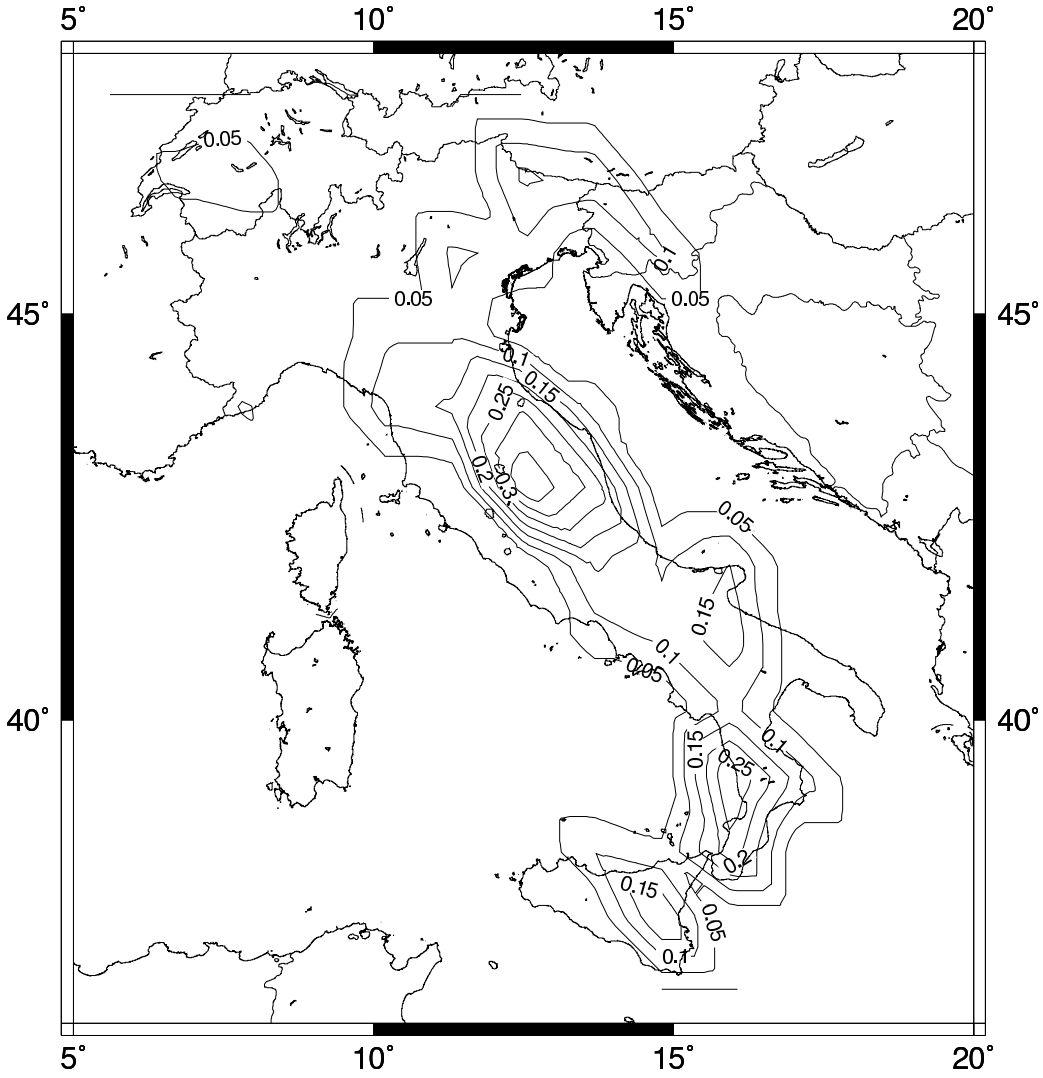

Figure 3.7: Map of the probability of occurrence for the next large earthquake in Italy calculated for $\tau = 10$ year (equation 3.1).

a large part of Northern-Central Apennines, and Central-Southern part of the Calabria; other smaller maxima of probability are located around Garda lake, in Friuli, and in part of Puglia-Basilicata and Sicilia regions.

As a final step, we evaluate the forecasting ability of the model through a retrospective forward modeling. For this purpose, we calculate the probabilities of the regions where earthquakes with $M \geq 5.5$ occurred since the 1950. The model can be considered "forward" because we calibrate it only by using data available immediately before each earthquake. In order to make easier to check the forecasting capability of the model, we bin the probabilities in four categories: "HIGH", "MEDIUM", "LOW", and "NEGLIGIBLE". The categories are assigned by ranking the probabilities of each circle. A rank equal to i means that there are $i - 1$ regions with higher probabilities. Then, we attach the category "HIGH" to the regions that have a probability rank below the median, the category "MEDIUM" for the regions above the median, the category "LOW" for the regions with less than 3 seismic events (for which we do not calculate the probability), and the category "NEGLIGIBLE" for the regions where no historical earthquakes occurred. In total, 46 circles experienced historical earthquakes, and 29 have 3 or more events. The results of the "forward" analysis are reported in Table 3.1.

In the period 1951-2002 (the *validation* data set), 16 out of 21 events occurred inside the 29 circles that experienced in the past 3 or more events: 13 inside circles with "HIGH" probability, and 3 inside circles with "MEDIUM" probability. Remarkably, 10 out of 21 occurred in the 5 circles with the highest probabilities. We have positively checked the stability of the results by using the seismic catalogue in a shorter period (since 1700), and two different grids, a 9×9 ($D \approx 140$ km and $R = 100$ km) and a 19×19($D \approx 70$ km and $R = 50$ km).

3.2 Discussion and concluding remarks

The main goal of this work was to apply the new nonparametric and multivariate method, proposed in chapter 2, in order to characterise the

spatio-temporal distribution of large earthquakes. The method has several advantages compared to more traditional approaches. It allows accounting for many different factors that can potentially influence the spatio-temporal distribution of earthquakes, and it provides a direct tool to test the relative importance of these factors. Moreover, the method is almost completely nonparametric. Compared to other stochastic models proposed in the past (i.e., Ogata, 1988; Kagan & Jackson, 2000) the method does not assume any *a priori* stochastic model for the time behaviour. Therefore, the empirical time behaviour found by the model (see, for instance, figure 3.3) furnishes a direct and formal check of the reliability of antithetic models of earthquake process generation, such as the ones based on seismic gap, Poisson, and cluster hypotheses.

The technique has been then applied to the Italian historical large earthquakes. The first result obtained is that large earthquakes ($M \geq 5.5$) in Italy tend to cluster in time and space. The time length of the cluster may reach few years; after this time the distribution of the earthquakes becomes a Poisson distribution. Note that the time clustering seems to be longer than what expected for a typical aftershock sequence (e.g., Gasperini & Mulargia, 1989; Console & Murru, 2001; see also Helm-

	5 circles with highest Pr.	Areas with HIGH Pr.	Areas with MEDIUM Pr.	Areas with LOW Pr..	Areas with NEGLIGIBLE Pr.
#events since 1951	10	13	3	3	2
Spatial coverage[1]	6%	17%	16%	19%	48%
Spatial coverage[2]	18%	33%	30%	37%	0%

[1] Calculated over the shaded area shown in figure 3.1
[2] Calculated over the total area of the 46 circles where large earthquakes occurred in the past 4 centuries

Table 3.1: Results of the retrospective forward simulation. The number of forecast earthquakes for each kind of binned probability circle (see text) are reported together with two different type of spatial coverage.

stetter & Sornette, 2003). Remarkably, we do not find any evidence of increasing hazard function. This result seems to rule out any statistical distribution characterised by an almost constant recurrence time, as the Weibull with $\beta > 1$ (Nishenko, 1985), the lognormal (e.g., Nishenko & Buland, 1987; Michael & Jones, 1998), the Brownian Passage Time (Ellsworth *et al.*, 1998), and all the other statistical distributions with an increasing hazard function. This is also in antithesis to the seismic gap model (McCann *et al.*, 1979) which assumes that recent earthquakes deter future ones. This result is instead in agreement to what found through different techniques by Ogata (1998), Kagan & Jackson (2000), and Rong & Jackson (2002) in other seismic regions.

With regards the spatial dimension of the cluster, this is in part imposed by the choice of R. This choice was mainly driven by two antithetical practical aspects; the need to have the largest number of circles having three or more events, but with the surface as small as possible. The last requirement is necessary to have almost tectonically homogeneous circles, and to make earthquake forecasting useful for hazard studies. In any case, in spite of we have observed comparable time clustering also by using different grids (see above), we argue that such time behaviour may depend on the spatial scale adopted, since it is hardly conceivable that the same time clustering holds at the two extremes of the spatial domain, i.e., for a single fault and for the whole earth. In particular, since the average of the IETs for each circle is definitely lower than the time expected to reload a fault through tectonics (in Italy, about 10^3 years; see, e.g., Pantosti *et al.*, 1993), we speculate that the long-term clustering found at the spatial scale defined by the dimension of the circles might be the result of the interaction among a probably high number of seismogenic faults, some of them close to the rupture. In this perspective, at such a spatio-temporal scale, the interaction between faults seems to be more relevant for earthquake forecasting purposes, than the behaviour of a single seismogenic fault.

Another interesting result obtained is that the length of the inter-event times is independent of the magnitude of the earthquakes. In other

words, the magnitude of the last event does not modify the probability to have another large event. From a physical point of view, this means that there is no evidence in favor of some kind of "time predictable" model (Shimazaki & Nakata, 1980; Papazachos & Papadimitriou, 1997). This result is also in contradiction to the models characterised by a time behaviour that depends somehow from the magnitude of the earthquakes (e.g., Ogata, 1988; Kagan & Jackson, 2000). Remarkably, this might lead to suggest that the Gutenberg-Richter law does not always hold, at least for large events occurring in small spatio-temporal windows.

The model has been also used to calculate the map of the probability of occurrence for the next large earthquakes in Italy. The same calculation has been made on a validation data set (1951-2002) through a retrospective forward analysis. This has allowed to evaluate the forecasting ability of the model, and to check its statistical reliability. The satisfactory forecasting ability provides a further evidence that the earthquakes tend to occur in the regions that experienced many large events in the past (e.g., Kagan & Jackson, 1991, 2000), also recently (e.g., Kagan & Jackson, 1999). As stated before, this is in contradiction with any sort of seismic gap hypothesis.

As a final consideration, we remark that the map of the probability of occurrence for the next large earthquakes reported here should be regarded only as a first attempt in this direction. Here, in fact, we have only considered information coming from a seismic catalogue disregarding all the geologic/tectonics variations that characterise the Italian territory. A reliable regionalization, the state of stress, the number and kind of tectonic structures, the modeling of possible interactions among different regions, are only some of the parameters to add to the vector $\mathbf{z}$ (see, i.e., equation 2.1 in chapter 2) that might improve significantly the maps reported in figures 3.6 and 3.7. Future works will be addressed on this direction.

Chapter 4

Application to the Italian territory: seismo-tectonical grid

In this second application of the PHM to the Italian seismicity, a seismological grid for the Italian territory, able to account for heterogeneities in the spatial domain, is defined. The main goal of this work is to provide a probability map for the next moderate to large earthquakes ($M \geq 5.5$) in Italy. The method has been applied to Italian seismicity of the last four centuries for earthquakes with $M \geq 5.5$. The Italian territory has been divided in 61 irregular zones representing areas with homogeneous tectonic regime resulting from active stress data. Besides the magnitude and the time of the earthquakes, the model includes information on the tectonic stress regime, the homogeneity of its orientation, the number of active faults, the dimension of the area, the homogeneity of the topography. The time distribution of $M \geq 5.5$ earthquakes appears clustered in time for few years after an event, then the distribution becomes similar to a memoryless Poisson process, leading to a time dependent probability map. This map shows that the most likely regions where the next large earthquakes may occur are Friuli, Umbria -Marche, and part of Southern Apennines and Calabrian arc.

4.1 Introduction

Italy is characterised by a generally high seismic risk. Although the largest earthquakes range only from medium to large magnitudes (up to about 7), the high density of inhabitants and the age and quality of the buildings make the vulnerability rather high. The memory of the casualties caused by an earthquake with $Mw = 5.8$ occurred in 2002 in Molise region (Southern Italy) is still recent.

The basic element for seismic hazard and, consequently, for seismic risk assessment, is sound modeling of the spatio-temporal distribution of earthquakes. Such a model allows for reliable long-term forecasting of future seismic activity, and extracting useful information on the physics of the earthquake occurrence process.

In most of the work devoted to this topic (see Working Group On California Earthquake Probabilities, 2003 as an exception), the spatio-temporal modeling of the earthquakes is usually accomplished by considering only the information reported in earthquake catalogues, such as time of occurrence, location (e.g., Posadas *et al.*, 2002; Stock and Smith, 2002), and sometimes magnitude and focal mechanism applied on regular grid (e.g., Kagan and Jackson, 2000). In practice, this means neglecting all the tectonic and geologic heterogeneities that characterise seismic regions with complex tectonic setting like Italy. The only way in which geological information has been taken into account in some past modeling of Italian seismicity (e.g., Boschi *et al.*, 1995; Pace *et al.*, 2002) is through the use of seismotectonic spatial zonation (e.g., Meletti *et al.*, 2000; Scandone and Stucchi, 2000), even though the rules adopted to build the zonation are not always clearly stated.

In this application of the PHM, a new regionalization of the Italian territory is developed. In particular, we provide a model of the spatio-temporal distribution of the destructive earthquakes in Italy ($M \geq 5.5$) based on a regionalization and by using the statistical strategy of analysis developed in this thesis. For "regionalization" we intend the definition of areas that are homogeneous with regard to seismic behaviour and to

kinematics and orientation of the stress field. Besides considering the usual data of the earthquake catalogues (CPTI Working Group, 1999, 2004; INGV-CNT seismic Bulletin at ftp://ftp.ingv.it/bollet/), the statistical technique is able to account for any tectonic/physics factors that can potentially influence the spatio-temporal distribution of the events, and it tests their relative importance. The final purpose is to provide a probability map of earthquakes with $M \geq 5.5$ in Italy for the next ten years. Besides being an important ingredient for hazard (and risk) assessment, it is worth noting that such a map can also be used as a "reference" model to check the reliability of any possible "predictive strategy" (see, for instance, Marzocchi *et al.*, 2003A).

4.2 Hazard Regionalization

We develop a regionalization of Italy in order to account for heterogeneities both in the tectonic domains and in the spatial distribution of earthquakes. The regionalization consists of drawing by hand areas homogeneous with regard to kinematics and orientation of the active stress field and to seismic behaviour (figures 4.1 and 4.2, and see also table 4.1). The subjective drawing of the boundaries of each zone is also guided by the knowledge of regional tectonic features (figure 4.3). Such a regionalization has two main rationales: first, the distribution of earthquakes from a homogeneous region can provide insights on the physics of the process that generates those seismic events. The second is practical: the use of well defined zones of limited extent allows a significant decrease of the spatial coverage of the forecasting (see section 4.4). Note that, the regionalization used implies that a future $M \geq 5.5$ earthquake within a specific zone is expected to have a defined mechanism of rupture.

The procedure used to define the zones is primarily based on the integration of active stress indicators (borehole breakouts, earthquake focal plane solutions, and seismogenic faults, see figure 4.1), with the distribution of seismicity in terms of location and magnitude (figure 4.2) and with the main tectonic structures of Italy (figure 4.3). The integration

of these data allows us to determine kinematics and variations of the present-day stress regime for a large part of Italy. Although the available data are not homogeneously distributed, most are concentrated in portions of the country critical for seismic hazard and thus useful and adequate for our purpose.

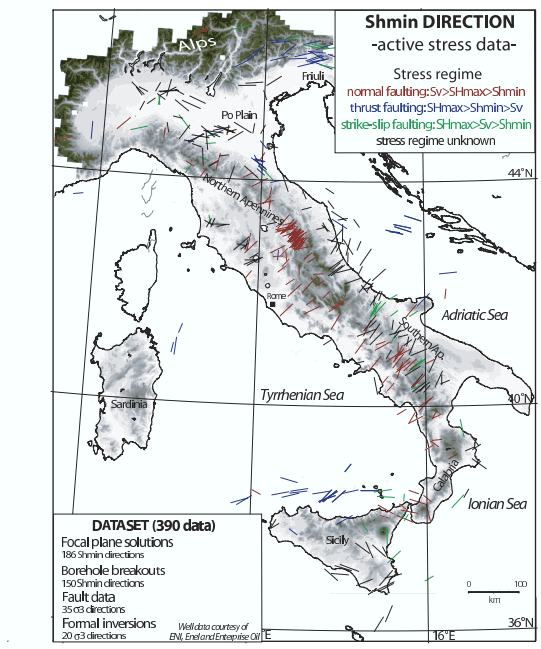

Figure 4.1: Data set used to build up the Italian hazard regionalization: S_{hmin} orientation and tectonic regime from active stress data (see references in the text).

For the sake of example, we report the basic criteria used to define the three zones, marked as U-M, L-A and CA in figure 4.4. Although these zones are all within the central-southern Apenninic region and are characterised by the same active tectonic regime and orientation, same topographic setting, and by good quality of data, we decide to keep them separate for the following reasons. At large scale, the geologic and tectonic setting are different (figure 4.3). The Apennines consist of an arcuate thrust belt with convexity towards the Adria-Africa foreland, where the thrusts show different size and curvature that progressively change their orientation (NS and NNW in U-M zone, NW-SE in L-A and CA zones). The degree of shortening varies irregularly according to the

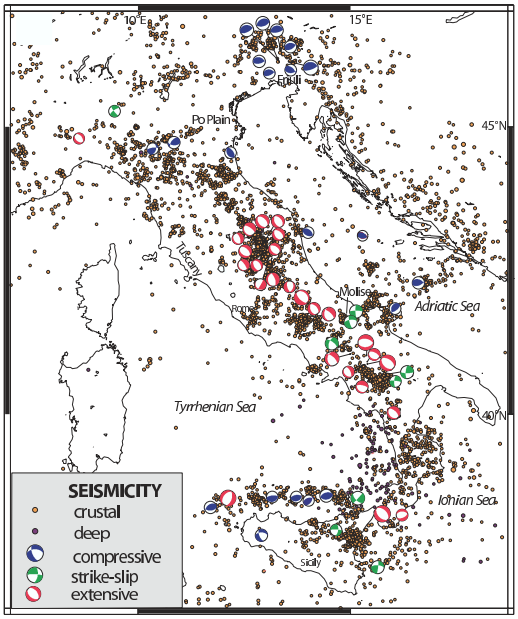

Figure 4.2: Data set used to build up the Italian hazard regionalization: Seismicity distribution of $M > 3$ earthquakes (period 1983-2003) and Centroid Moment Tensor fault plane solutions ($M \geq 5$), from INGV and CMT catalogs (scaled by magnitudes);

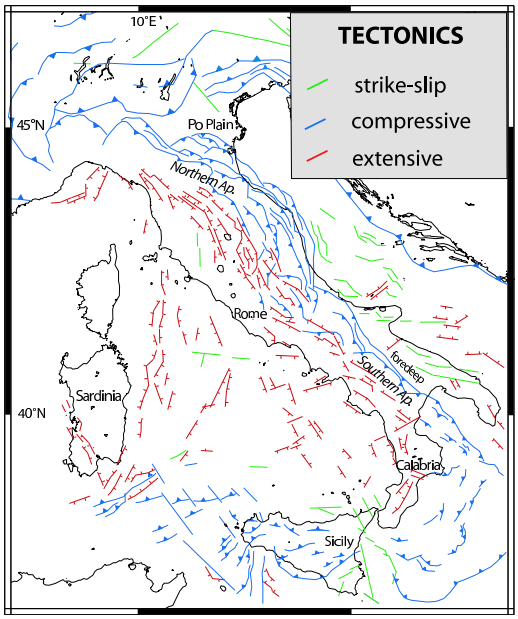

Figure 4.3: Data set used to build up the Italian hazard regionalization: Major structural features of Italy (simplified by Bigi *et al.*, 1991).

inherited paleogeography (U-M belongs to the pelagic basins, L-A and CA belong to the platform domain), contrasting rheology and differential sinking and roll-back of the subducting plate. We also subdivide the central-southern region because the seismicity is not widely diffused over the entire sector, but it appears more concentrated in U-M and CA zones. Relatively frequent seismicity with $M > 6$ magnitude events occurs in LA and CA zones (see table 4.1) while moderate magnitude ($M \leq 6$) and rare large magnitude earthquakes ($M > 6$) seem to characterise U-M zone (CPTI Working Group, 1999, 2004). As mentioned above, zones with a good data coverage, and well defined and limited extent allow a better localization of the forecasting.

The hazard regionalization (figure 4.4) broadly reflects the kinematic behaviour occurring in Italy (Anderson and Jackson, 1987; Patacca and Scandone, 1989; Westaway, 1992). Several areas are dominated by compression, such as the eastern Alps, the foredeep of the northern Apenninic arc, and a narrow band offshore of Sicily region. The active compression occurring in the southern Tyrrhenian sea shows a kinematics in agreement with the northward convergence of Africa relative to the Eurasian plate; north-south compression in the Alps is related to the relative motion of the Adriatic microplate, suggesting that this latter moves coherently with Africa. A long internal band running from the northern Apennines to the inner Calabria-Sicily arc shows a very well defined extensional stress field, with a rate of 2.5 - 5.0 mm/yr (Hunstad *et al.*, 2003). Although strongly debated, this process has often been attributed to the presence of subduction systems extending from Sicily to the northern Apennines. The hypothesis of an active retreating slab of the Adria microplate below the northern Apennines (Malinverno and Ryan, 1986), indicated by the presence of earthquakes down to 90 km depth (Selvaggi and Amato, 1992), is also referred as the driving mechanism for the ongoing convergence in the easternmost arc (Lucente *et al.*, 1999; Amato and Cimini, 2001; Piromallo and Morelli, 2003). The southern Apenninic foredeep and the Gargano area are characterised by a transcurrent regime clearly marked by important E-W strike slip faults

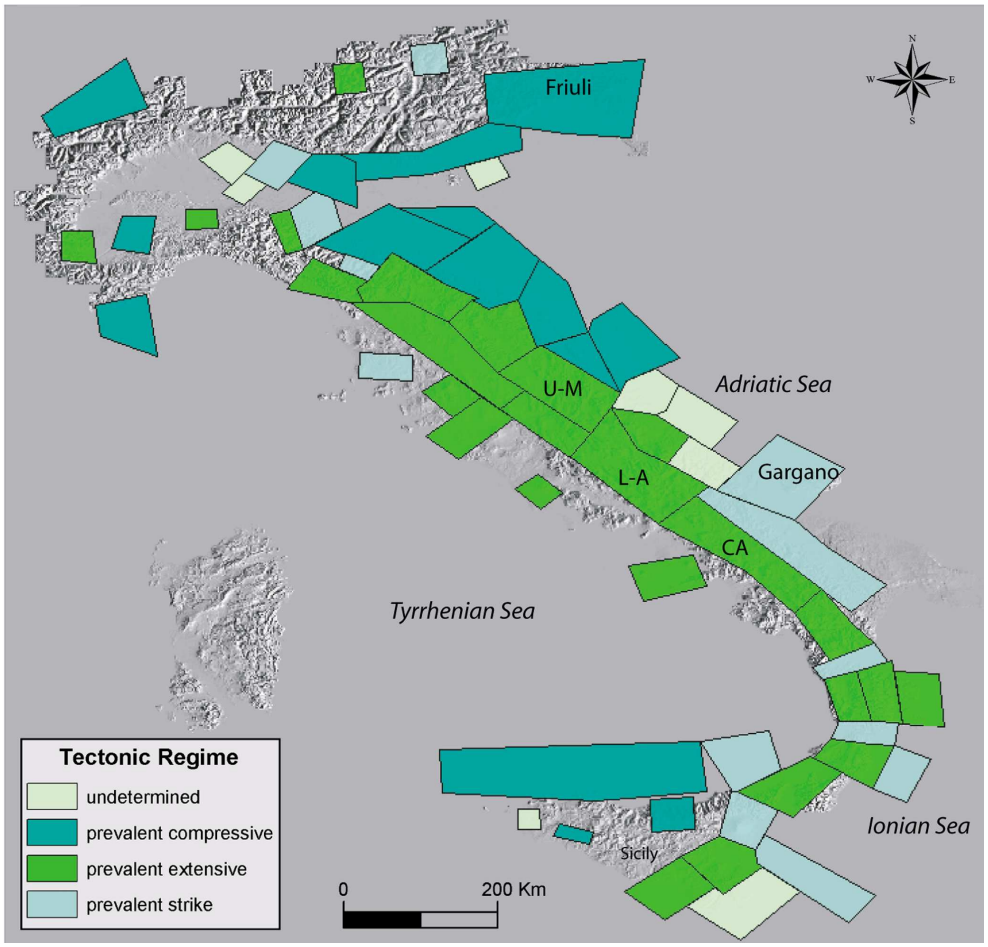

Figure 4.4: A total of 61 zones defines the resulting hazard regionalization. For each zone is shown the tectonic regimes. Abbreviations: U-M: Umbria-Marche; L-A: Lazio-Abruzzi; CA: Campania. The map is elaborated with GIS software and projected in GCS-WGS84.

Date and Magnitude of the $M \geq 5.5$ Earthquakes Since 1600 A.D.[a]

Zone 1		Zone 2		Zone 3		Zone 4		Zone 5	
Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M
1690/12/04	6.0	1688/06/05	6.7	1639/10/07	6.3	1659/11/05	6.5	1638/03/27	7.0
1700/07/28	5.7	1694/09/08	6.9	1639/10/15	6.5	1743/12/07	5.7	1767/07/14	5.8
1776/07/10	5.8	1702/03/14	6.3	1703/01/14	6.8	1783/02/07	6.6	1835/10/12	5.9
1788/10/20	5.6	1732/11/29	6.6	1703/01/16	5.6	1791/10/13	5.9	1854/01/12	6.2
1812/10/25	5.6	1805/07/26	6.6	1703/02/02	6.7	1905/09/08	7.1	1870/10/04	6.2
1873/06/29	6.3	1826/02/01	5.6	1730/05/12	5.8	1928/03/07	5.9	1913/06/28	5.5
1928/03/27	5.7	1853/04/09	5.9	1747/04/17	5.9	1947/05/11	5.6		
1936/10/18	5.9	1910/06/07	5.8	1751/07/27	6.3				
1976/05/06	6.4	1962/08/21	5.6	1832/01/13	5.7				
1976/09/11	5.5	1962/08/21	6.2	1838/02/14	5.5				
1976/09/15	6.0	1980/11/23	6.9	1859/08/22	5.6				
1976/09/15	5.9			1979/09/19	5.9				
1998/04/12	5.6			1997/09/26	5.9[b]				
				1997/09/26	5.9				
				1997/10/14	5.7[b]				

Zone 6		Zone 7		Zone 8		Zone 9		Zone 10	
Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M
1755/12/09	5.9	1654/07/23	6.2	1627/07/30	6.7	1731/03/20	6.3	1786/12/25	5.5
1855/07/25	5.8	1762/10/06	5.9	1627/07/30	5.8	1851/08/14	6.6	1875/03/17	5.7
1905/04/29	5.7	1904/02/24	5.5	1627/08/07	6.0	1930/07/23	6.7	1916/05/17	5.8
1946/01/25	6.1	1915/01/13	7.0	1627/09/06	5.8	1990/05/05	5.5	1916/08/16	5.9
1946/05/30	5.7	1984/05/07	5.7	1646/05/31	6.2				
				1875/12/06	6.1				
				2002/10/31	5.6				
				2002/11/01	5.8[b]				

Zone 11		Zone 12		Zone 13		Zone 14		Zone 15	
Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M
1726/09/01	5.5	1661/03/22	5.8	1690/12/23	5.6	1741/04/24	6.1	1854/12/29	5.7
1980/05/28	5.6	1688/04/11	5.9	1930/10/30	5.9	1799/07/28	5.9	1887/02/23	6.3
2002/09/06	5.6	1781/04/04	5.8	1943/10/03	5.8	1873/03/12	5.9	1963/07/19	5.9

Zone 16		Zone 17		Zone 18		Zone 19		Zone 20	
Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M
1786/03/10	6.0	1768/10/19	5.8	1708/01/26	5.5	1695/02/25	6.6	1781/06/03	6.2
1818/02/20	6.0	1918/11/10	5.7	1836/11/20	5.8	1891/06/07	5.6	1789/09/30	5.7
1978/04/15	6.1	1919/06/29	6.2	1857/12/16	7.0	1901/10/30	5.6	1917/04/26	5.7

Zone 21		Zone 22		Zone 23		Zone 24		Zone 25	
Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M
1783/02/05	6.9	1706/11/03	6.6	1837/04/11	5.5	1638/06/08	6.6	1693/01/09	5.9
1783/02/06	5.9	1933/09/27	6.6	1914/10/27	5.7	1832/03/08	6.5	1693/01/11	7.4
1894/11/16	6.1	1950/09/05	5.6	1920/09/07	6.5	1836/04/25	6.3	1818/03/01	5.5
1908/12/28	7.2								

(continued)

Zone 26		Zone 27		Zone 28		Zone 29		Zone 30	
Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M
1626/04/04	6.1	1796/10/22	5.5	1823/03/05	5.8	1968/01/15	5.6	1970/07/15	5.5
1783/03/01	5.9					1968/01/15	6.1		
1783/03/28	6.9								
Zone 31		Zone 32		Zone 33		Zone 34		Out of Zones	
Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M	Y/M/D	M
1802/05/12	5.5	1695/06/11	5.7	1883/07/28	5.7	1932/01/02	5.5	1644/02/15	5.9
								1808/04/02	5.6
								1828/10/09	5.6
								1834/02/14	5.5
								1846/08/14	5.6
								1907/10/23	5.9

Table 4.1: [a] Data are extracted from: CPTI Working Group (1999) for the period 1600-1992, INGV-CNT seismic Bulletin and CPTI Working Group (2004) for 1992-2004;[b] Data from NEIC (http://neic.usgs.gov/neis/epic/epic.html).

and by deep seismic hypocenters (depth $\geq$ 25 km).

4.2.1 Data Sets

For the specific purpose of this regionalization we use only data with known quality reported in the literature, or defined by our analysis, which are commonly accepted by different authors. For this reason, the data considered in the regionalization are only a part of the larger amount of published data on active stress regime.

The map in figure 4.1 represents the most up-to-date, complete, and reviewed catalogue of active stress indicators of Italy (see also figure 4.5). This is used to define the present-day stress field of the country. The map shows the directions of the minimum horizontal stress (S_{hmin}, corresponding to σ_2 or σ_3, with $\sigma_1 > \sigma_2 > \sigma_3$; compressive stress is assumed positive) inferred from borehole breakouts analysis, earthquake fault plane solutions and seismogenic fault data. It was compiled by updating the existing active stress map (Montone *et al.*, 2004), with new stress information derived from seismogenic faults.

A detailed description of the different stress measurement techniques

and methods used to obtain the stress directions in this work, can be found in Zoback (1992) and Montone *et al.*, (1999, 2004 and references therein) as concerns the Italian data set.

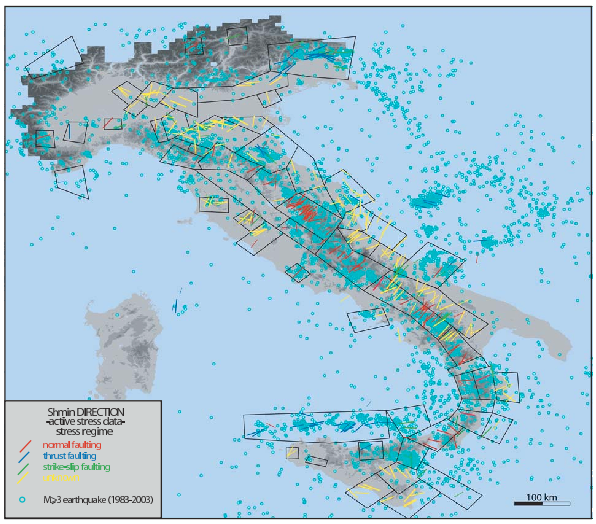

Figure 4.5: Plot of the active stress indicators to define the present-day stress field of Italy (earthquakes fault plane solution, borehole breakout data, seismogenic fault data) and of the $M > 3$ earthquakes that occurred berween 1983 and 2003. The integration of these punctual data and the knowledge of regional tectonic features (see figure 4.3) guided the drawing of the 61 zones of the hazard regionalization (black polygons).

A total of 390 data points are used to construct figure 4.1. These include the following:

- Borehole breakout data: Measurements are from 150 deep wells that are spread across Italy (Montone *et al.*, 2004). Major concentrations of these data are along the southern Apenninic belt, along the Apenninic foredeep, and also in the Po Plain. The maximum depth of the wells is 7.5 km with an average depth of the breakout zones around 3-4 km.

- Earthquake fault plane solutions: Although indicative of strain and not directly of stress, focal mechanisms are commonly used as a first approximation to indicate S_{hmin}. For this reason, we use only well constrained focal mechanisms, obtained for events of $M \geq 4.0$,

that most likely represent the response to the regional stress field rather than local structural complexities. The earthquake data set is mainly represented by 155 Centroid Moment Tensor solutions ($M \geq 4.0$, period updated to 2003) and also by 31 focal mechanisms obtained from P waves polarities ($4.0 \geq M \geq 7.0$, period 1908-1996) and by 20 stress directions determined from formal inversions of P, T and N axis of diffuse seismicity (Montone *et al.*, 2004).

- Seismogenic fault data: To determine stress regime and direction, we also consider the wide data set available on seismogenic faults (e.g., Galadini *et al.*, 2001, and references therein; Valensise and Pantosti, 2001, and references therein; Galli and Bosi, 2002, 2003), that were marginally used in the map compiled by Montone *et al.* (2004). We remark that these faults are not necessarily the ones that produced the earthquakes used in the statistical catalogue. In general, these literature sources synthesize information on fault orientation and sense of slip. Our selection of the fault data is driven by two choices: the seismogenic faults have to be clearly defined by geologic and geophysical observations, and the fault parameters have to have a general consensus. This means that we consider faults whose geometries are similarly described in more than one literature source. The inclusion of these fault data allows improving the stress-field estimation in terms of quality and distribution, also because they refer to the whole crustal volume, differently from the punctual breakout data. A total of 35 faults is selected and inserted in the stress map.

The boundaries of the zones are also inferred using other constraints (figure 4.4):

- the spatial of instrumental seismicity ($M > 3$, period 1983-2003; INGV seismic Bulletin at ftp://ftp.ingv.it/bollet/) and the characteristic magnitude of the events;

- the knowledge of the regional tectonic setting (figure 4.3) (Bigi *et al.*, 1991).

In figure 4.5 part of the data set (punctual data) is plotted along with the boundaries of the zones.

4.3 Estimate Covariate for Each RV

PHM is applied to the Italian catalogue of the $M \geq 5.5$ earthquakes in the last four centuries (1600-2003) (CPTI Working Group, 1999, 2004; INGV-CNT seismic Bulletin ftp://ftp.ingv.it/bollet/). The catalogue is considered complete for this time-magnitude window (see chapter 3). Once the regionalization is defined we estimate the covariates for each RV and they are inserted, as numerical code, in the model of spatial-temporal earthquake distribution (see chapter 2). In particular, at each IET and CT, we attach a vector $\mathbf{z}$ with 7 components, the logarithm of the rate of occurrence, the magnitude of the earthquake from which we calculate the IET and CT, the predominant tectonic regime, the number of seismogenic faults, the homogeneity of the stress orientation, the homogeneity of the topography, and the extent of the area.

A detail description of the covariates follows

- Logarithm of rate of earthquakes occurrence: It is the natural logarithm of the ratio of number of sequences of $M \geq 5.5$ earthquakes occurred inside the areas (not on individual faults) divided by 404 years, period covered by the seismic catalogue (1600-2003). Each sequence is composed by earthquakes with $M \geq 5.5$ occurred in a time window of three months and at a spatial distance less than 30 km (see CPTI Working Group, 1999)[1].

[1]The use of the natural logarithm of the rate of occurrence has a technical meaning. For PHM the survivor function is $S(t;\mathbf{z}) = S_0(t)^{\exp(\mathbf{z}\boldsymbol{\beta})}$; its shape, for large t, is similar to an exponential distribution where $S_{exp}(t) = \exp(-\lambda t)$ with λ the rate of occurrence. If $S(t;\mathbf{z}) = \exp(-\lambda^\star t)$, we want to see the parameters for which this relation holds:

$$S_0(t)^{\exp(\mathbf{z}\boldsymbol{\beta})} = \exp(-\lambda^\star t) \tag{4.1}$$

for one dimensional vectors $\mathbf{z}$ and $\boldsymbol{\beta}$, representing the covariate relative to the rate of occurrence and its weight, respectively. This relation is valid only if we consider $z_1 = \ln(\lambda)$ with $\beta_1 \simeq 1$ and $\lambda^\star \simeq 1$. Vector $\boldsymbol{\beta}_1$ is estimated with MLE, see below, while a linear regression of the $S_0(t)$ gives $\lambda^\star = 1.02 \pm 0.06$.

- Earthquake magnitude: The magnitude of the event from which the RV is calculated.

- Prevalent stress regime within each zone: It is attributed on the basis of the dominant tectonic style (1: extensional, -1: compressional, 0: strike-slip).

- Homogeneity of the stress orientation: We consider the standard deviation(σ) of the S_{hmin} data inside the zone where the RV is sampled (1: very low, $\sigma \geq 31°$; 2: low, $21° \leq \sigma \leq 30°$; 3: medium, $16° \leq \sigma \leq 20°$; 4: high, $\sigma \leq 15°$).

- Number of seismogenic faults: We count the number of faults that has been active in the past and/or might be the potential sources for a $M \geq 5.5$ earthquake within each zone. We define 3 classes (0: no fault; 1: less than or equal to 3; 2: more than 3).

- Homogeneity of the topography: We consider the topographic relief within the zone and then attribute a homogeneity degree (1: most of the territory is characterised by abrupt and high elevation changes; 2: the opposite).

- Area of the zone: We estimate the area in km^2 of each zone.

Besides the earthquake magnitude, all the other covariates are relative to the areas where the IET and CT are sampled.

In table 4.2 we list the covariates for the areas that experienced at least one large earthquake since 1600 A.D. For each zone we also list a quality code that is attributed on the basis of the reliability of the data points and by considering their spatial coverage.

4.4 Results and Discussion

The application of the PHM to the $M \geq 5.5$ Italian seismicity and the usage of equations 2.10-2.11 of chapter 2 show that only the first component of the weight vector $\boldsymbol{\beta}$ (β_1, associated to the logarithm of the rate

Covariate and Quality Code Relative to Cosidered Zones[a]

Zone	Log Rate of Occurrence	Prevalent Stress Regime	Homogeneity of the Stress Orientation	Fault Code	Topographic Homodeneity	Area km^2	Zone Quality
1	1.002	-1	3	2	1	13238	A
2	0.906	1	3	2	1	7760	A
3	0.906	1	4	2	1	6757	A
4	0.550	1	1	1	1	3020	C
5	0.396	1	1	1	1	2124	B
6	0.213	-1	1	1	2	5342	C
7	0.213	1	4	2	1	8679	A
8	-9.950D-03	0	4	2	1	7654	A
9	-9.950D-03	0	3	1	1	8811	B
10	-9.950D-03	-1	1	1	1	4837	C
11	-0.298	-1	3	2	1	18329	A
12	-0.298	-1	2	2	1	7345	B
13	-0.298	-1	3	1	1	5565	B
14	-0.298	-1	1	1	1	2241	B
15	-0.298	-1	1	1	1	3269	C
16	-0.298	0	1	1	1	2857	C
17	-0.298	1	2	1	1	5091	C
18	-0.298	1	4	2	2	3540	B
19	-0.298	-1	4	1	1	6619	B
20	-0.298	1	2	1	1	5189	B
21	-0.298	1	4	1	1	4018	B
22	-0.298	1	4	1	2	2893	B
23	-0.298	1	1	1	1	2667	B
24	-0.298	1	1	1	2	2634	B
25	-0.703	1	4	1	1	3500	C
26	-1.396	0	1	1	2	1864	C
27	-1.396	-1	2	1	1	7464	B
28	-1.396	-1	1	1	1	1906	C
29	-1.396	-1	1	1	2	621	B
30	-1.396	0	1	1	1	2525	C
31	-1.396	0	1	1	2	2306	C
32	-1.396	1	4	1	1	3434	B
33	-1.396	1	1	1	1	3114	C
34	-1.396	1	1	1	1	3102	C

Table 4.2: [a] Zone: 34 out of 61 zones numbered from higher to lower probability value (see figure 4.8); Log Rate of Occurrence; Prevalent Stress Regime: 1, prevalent normal; 0, prevalent strike-slip; -1, prevalent thrust; Homogeneity of the Stress Orientation: 4, high ($\leq 15°$); 3, medium ($16° - 20°$); 2, low ($21° - 30°$); 1, very low ($\geq 31°$ or insufficient data for calculation); Fault code: 1, number of faults ≤ 3; 2, number of faults > 3; Topographic Homogeneity: 1, Not Homogeneous; 2, Homogeneous; Zone Quality: A, very good; B, good; C, low;

of occurrence z_1) is significantly different from zero, being $\beta_1 = 1.2 \pm 0.2$. In practice, this means that the rate of occurrence seems to be the only important covariate (of the considered ones) in modeling the spatio-temporal distribution of large earthquakes.

As a further check on the importance of each covariate, we also run the program by using individual covariate one at a time. That is at each RV we attach a covariate vector having only one component. Compared to our previous case, besides the weight β_1 associated to the logarithm of the rate of occurrence, also the weight associated to the number of active faults appears statistically different from zero. The fact that the number of active faults is important only when considered alone indicates that the rate of occurrence may be linked not only to the tectonic rate, but also to the average number of active faults in the area. The analysis of all the covariates gives importance only to the rate of occurrence maybe because the latter is better quantitatively defined, being less grouped (see, for instance, table 4.2).

Figure 4.6 shows the form of the baseline $\lambda_0(t)$ for the last four centuries (see equation 2.23 in chapter 2). This function mimics the time behaviour of the instantaneous conditional probability of earthquake occurrence as a function of the time elapsed since the most recent event (the CT). It is important to stress that this function is not imposed a priori by our model, but it comes directly from the data. In the time window considered, the trend of the hazard function is decreasing in the first few years after an earthquake, then it gradually becomes almost flat as expected for a Poisson process. In practice, this means that seismic clustering is a prominent aspect of the time distribution of $M \geq 5.5$ earthquakes, at least over the first few years following a moderate to large earthquake; afterwards the process becomes almost time independent (e.g., Kagan and Jackson, 2000; and the result in chapter 3).

Remarkably, the length of such a cluster is significantly longer than the typical duration of aftershock sequences in Italy (until few months; see Console *et al.* 2003 and chapter 5). The latter are usually explained by the Omori law generated by an elastic interaction between faults (Di-

etrich, 1984). We speculate that the longer time clustering found here may be caused by a viscoelastic coupling between faults (Piersanti *et al.*, 1997; Pallitz *et al.*, 1998; Freed and Lin, 2001; Marzocchi *et al.*, 2003B). Note that, we do not find any significant evidence of seismic gaps over a time scale of few centuries.

Before estimating the probability of earthquakes we evaluate the validity of the model (see figure 4.7). The data belonging to the *learning* (1600-1950) and *validation* (1951-2003) data set are normalized through equations 2.24 and 2.25 in chapter 2, and compared to an exponential distribution with average equal to 1 (the null hypothesis). In both cases, the goodness-of-fit test given by a one-sample Kolmogorov-Smirnov test (see Appendix G) does not reject the null hypothesis of a common distribution. In practice, the model works well in both data sets.

The knowledge of the hazard function implies the knowledge of the survivor function and, therefore, the estimation of the probability for each zone defined in the regionalization through the same equation as

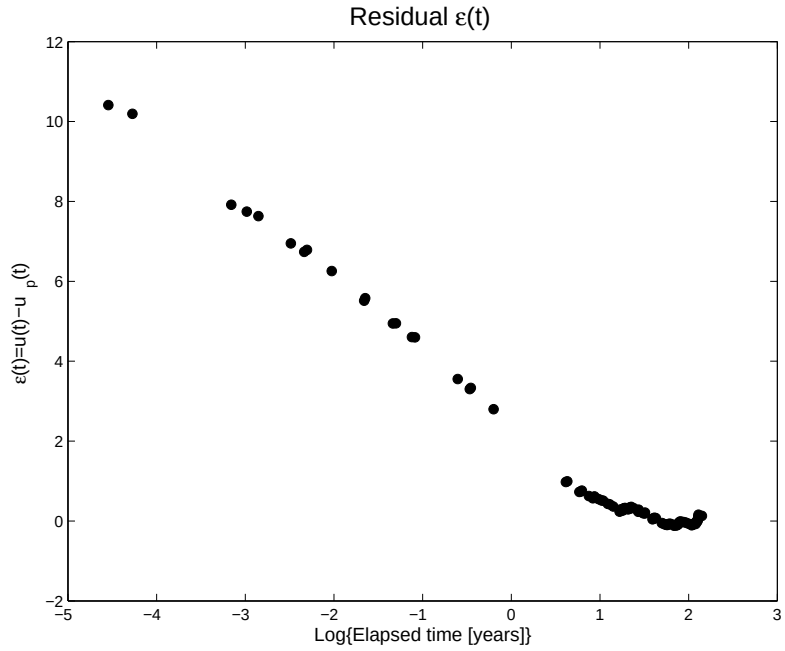

Figure 4.6: Plot of the residuals of the model as a function of the time elapsed since the most recent event (CT). The trend of the residuals is indicative of the trend of the hazard function.

the one of 3.1 in chapter 3. This result allows constructing a probability map of the next $M \geq 5.5$ earthquake in ten years of Italy. Ten years is a reasonable time window for forecasting, if we consider the length of the seismic catalogue used (four centuries).

In figure 4.8 we report 34 out of 61 zones (defined in figure 4.4) that have experienced at least one $M \geq 5.5$ event since 1600 A.D., and the location of the earthquakes. Each zone is color-coded for 10 year probabilities, which range from 2% to 27%. For the other 27 zones (and the rest of territory) we arbitrarily associate a probability $< 1\%$, because we cannot rule out that next future earthquakes will occur in these areas (we provide a check of this hypothesis in the following).

The four regions with the highest probabilities (in the range 15% - 30%) are the Friuli region (zone 1 in table 4.2), part of the Southern Apennines (zone 2 in table 4.2), Umbria-Marche region (zone 3 in table 4.2), and part of the Calabrian arc (zone 4 in table 4.2). The combined area of these four zones is 13% of the total area covered by the hazard regionalization. On the basis of the active stress data analysis, zone 1 has

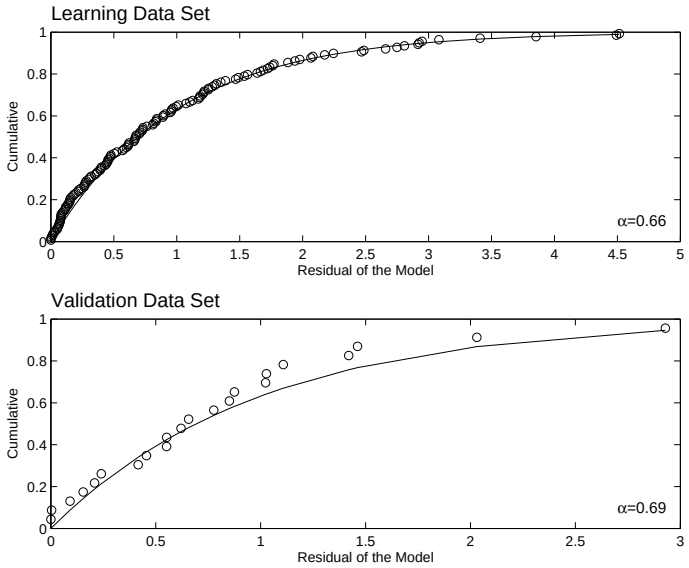

Figure 4.7: (a) Empirical (dotted line) and theoretical (solid line) cumulative functions for the *learning* data set. The parameter α is the significance level at which the null hypothesis can be rejected. (b) The same as for (a), but relative to the *validation* data set.

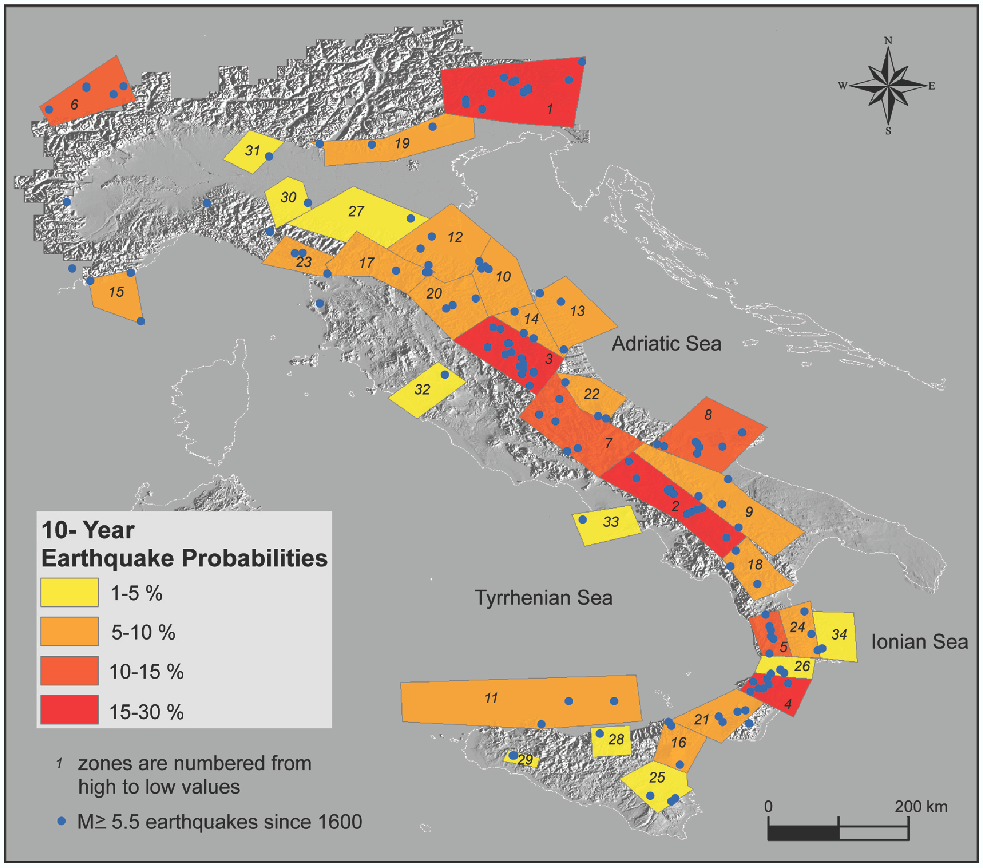

Figure 4.8: Probability map of moderate to large earthquakes ($M \geq 5.5$) in Italy for the next 10 years.

a 27% probability of occurrence of an earthquake produced by an E-W oriented fault with a likely thrust/reverse mechanism of rupture. Zones 2 and 3 both have a 25% probability of occurrence of earthquakes produced by a NW-SE oriented faults with an extensional rupture mechanism. Zone 4 has a 17% probability of occurrence of an earthquake produced by NE-SW oriented fault with an extensional mechanism of rupture.

The use of hazard regionalization allows a better spatial definition of probability, as it can be seen by comparing figure 4.8 with figure 3.7 in chapter 3. The comparison between the figures shows that the peaks in the two maps are roughly in the same position. Note that in this work the whole spatial coverage is significantly reduced and better defined as illustrated by comparing the probabilities in Central Italy and in Calabria region. Moreover, the use of the hazard regionalization allows obtaining a more stable probability map, than that one obtained by using a regular grid, for reasonable variations of the shape and/or dimension of the zones.

Compared to other Italian maps where a Poisson process is assumed (e.g., Albarello *et al.*, 1999; Akinci *et al.*, 2004), the main difference is that our probability map is time dependent, at least for few years after an earthquake. In other words, the probability map depends on the time in which it is calculated. Note also that the time dependent law found here is opposite to the one assumed in other papers (e.g., Pace *et al.*, 2002), where the conditional probability of earthquake occurrence (i.e., the hazard function) tends to increase with the time since the most recent event (e.g., Shimazaki and Nakata, 1980; Nishenko, 1985; Ellsworth *et al.*, 1998). This kind of opposite behaviour may be appropriate for single seismogenetic faults. Here, instead, we consider areas composed by multiple sources and the time behaviour observed is mainly linked to the (viscoelastic) interaction between these faults.

As final consideration, we estimate the forecasting ability of the model by looking at the probability rank of the zones where the earthquakes of the *validation* data set occurred. This is a retrospective forward analysis, because the data used for this calculation are never used to calibrate the model. In table 4.3, we report the number of times in which a $M \geq$

5.5 earthquake occurred in zones with high, medium, low and negligible probability. The grouping of the probabilities is described in the caption of table 4.3. In detail, 13 out of 23 events occurred in the 3 zones with the highest probabilities at that time, while only 2 out of 23 occurred in areas with a negligible probability.

Probabilities	Events Since 1951
Three zones with highest probability	13
Zones with high probability[a]	17
Zones with medium probability[b]	3
Zones with low probability[c]	1
Zones with negligible probability[d]	2

Table 4.3: Results of the Retrospective Forward Forecasting:
[a] High probability: $P > P_{66perc}$.
[b] Medium probability: $P_{33perc} < P \leq P_{66perc}$.
[c] Low probability: $P \leq P_{33perc}$.
[d] Negligible probability: no $M \geq 5.5$ earthquakes occurred since 1600.

We can use these results to check the main assumption of our model, that is, the earthquakes tend to occur in the same areas where they occurred in the past (in fact, we calculate the probabilities only for zones where at least one $M \geq 5.5$ earthquake occurred in the last 4 centuries). In particular, 2 out of the last 23 events occurred in areas not hit by previous moderate to large events. This means that we have approximatively 91% of probability that the next large event will occur inside one of the 34 zones where the probabilities are calculated.

4.5 Conclusions

The main goal of the present work has been to provide a probability map of the next $M \geq 5.5$ earthquakes in Italy, which presents a fundamental ingredient for seismic hazard assessment. We have used a nonparametric multivariate statistical model that is able to account for seismological/geological parameters simultaneously. The method has been applied to Italy, by using a regionalization based on tectonic parameters, the seismic catalogue of the last four centuries, and several seismological and

geological data. The results have indicated a time clustering of the earthquakes for few years, leading to a time-dependent probability map. We have found that the four zones with highest probability of occurrence in the next 10 years are Friuli region ($\sim$ 27%), part of the Southern Apennines ($\sim$ 25%), Umbria-Marche region ($\sim$ 25%), and part of the Calabrian arc ($\sim$ 17%).

Chapter 5

Some insights on the time clustering of large earthquakes in Italy

The aim of this work is to investigate on the clustering properties of the large earthquakes which occurred in Italy in the last four centuries. In particular, we compare the results of PHM applied to the catalogue of large earthquakes in Italy as shown in chapter 3, and to a synthetic catalogue generated through a specific ETAS model, successfully applied to describe aftershock sequences. The results disclose a longer clustering time for real large earthquakes, suggesting that the physical process that governs aftershock sequences and the occurrence of large earthquakes may be different. Alternatively, the results can be explained by suggesting that the ETAS model, used to describe aftershock sequences, is not a suitable tool to model the seismicity as a whole.

5.1 Introduction

The spatio-temporal distribution of large earthquakes is a fundamental ingredient for seismic hazard assessment. Remarkably, despite the importance of the issue and many efforts made in the past, so far shared conclusions could not be reached and different statistical distributions are still used for large earthquake forecasting. The most striking recent example is the report concerning the seismic hazard assessment for the

Northern California (Working Group on California Earthquake Probability, 1999), where quite different (and opposite) models were used for the calculations (e.g. Poisson, Brownian Passage Time, Time Predictable).

The application of the PHM to the Italian seismicity indicate the presence of a time clustering of large earthquakes (see chapter 3 figure 3.4 and chapter 4 figure 4.6). In spite of a first order similarity with the time behaviour of the ETAS model (e.g., Ogata, 1988) used to successfully model aftershock sequences (e.g., Console *et al.*, 2003 for Italian seismicity), it is remarkable to note that the length of the time clustering for large earthquakes (few years) seems to be larger than the clustering time of aftershocks sequence. This difference, if confirmed, could have important theoretical implications. In particular, a difference in the clustering properties may suggest the existence of different physical mechanisms for aftershock and large earthquake occurrences, and/or that the specific ETAS model used to describe the aftershock sequences is not suitable to model some major feature of the seismicity.

Here we compare the clustering properties of large earthquakes found by chapter 3, with those of the ETAS model used to described seismicity of small magnitudes (Console *et al.*, 2003). In particular, we apply the PHM to a synthetic catalogue of large earthquakes generated using the ETAS model. ETAS model parameters have been estimated from the spatio-temporal distribution of the events with $M \geq 2.0$ in Italy in the time interval 1987-2000 (Console *et al.*, 2003). This allows us to prove if the hazard function coming from the synthetic catalogue is different from the one observed for real large earthquakes. In other words, we check the validity of the specific ETAS model used for aftershock sequences also to describe the spatio-temporal distribution of large events.

5.2 ETAS model and simulation of synthetic catalogue

The Epidemic Type Aftershocks-Sequences (ETAS) model is a stochastic marked point process representing the occurrence of earthquakes of size

larger than, or equal to, a threshold magnitude M_0, in a region and in a period of time (Ogata, 1988). As in other triggering models, it is based on the principle that earthquakes are clustered in time and space because of occurrence of aftershocks; but, unlike those, it solves the debated problem to find the best way to identify clusters and to classify events (between mainshocks, aftershocks, foreshocks, ...). In fact, even if it considers the overall seismicity as the superposition of a background activity and of seismicity induced by previous earthquakes, its application to real data does not request any discrimination of events. We do not discuss in details characteristics of this parametric model: a complete description of its formulation can be found in Ogata (1988, 1998).

An application of ETAS model to the Italian seismicity was provided by Console and Murru (2001) and by Console et al. (2003). Considering as marks magnitude (m_i) and epicentral coordinates (x_i, y_i), they inferred the following expression of conditional intensity function, based on history of occurrence $H_t = \{(t_i, m_i, x_i, y_i); t_i < t\}$

$$\lambda(t, m, x, y|H_t) = \left[\mu g(x, y) + \right.$$
$$\left. + \sum_{i;t_i<t} \frac{K e^{\alpha(m_i - M_0)}}{(t - t_i + c)^p} \frac{(q-1)d^{2(q-1)}}{\pi[(x - x_i)^2 + (y - y_i)^2 + d^2]^q} \right] \beta e^{-\beta(m - M_0).} \quad (5.1)$$

where $g(x, y)$ is the spatial density function of background events.

For details on formulation of function and on significance of parameters we refer to Console et al. (2003). The parameters of the model were estimated by the Maximum Likelihood Method on the national instrumental catalogue, collected by INGV (Istituto Nazionale di Geofisica e Vulcanologia), for period 1987-2000. The values obtained are: $\beta = 0.997 \cdot \ln(10)$, $\mu = 0.0613$, $K = 0.0014$, $c = 0.0068$, $p = 1.0580$, $\alpha = 0.9740$, $d = 3.07$, and $q = 1.828$.

These parameters are used to simulate a synthetic catalogue. It is developed following the thinning simulation procedure, outlined by Ogata (1981) for the Hawkes processes, of which ETAS model is an application, and than adjusted by himself to ETAS model (Ogata, 1998). In every realization events are simulated sequentially: first the time and then the

magnitude and the epicentral coordinates are obtained. This method involves simulating the time to the next event, using a rate equal to an upper bound of the intensity function, and calculating the intensity at this point. The ratio of this rate with the upper bound is compared with a uniform random number to determine if the time is retained or not. Then epicentral coordinates and magnitude are simulated according with density functions chosen.

The synthetic catalogue (from now on SC) is simulated with the same cutoff magnitude used to estimate parameters by real catalogue ($M_0 = 3.0$); then only events with magnitude larger than, or equal to, 5.5 were selected. The interval time is the same of the historical real catalogue (1600-2002) used in chapter 3 to test the PHM model. The events obtained are 131.

5.3 PHM applied to ETAS Catalogue

We apply PHM to SC to compare these results with the one obtained using the real catalogue (from now on RC). As in the previous work (see chapter 3), the Italian territory has been divided into 13×13 grid between the latitudes 36-48 N, and longitudes 5-20 E. Each node of the grid is the center of a circle. In order to cover the whole area the radius R of the circle is set about $D/\sqrt{2}$, where D is the distance between two nodes. For such grid, $D \approx 100$ km and $R = 70$ km.

In the following step, we select the circles that contain at least 3 earthquakes; 31 areas have been analysed for SC, and 29 for RC. This choice makes the statistical analysis performed for the ETAS catalogue homogeneous with the one done to the real Italian catalogue in chapter 3. Moreover, we emphasize that this technique is robust because it considers all the spatially inhomogeneous data simultaneously to build the statistical model. For each one of these circles we calculate the IETs among the earthquakes inside the circle, one CT relative to the time elapsed since the last event, and a two-dimensional vector of covariates $\mathbf{z}$, attached to each RV, bearing the information about the event itself and the area

where this is sampled. Specifically, $\mathbf{z}$ is composed by the logarithm of the rate of occurrence, i.e., the total number of clusters which occurred inside the circle divided by 403 years (1600-2002), and by the magnitude of the event from which we calculate the IET and the CT. By using equations 2.10 and 2.11 in chapter 2, we obtain $\beta_1 = 1.2 \pm 0.2$ and $\beta_2 = 0.2 \pm 0.2$. As for RC (see chapter 3) β_2 is not significantly different from zero, thus the magnitude of the events is not important for the distribution of the IETs. On the contrary, β_1 is positive and significantly different from zero.

Figure 5.1 comparises between the residuals $\varepsilon(t)$ (see equation 2.23 in chapter 2) for SC and RC, respectively, is reported. The trend of $\varepsilon(t)$ is

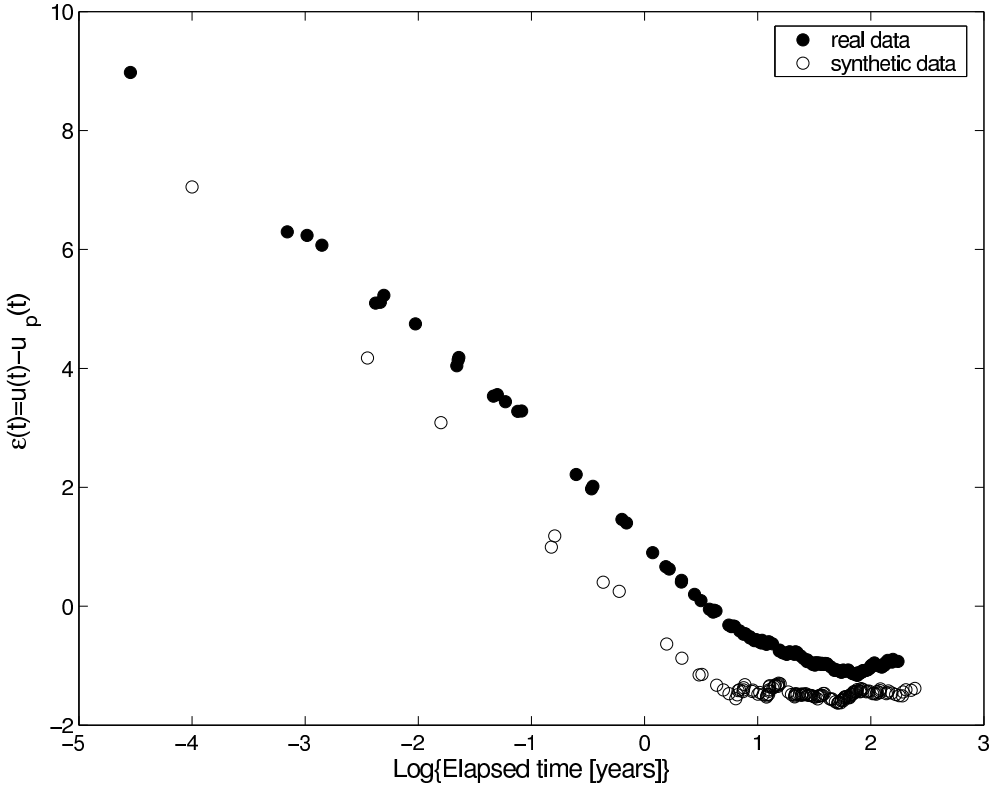

Figure 5.1: Plot of the residuals $\varepsilon(t)$ (see equation 2.23 in chapter 2) for real and synthetic catalogues as a function of the time elapsed since the most recent event.

comparable to the trend of $\lambda_0(\cdot)$; in particular the use of cumulative to calculate $\varepsilon(t)$ makes it a filtered version of $\lambda_0(t)$. In both cases a negative trend is shown. In other words, earthquakes tend to be more clustered

than in a Poisson process, where $\lambda_0(\cdot)$ is a constant. This is what we expect to get from the ETAS model, since it is constructed imposing a cluster of events.

An important result is that for SC the time clustering seems to be shorter than for RC. Both catalogues show a negative trend (indicating cluster) that becomes almost flat for longer times as in a Poisson process; figure 5.1 shows that the flattening occurs before for the SC than for RC. We suggest that this longer clustering for RC may be due to the fact that the interaction between earthquakes lasts longer than that imposed by the specific ETAS model used. Note that in both cases there is no evidence of a positive trend as expected for the gap seismic hypothesis.

A quantitative test of the difference between RC and SC is performed through a two-sample Kolmogorov-Smirnov (see Appendix G) of the empirical survivor functions of the PHM reported in figure 5.2. The null hypothesis of equal distribution is rejected at a significance level $\alpha < 0.01$.

In figure 5.3 a graphs representing the goodness-of-fit of the PHM model applied to SC is reported. Here the same strategy described in 2.3 in chapter 2 is adopted. As for the study of RC (chapter 3), the time interval 1600-1950 (109 earthquakes) is used for the *learning* phase, i.e., to set up the model. The time interval 1951-2002 (22 earthquakes), instead, is used for the *validation* phase, i.e., to verify the model with an independent data set. The goodness-of-fit is quantitatively evaluated through a one-sample Kolmogorov-Smirnov test (e.g. Gibbons, 1971). The significance levels at which the null hypothesis of equal distributions is rejected are reported in figure. They are both > 0.95.

5.4 Concluding Remarks

A main goal of this work has been to investigate on the capability of a specific ETAS model, successfully used to describe aftershock sequences, to model also the spatio-temporal distribution of large earthquakes ($M \geq 5.5$) which occurred in Italy in the last four centuries. To this purpose, we

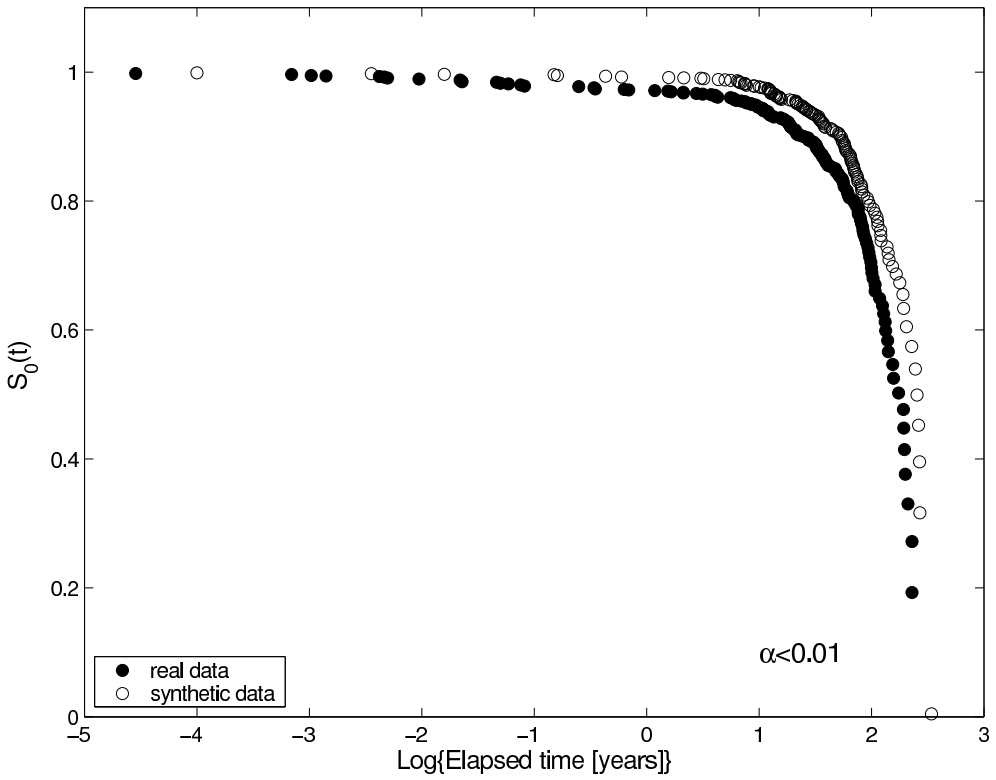

Figure 5.2: Plot of empirical survivor functions for real and synthetic catalogues. The parameter α represents the significance level at which the null hypothesis of a common parent distribution can be rejected.

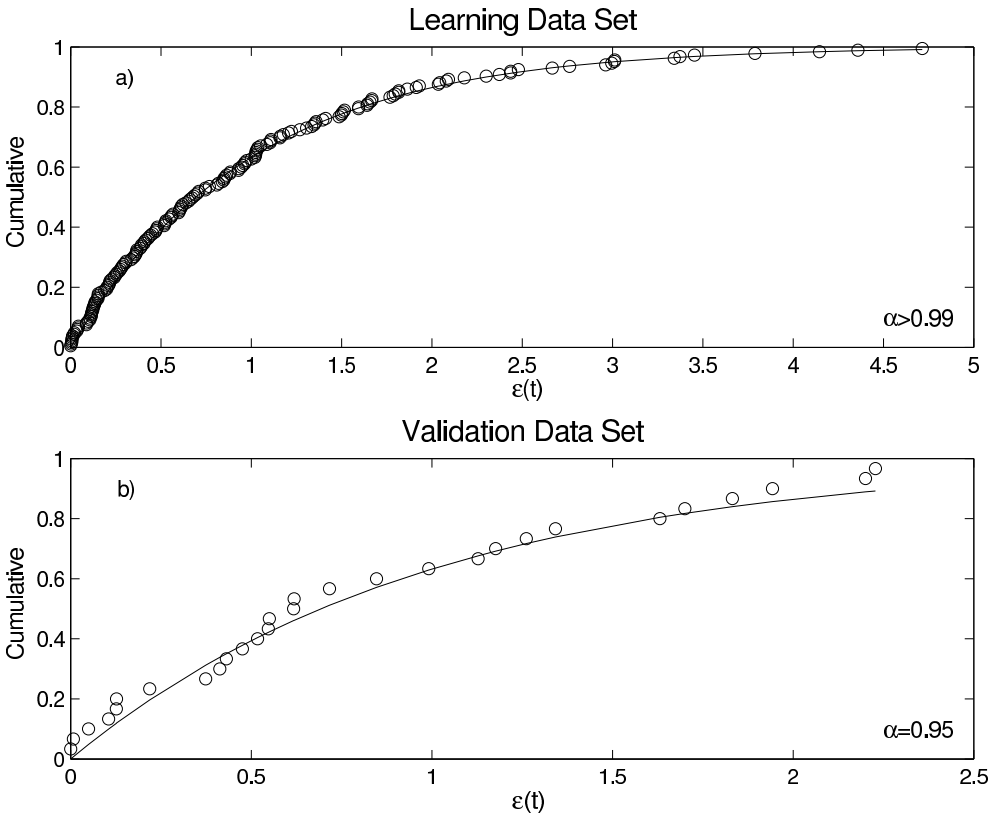

Figure 5.3: (a) Empirical and theoretical (solid line) cumulative functions for the *learning* data set. The parameter α is the significance level at which the null hypothesis (Poisson hypothesis) can be rejected (see text for more details). (b) The same as (a), but relative to the *validation* data set.

have compared the hazard functions obtained by the real catalogue and by a synthetic catalogue generated by an ETAS model, the parameters of which have been estimated by Console *et al.* (2003) by analysing aftershock sequences. The two catalogues are both clustered in time, but the length of the clustering seems to be significantly different. For the real catalogue it reaches few years and after this time the distribution of the earthquakes appears to follow a Poisson distribution. In the synthetic catalogue the cluster is shorter: the negative trend lasts for few months after the event, and then it becomes flat as in a Poisson distribution. Thus, the cluster imposed by the specific ETAS model used (described in section 5.2) seems to be shorter than the one in the real catalogue. A plausible reason might be that the interaction between earthquakes may last for longer than the typical characteristic time of aftershock sequences.

Chapter 6

Application to the whole World

In this chapter the results of applying PHM to the world seismicity are reported.
Predicting individual earthquakes is not possible at the moment (see discussions in Geller *et al.*, 1997 and Wyss, 1997), but long-term probabilistic forecasts can be validated and provide useful information for managing earthquake risk. In this picture, the aim of the work outlined in this chapter is to define the areas that may most probably generate a large earthquake in the next years. As extensively argued in chapter 1, the distribution of large events is a widely debated issue and different and sometimes antithetical models are used in earthquakes forecasting (e.g. Vere-Jones, 1970; Shimazaki & Nakata, 1980; Nishenko, 1985; Boschi *et al.*, 1995; Ellsworth *et al.*, 1998; Ogata, 1998; Kagan & Jackson, 2000; Stock & Smith, 2002; Posadas *et al.*, 2002, and many other references therein). In some stalies, the adaption of a statistical model is complemented by a proper validation of the model itself An aspect that sometimes deficits in the choosing of the statistical model is a proper validation of the model (e.g., Kagan & Jackson, 1991).

We analysed the world seismicity with M$\geq$ 7.0 from 1900 to 2004, using a geometrical grid. The main feature that characterise the spatio-temporal distribution of large earthquakes is the cluster. This means that the probability of having an earthquake is high immediately after

the event, and drops in a few years to a constant value. As previously mentioned, this behaviour is in contradiction with other models used in earthquake forecasting, like, for example, the "gap hypothesis" (e.g., McCann *et al.*, 1979), the "time-predictable" model (e.g., Shimazaki & Nakata, 1980), and the Brownian Passage Time distribution (e.g., Ellsworth *et al.*, 1998). We calculate a time-dependent probability map for the next i) one year and ii) 10 years. The time-dependent probability maps show that the areas with the highest probability of the next event occurrence are the Kurile Island, the Solomon Islands and the western cost of Mexico.

6.1 The seismic catalogue

The spatio-temporal distribution of large earthquakes in the last century (1900-2004) is investigated through the analysis of the Pacheco and Sykes catalogue (1992) for events from 1900 to 1989 and the CMT catalogue for the remaining time window (Dziewonski & Woodhouse, 1983; Dziewonski *et al.*, 1987; CMT Catalog, available at http://www.seismology.harvard.edu/projects/CMT/). Only shallow earthquakes with depth ranging from 0 to 70 km and $M_s \geq 7.0$ are used, for a total amount of 868 events. A catalogue for such time and magnitude windows is completed, as shown in figure 6.1, where the cumulative number of events as a function of time is plotted.

6.2 The grid

Earthquakes are not homogeneously spread all over the Earth surface, but their spatial distribution reflects a cluster behaviour representing tectonic and geological features. This means that the probability of having an earthquake is not the same in all the territory, but it has a marked spatial variability. For the purpose of this work, the Earth surface is divided into areas of equally areal extension. Our grid is therefore a regular grid that divides the Earth surface into cells. Each node of the grid is the center of a circle and, in order to cover the whole Earth, its radius R is set equal

to the mean value of half diagonal of the cell. In this way, it is possible to take into account the inhomogeneities in the spatial distribution of the events. We analyse all the areas where at least one event occurred in the last century. In this application we present the results for a radius $R = 300$ km that seems a reasonable dimension in relation to the spatial extension of the faults that generate $M_s \geq 7.0$ events. As an example, in figure 6.2 we report the epicentral distribution of the events in the catalogue and the areas analysed with a radius $R = 300$ km.

CMT catalogue (Dziewonski & Woodhouse, 1983; Dziewonski *et al.*, 1987; CMT Catalog, available at http://www.seismology.harvard.edu/projects/CMT/) is used to describe the tectonical inhomegeneity of such a wide territory. The purpose is to identify the tectonic regime for each selected area using the moment tensor. As shown by Kostrov (1974), the average stress tensor of a volume can be determined from the sum of the moment tensor of the events inside that region. In particular, the sum of the moment tensor is the sum of the relative component of the moment tensor of each event inside that area. Once the rake of such an average moment tensor is obtained, it is binned into three classes: prevalent compressive, prevalent extensive and prevalent strike-slip. Figure 6.3 shows the areas under study and their relative regime for a radius of $R = 300$ km. This result is stable for different radii.

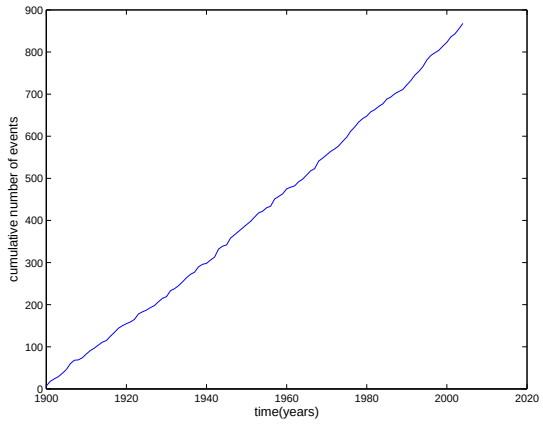

Figure 6.1: Plot of the cumulative number of events with $M_s \geq 7.0$ in the time window 1900-2004.

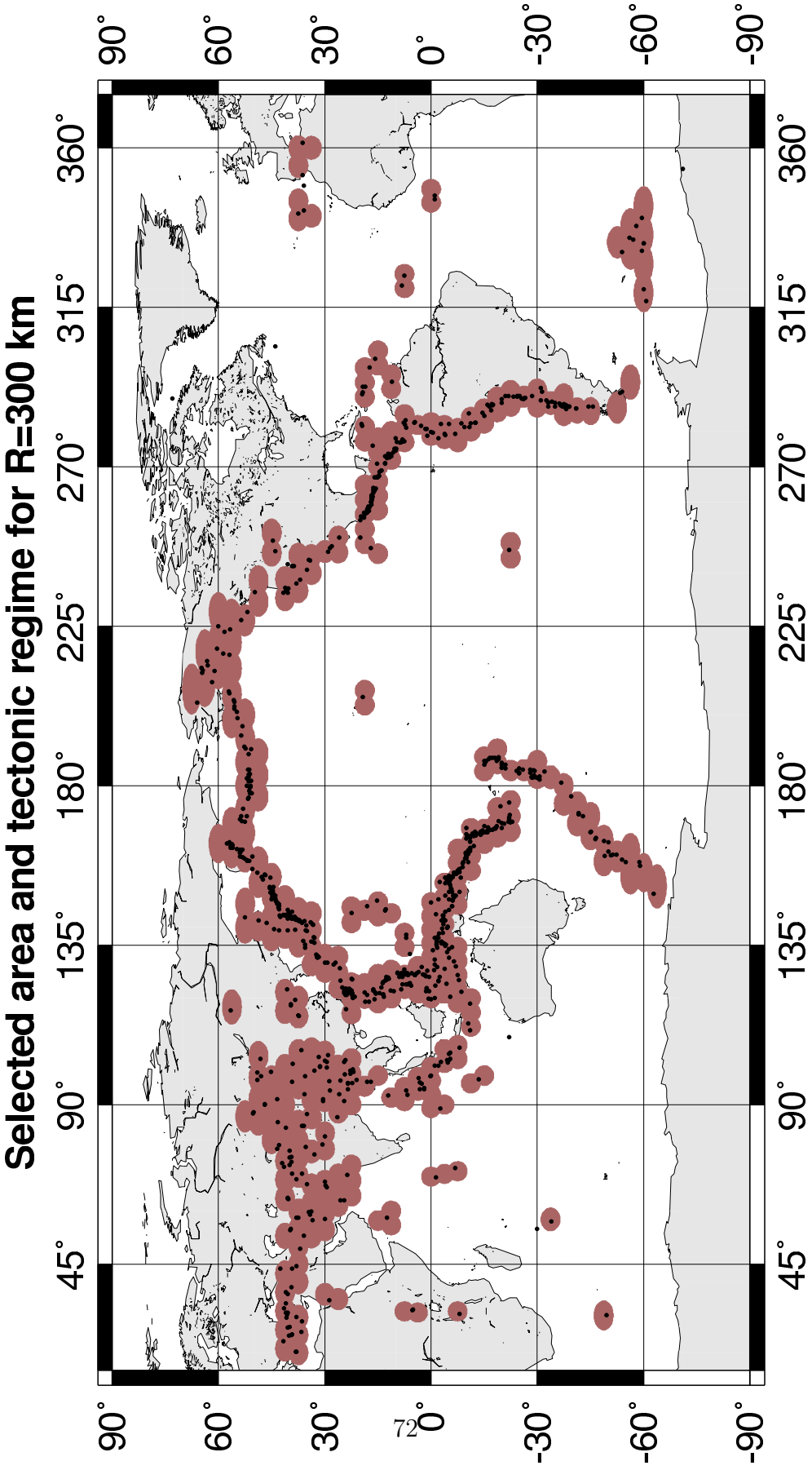

Figure 6.2: Selected areas for a radius $R = 300$km and Epicentral distribution of events with $M_s \geq 7.0$ from 1900 to 2004

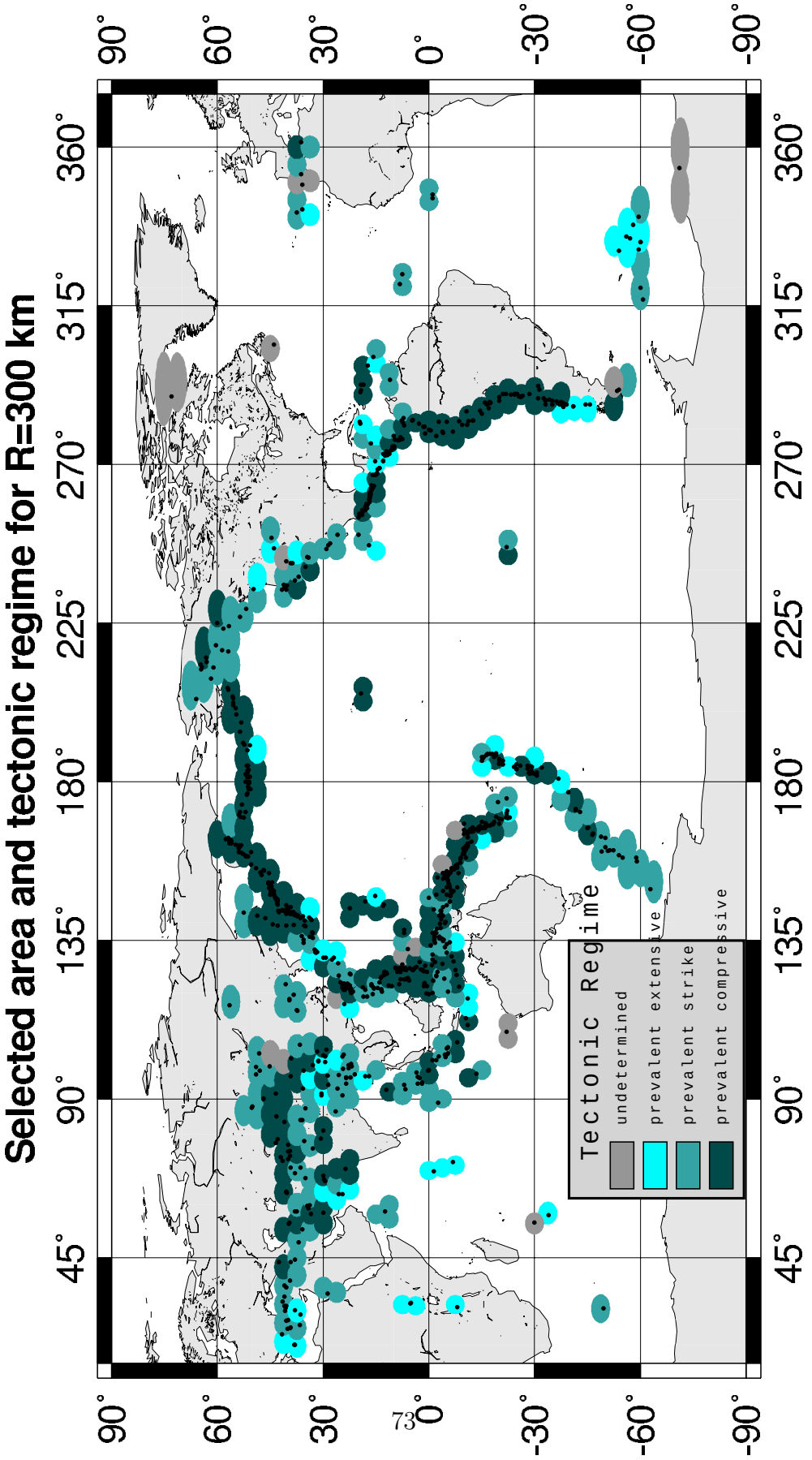

Figure 6.3: Selected areas and their prevalent tectonic regime: prevalent compressive, extensive or prevalent strike-slip. In gray, areas with undefined tectonic regime

6.3 Application to the world seismicity and results

Once the grid is set up for a particular radius R, all the areas where at least one event occurred in the last century are taken into consideration. For each of these circles we calculate i) the IETs among the earthquakes inside the area, ii) one CT relative to the time elapsed since the most recent event and the end of the catalogue, and iii) the vector of covariates $\mathbf{z}$ attached to each RV that may bear the characteristics of the event as well as the spatial/tectonic characteristics of the area where it is sampled. Here, $\mathbf{z}$ is a three-dimensional vector. Its first component is the natural logarithm of the rate of occurrence λ, calculated as the ratio of the number of sequence of $M_s \geq 7.0$ earthquakes occurred inside the areas to the time period covered by the seismic catalogue (104 years). The earthquake magnitude of the event from which we calculate the RV is the second component of $\mathbf{z}$, and the third is the prevalent stress regime within each area.

Using Maximum Likelihood Estimation strategy described in chapter 2, it turns out that only the first component of the weight vector (β_1, associated to the logarithm of the rate of occurrence z_1) is significantly different from zero, being $\beta_1 = 1.1 \pm 0.1$. This means that the rate of occurrence seems to be the only covariate (among the considered ones) that can influence the spatio-temporal distribution of large earthquakes. Remarkably, the magnitude of the previous event seems not to modify the probability of occurrence of next events, even in case of large ones. This is crucial in understanding the role played by the magnitude in earthquakes occurrence, and it neglects any kind of "time predictable" model for the spatio-temporal distribution of large events $M_s \geq 7.0$. (e.g. Ellsworth *et al.*, 1998).

Another important result is the trend of the base-line hazard function $\lambda_0(t)$ versus time, shown in figure 6.4 in terms of residuals. We remark that the time behaviour of $\lambda_0(t)$ is not imposed *a priori* by our model, but it derives directly from the data. In the considered time window, the

trend of the hazard function is a decrease for almost 10 years after an earthquake, a time window longer if compared to the one for aftershocks sequences, which may last for few weeks or months (Kagan, 1991). After that, the hazard function gradually reaches a constant as expected for a Poisson process. In practice, this means that seismic clustering is a prominent aspect of the time distribution of large earthquakes, at least over few years following a large earthquake; afterwards the process becomes almost time independent (e.g., Kagan and Jackson, 2000).

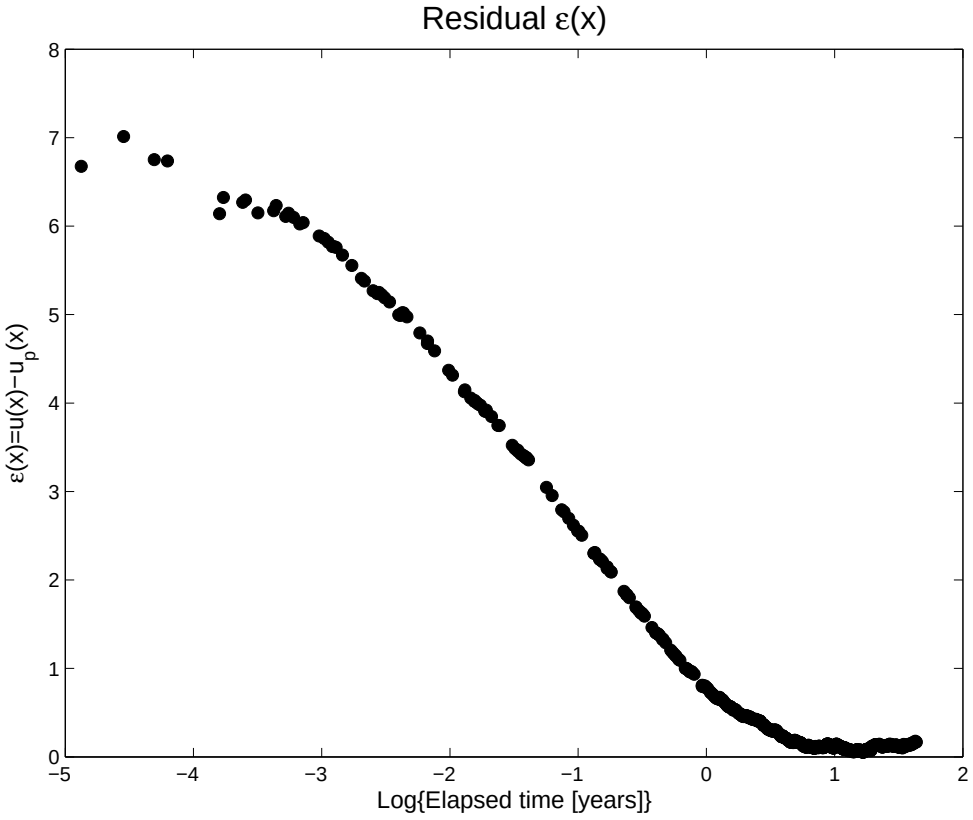

Figure 6.4: Residuals of the model (see equation 2.23 in chapter 2). The trend of the residuals is indicative of the trend of the hazard function.

In figure 6.5 the goodness-of-fit for the model is out lined. The time interval 1900-1985 forms the *learning* data set (688 events) and it has been used to set up the model; the remaining part of the catalogue, the time window 1986-2004, instead, is used to verify the model with an independent data set, the *validation* data set (180 events). The data

were normalized as explained in chapter 2 equations 2.24 and 2.25 and compared to an exponential distribution with rate set to 1. The null hypothesis H_0 of equal distributions is quantitatively evaluated through a one-sample Kolmogorov-Smirnov test (see Appendix G), and in both cases the significance level α, reported in figure, is high enough to allow to not reject H_0.

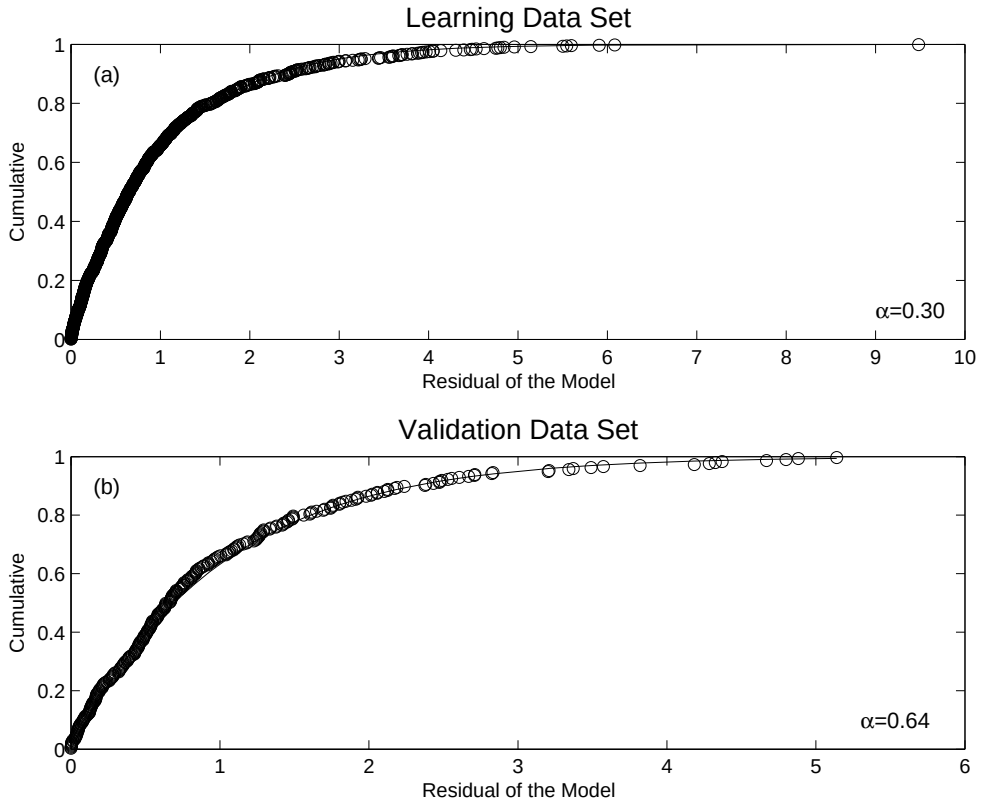

Figure 6.5: (a) Empirical (dotted line) and theoretical (solid line) cumulative functions for the *learning* data set. (b) The same as for (a), but relative to the *validation* data set.

As a final application, we evaluate the probability of occurrence in each area through equation 3.1 in chapter 3. We then construct a time-dependent probability map for the next $M_s \geq 7.0$ earthquakes in the next 1 year and 10 years, see figures 6.6 and 6.7. The areas with higher probabilities are the Kurile Island, the Solom Islands and the western cost of Mexico.

As previously introduced, the spatial distribution of the events is not

homogeneous, not even inside the regions of the world with a high density of events (e.g. the Himalayan region). Some of the regions are well constrained by the selected circles (see e.g. the western coast of South America in figure 6.2 and 6.3), while other regions present a more spread spatial distribution of the selected circles (see e.g. the Central Asia region). For these latter regions, the spatial coverage is not precise enough to grant an accurate spatial forecasting ability. Like the models based on past seismicity to forecast the future one, strong isolated earthquakes are hard to forecast, see the discussion in chapter 1.

A further implementation of PHM is to assign an error to the vector $\mathbf{z}$. For example in the areas with only one event the value of $\mathbf{z}_1$ is strongly effected by a possible lack of data in the seismic catalogue.

6.4 Other approaches to the World Seismicity

Within the Ph.D. research activity, different types of analyses about the world seismicity have been performed and are still in progress. A brief outline of just one of this analysis, the strata model, is presented in the following.

6.4.1 Strata Model

We tried to investigate the relevance of the tectonic domain in earthquakes generation. One assumption of PHM, described in chapter 2, is that the base-line hazard function $\lambda_0(\cdot)$ is the same for all the areas, and that the parameters $\mathbf{z}$ can only act multiplicatively on $\lambda_0(\cdot)$. Here we present a generalisation of PHM in which the base-line $\lambda_0(t)$ is allowed to change trend for the different value of a parameter.

PHM (equation 2.1 in chapter 2) requires a proportionality between the hazard function for any two covariate sets $\mathbf{z}_1$ and $\mathbf{z}_2$, that is

$$\lambda(t; \mathbf{z}_1) \propto \lambda(t; \mathbf{z}_2). \tag{6.1}$$

Although this situation is descriptive of many situations, sometimes

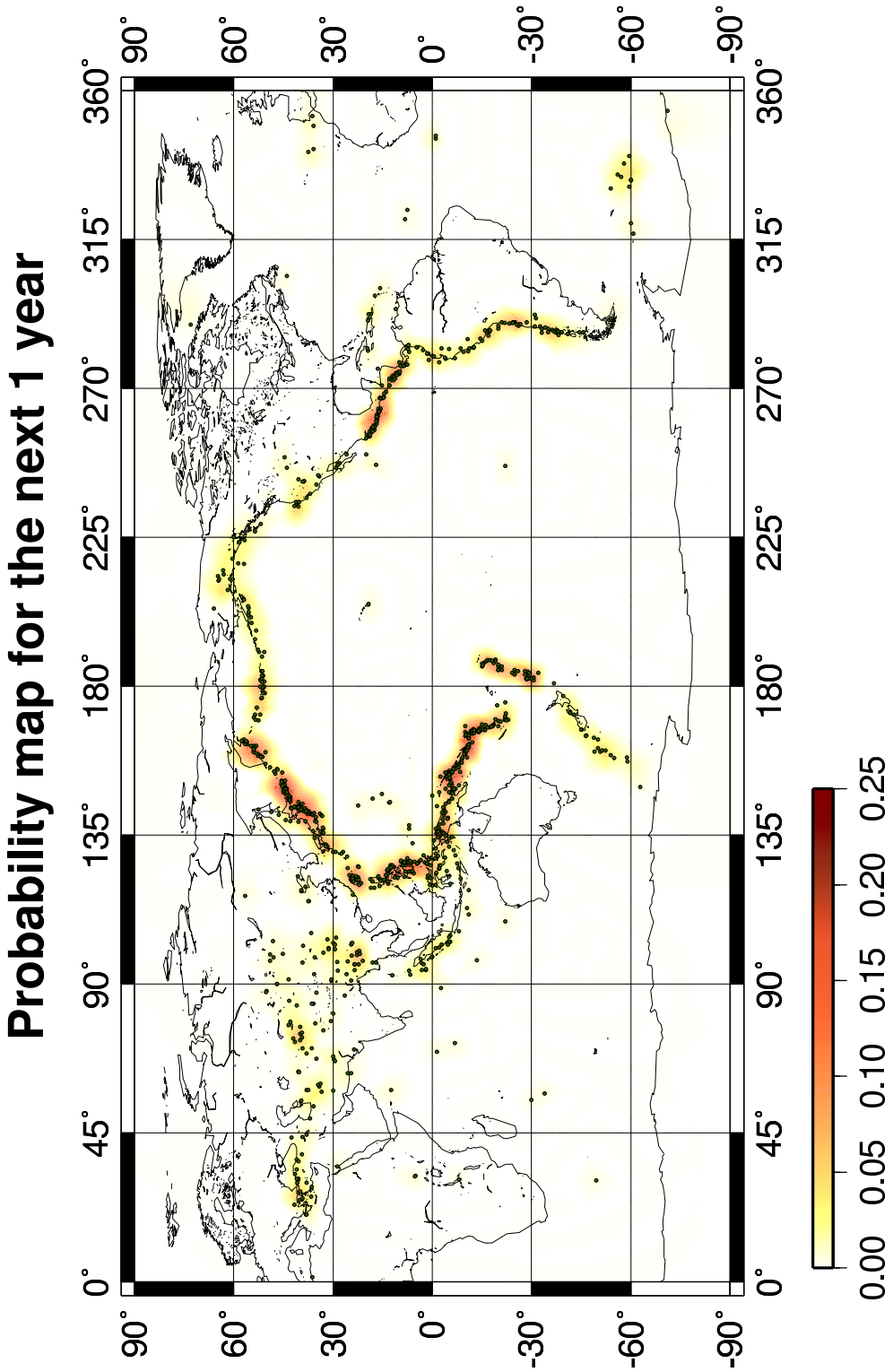

Figure 6.6: Time-dependent probabiliyy map for the next 1 year.

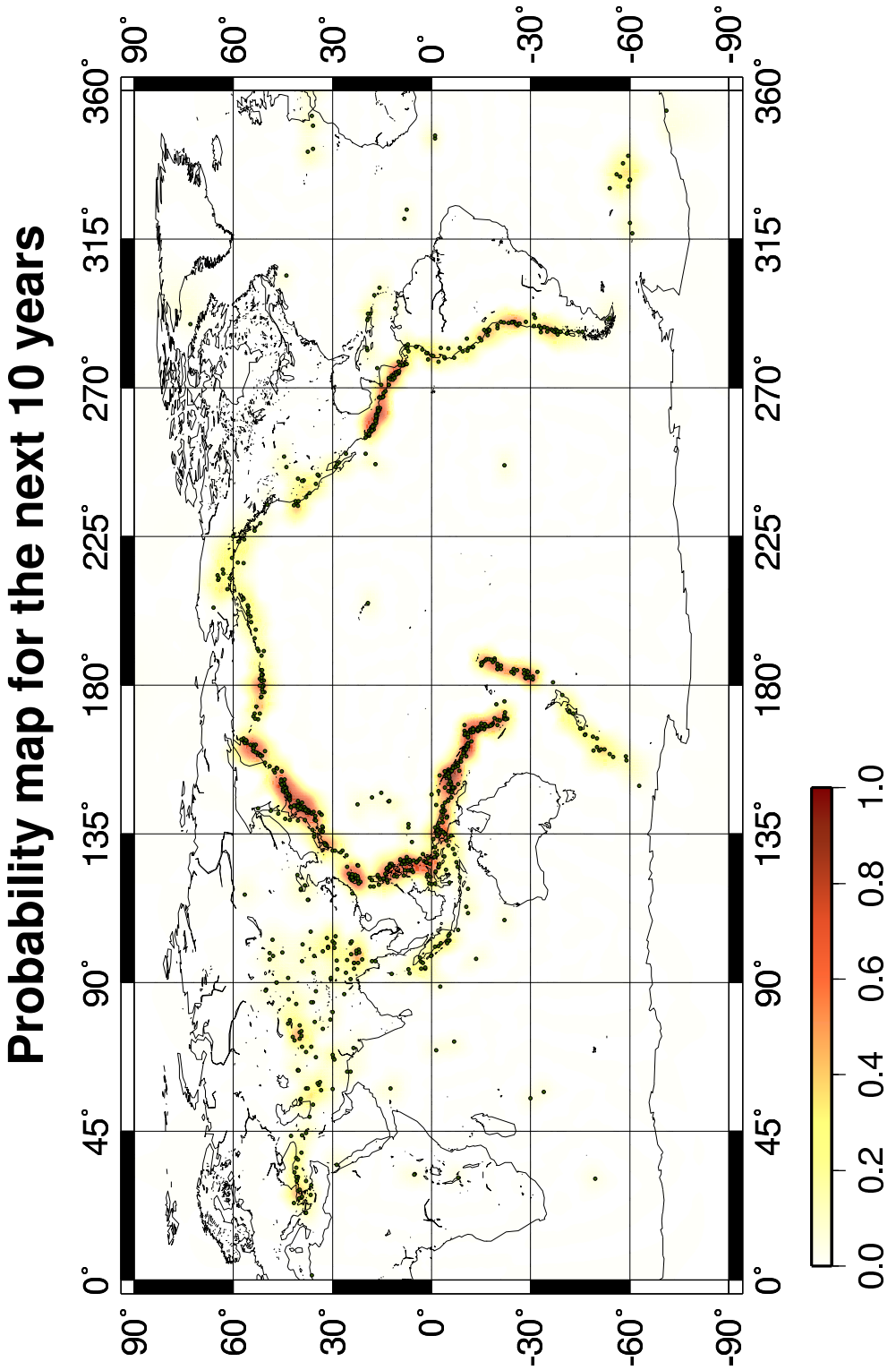

Figure 6.7: Time-dependent probability map for the next 10 years.

there are important parameters, the different values (or strata) of which produce hazard functions which differ markedly from proportionality. Consider a parameter that occurs on q strata and for which equation 6.1 may be violated. It is possible to define the hazard function for an individual in the j-th value of this parameter (Kalbfleisch & Prentice, 1980) as:

$$\lambda_j(t; \mathbf{z}) = \lambda_{0j}(t) \exp(\mathbf{z}\boldsymbol{\beta}) \tag{6.2}$$

for $j = 1, \cdots, q$ where $\mathbf{z}$ is the vector of the covariates for which equation 6.1 is descriptive. The base-line hazard functions, $\lambda_{01}(\cdot), \cdots, \lambda_{0q}(\cdot)$, for the q strata are allowed to be arbitrary and are completely unrelated.

Let $t_{j1}, \cdots, t_{jn_j}$ be the failure times on the n_j items in the j-th stratum and $\mathbf{z}_{j1}, \cdots, \mathbf{z}_{jn_j}$ be the corresponding covariates. Let $\mathbf{r}_j$ be the rank vector for the j-th stratum. The marginal likelihood of $\boldsymbol{\beta}$ is proportional to the marginal probability of $\mathbf{r}_1, \cdots, \mathbf{r}_q$, being invariant under the direct group of differentiable monotone increasing transformations acting on t inside each stratum. It can therefore be written as

$$L(\boldsymbol{\beta}) \propto \prod_{j=1}^{q} f_j(\mathbf{r}_j; \boldsymbol{\beta}), \tag{6.3}$$

where $f_j(\mathbf{r}_j; \boldsymbol{\beta})$ is the marginal likelihood (equation 2.6 in chapter 2) which arises out of the distribution of the ranks for the j-th stratum. The model can be generalized to take into account censoring and ties data (Kalbfleisch & Prentice, 1980). The general form of the marginal likelihood is

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{q} L_j(\boldsymbol{\beta}) \tag{6.4}$$

where $L_j(\boldsymbol{\beta})$ is the marginal likelihood of $\boldsymbol{\beta}$ arising from the j-th stratum alone. The first and second derivative of equation 6.4, corresponding to the score function and to the information matrix for $\boldsymbol{\beta}$, are sums over strata of equations 2.6 and 2.11 in chapter 2.

Once an extimate of $\boldsymbol{\beta}$ is obtained, the same strategy described in section 2.2 in chapter 2 can be used to give estimates of the survivor functions in each of the q strata separately.

In the application to the world seismicity, the same procedure described above has been repeated for different input area, and consequently different radii, from a minimum to $R = 300$ km to a maximum to $R = 2500$ km. Using Kostrov (1974) strategy, a prevalent tectonic stile was identified for each area (see above): compressive, extensive and strike-slip. The survivor functions of the three strata were compared with a two-sample Kolmogorov-Smirnov test (see Appendix G), showing no statistical differences between them.

Appendix A

Failure Time Distribution

The subject of our study is the modeling and the analysis of data that have as a principal end point the time until an event occurs. Such events are generically referred to as failures through the events.

Let T be a nonnegative random variable representing the failure time of an individual from a homogeneous population. The probability distribution of T can be specified in many ways, three of which are particularly useful in survival applications: the survivor function, the probability density function (pdf), and the hazard function. Interrelations between these three representations are given below for both discrete and continuous distributions.

The survivor function is defined for both discrete and continuous distributions as the probability that T is at least as great as a value t; that is,

$$S(t) = P(T \geq t), \qquad 0 < t < \infty. \tag{A.1}$$

Clearly $S(t)$ is a monotone nonincreasing function with $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$. The pdf and the hazard function are most easily specified separately for discrete and continuous T.

A.1 T (Absolutely) Continuous

The pdf of T is

$$f(t) = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t)}{\Delta t}$$
$$= \frac{-dS(t)}{dt}. \tag{A.2}$$

Conversely, $S(t) = \int_t^\infty f(u)du$ and $f(t) \ge 0$ with $\int_0^\infty f(t)dt = 1$. The domain of T is $[0, \infty)$. The hazard function specifies the instantaneous rate of failure at $T = t$ conditional upon survival to time t and is defined as

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}$$
$$= \frac{f(t)}{S(t)}. \tag{A.3}$$

It is easily seen that $\lambda(t)$ specifies the distribution of T since, from A.2,

$$\lambda(t) = \frac{-d \log S(t)}{dt} \tag{A.4}$$

and, *vice versa*, integrating and using $S(0) = 1$, we obtain

$$S(t) = \exp\big(-\int_0^t \lambda(u)du\big). \tag{A.5}$$

The pdf of T can be written

$$f(t) = \lambda(t) \exp\big(-\int_0^t \lambda(u)du\big). \tag{A.6}$$

Examination of A.5 indicates that $\lambda(t)$ is a nonnegative function with

$$\int_0^s \lambda(u)du < \infty \tag{A.7}$$

for some $s > 0$, and $\int_0^\infty \lambda(u)du = \infty$.

A.2 T Discrete

If T is a discrete random variable taking values $t_1, t_2, \cdots$ with associated probability function

$$f(t_i) = P(T = t_i), \qquad i = 1, 2, \ldots, \tag{A.8}$$

then the survivor function is

$$\begin{aligned} S(t) &= \sum_{j|t_j \geq t} f(t_j) \\ &= \sum f(t_j) H(t_j - t), \end{aligned} \tag{A.9}$$

where $H(t)$ is the Heaviside function

$$H(t) = \begin{cases} 0, & t < 0 \\ 1, & t \geq 0. \end{cases} \tag{A.10}$$

The hazard at t_j is defined as the condition probability of failure at t_j,

$$\begin{aligned} \lambda_j &= P(T = t_j | T \geq t_j) \\ &= \frac{f(t_j)}{S(t_j)}, \qquad j = 1, 2, \ldots. \end{aligned} \tag{A.11}$$

Corresponding to A.5 and A.6, the survivor function and the pdf are given by

$$S(t) = \prod_{j|t_j < t} (1 - \lambda_j) \tag{A.12}$$

and

$$f(t_j) = \lambda_j \prod_{1}^{j-1} (1 - \lambda_i). \tag{A.13}$$

Equation A.12 can be motivated as follows: $1 - \lambda_j$ is the conditional probability of surviving past time t_j, given survival up to t_j, since the hazard function is the probability of failure at time t_j, given that one has survived to t_j. The product corresponds to multiplying these conditional probabilities of not failure for all known failure times from zero up to the time of interest.

Appendix B

Some Continuous Parametric Distributions

In this appendix a gallery of the most significant and widely used statistical distributions in earthquake forecasting is shown, emphasising their characteristics.
In the following, let T be a random variable representing failure time and t a specific value in its range.

B.1 Exponential Distribution

The formula for the one-parameter probability density function (pdf) of the exponential distribution is

$$f(t) = \lambda \exp(-\lambda t) \qquad \lambda > 0 \tag{B.1}$$

where $\lambda = \frac{1}{\theta}$ and θ is the scale parameter, and also the mean and standard deviation of the distribution. The cumulative function is

$$F(t) = \int_0^t \lambda \exp(-\lambda u) du = 1 - \exp(-\lambda t) \qquad \lambda > 0, \tag{B.2}$$

where $F(\infty) = 1$, so that the total probability is 1, as it should be; the survivor function is

$$S(t) = 1 - F(t) = \exp(-\lambda t); \tag{B.3}$$

and the hazard function is

$$\lambda(t) = \frac{f(t)}{S(t)} = \lambda. \tag{B.4}$$

The feature of exponential distribution is that the hazard function is a constant over the range of T; that is the instantaneous failure rate is independent of t. The conditional probability of failure in a time interval is independent from the time since the previous event; this is referred to as the memoryless property of the exponential distribution.
In figure B.1 the pdf, cumulative, survivor and hazard functions for $\lambda = 0.5$ are shown.

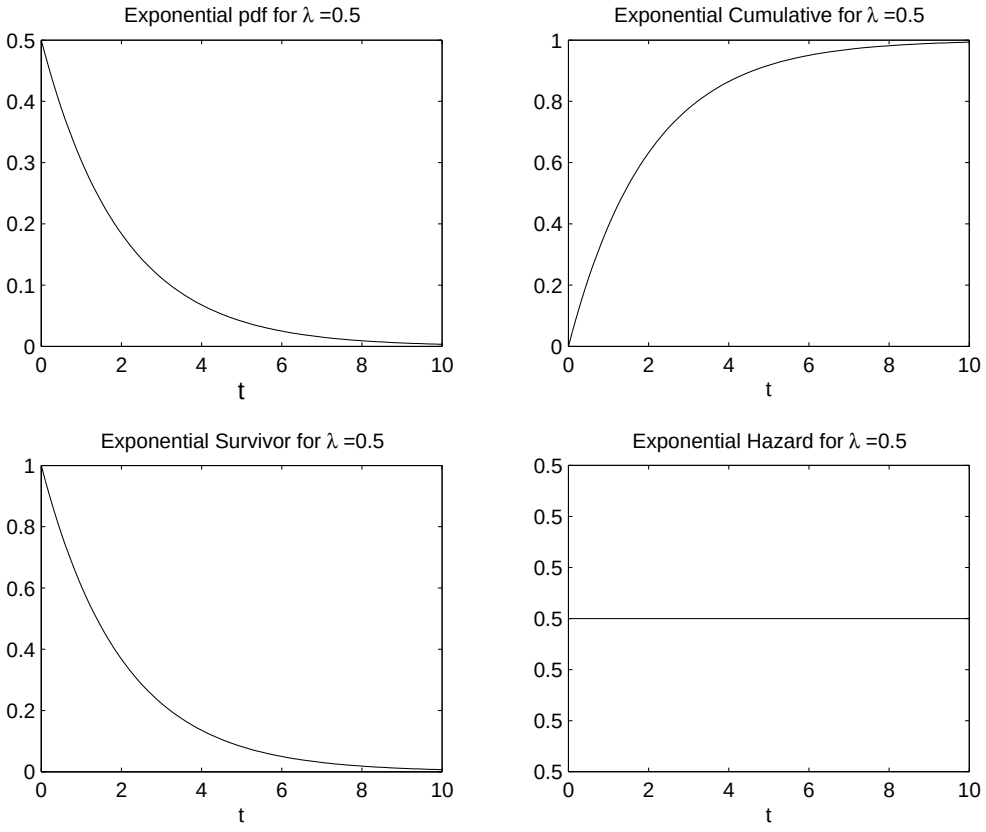

Figure B.1: The pdf, cumulative, survivor and hazard functions for an exponential distribution with $\lambda = 0.5$.

B.2 Weibull Distribution

The Weibull distribution is extensively used in reliability application to model failure time, especially in engineering applications. The formula

for the pdf of the two-parameter Weibull distribution is

$$f(t) = \lambda\beta(\lambda t)^{\beta-1}\exp[-(\lambda t)^{\beta}] \qquad \lambda, \beta > 0, \tag{B.5}$$

where λ is the inverse of the scale parameter θ, and β is shape parameter. The cumulative and survivor functions are:

$$F(t) = 1 - \exp[-(\lambda t)^{\beta}] \tag{B.6}$$

$$S(t) = \exp[-(\lambda t)^{\beta}] \tag{B.7}$$

and the hazard function is

$$\lambda(t) = \lambda\beta(\lambda t)^{\beta-1}. \tag{B.8}$$

The Weibull distribution is built imposing a failure rate $\lambda(t)$ that changes as a power of the age. Weibull hazard function changes shape in relation to the value of β. In particular, it is monotone decreasing for $\beta < 1$, increasing for $\beta > 1$, and reduces to the constant hazard (i.e., it becomes an exponential distribution) for $\beta = 1$. Figures B.2 and B.3 show the pdf, cumulative, survivor and hazard functions for different value of β: 0.5 , 1, 2 and 5.

B.3 Lognormal Distribution

A random variable T is lognormally distributed with scale parameter θ and shape parameter β, if $Y = \ln T$ is normally distributed with mean $\mu = \ln(\theta)$ and standard deviation $\sigma = \frac{1}{\beta}$. The general form of the pdf of the lognormal distribution is

$$f(t) = \frac{1}{\sqrt{2\pi}\beta t}\exp\big(-\frac{(\ln(\frac{t}{\theta}))^2}{2\beta^2}\big) \qquad \theta, \beta > 0 \tag{B.9}$$

where the scale parameter θ is also the median of the distribution. The cumulative, survivor and hazard functions involves the normal integral

$$\Phi(t) = \int_0^t \phi(u)du \tag{B.10}$$

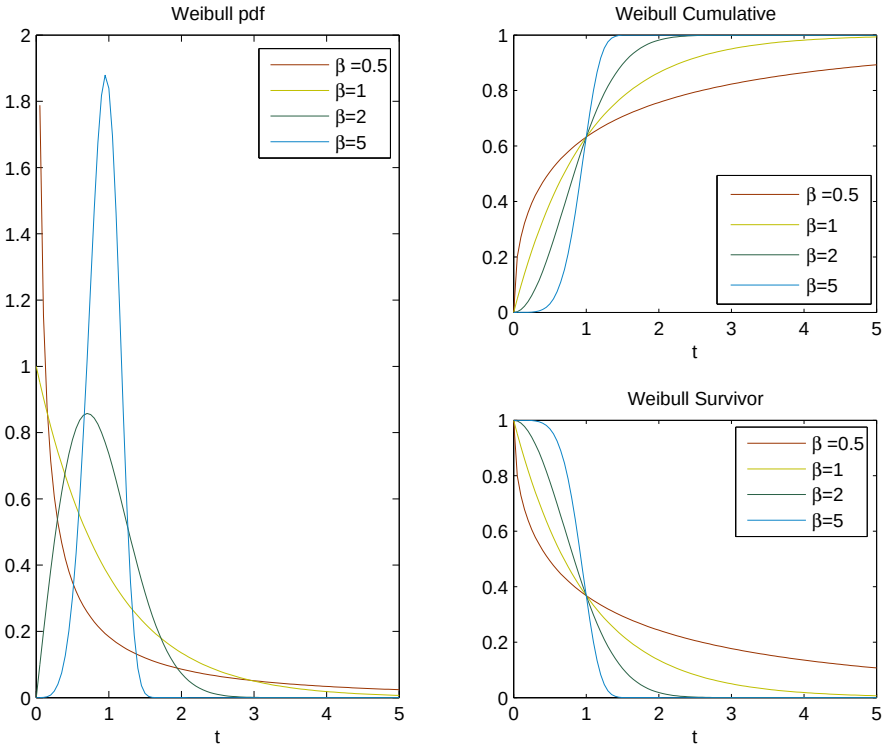

Figure B.2: The pdf, cumulative and survivor functions for a Weibull distribution with β equal to 0.5, 1, 2 and 5.

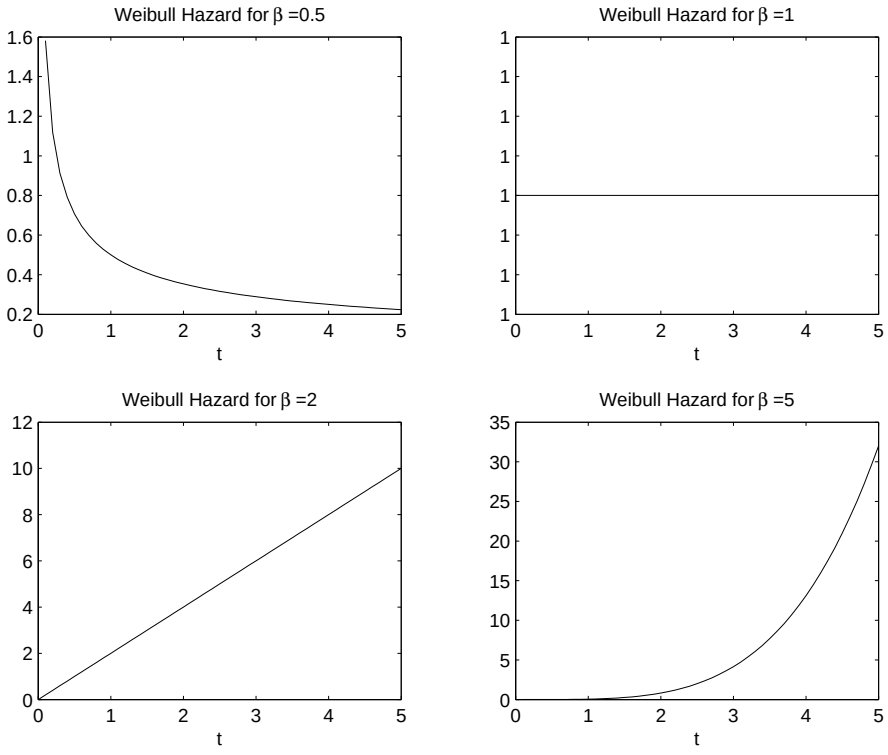

Figure B.3: The hazard function for a Weibull distribution with β equal to 0.5, 1, 2 and 5.

where $\phi(u)$ and $\Phi(u)$ are the pdf and cumulative of a standard normal distribution, respectively; they can be written as

$$F(t) = \Phi(\frac{\ln(\frac{t}{\theta})}{\beta}) \tag{B.11}$$

$$S(t) = 1 - \Phi(\frac{\ln(\frac{t}{\theta})}{\beta}) \tag{B.12}$$

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{B.13}$$

Lognormal hazard function is fairly complicated; as t increases, $\lambda(t)$ first increases to a maximum and then decreases, approaching zero as t becomes large.

Figures B.4 and B.5 illustrate the lognormal distribution for different shape parameters β.

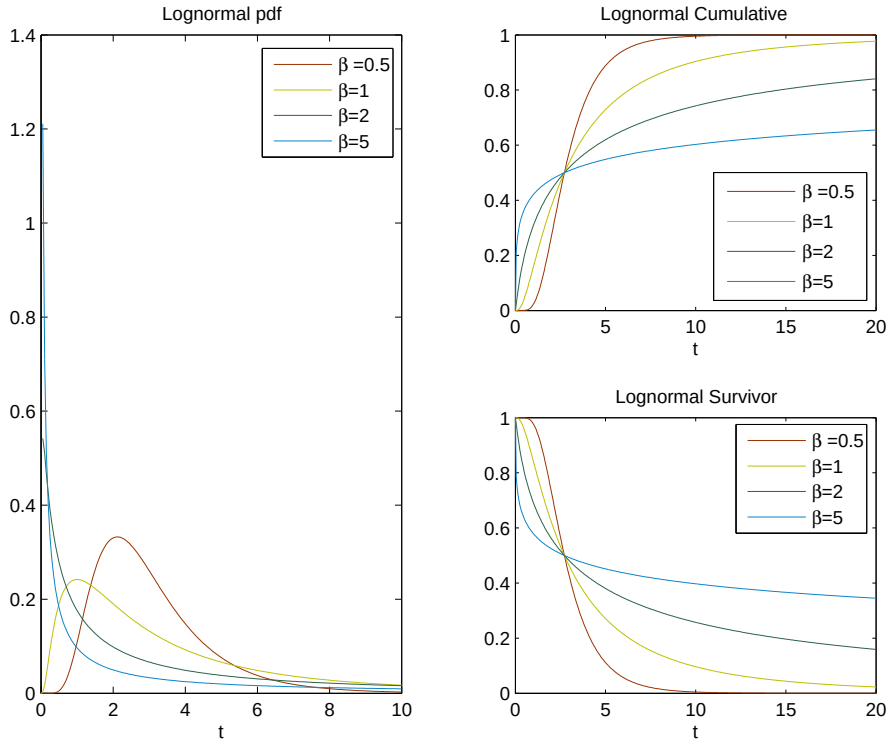

Figure B.4: The pdf, cumulative and survivor functions for a lognormal distribution with β equal to 0.5, 1, 2 and 5.

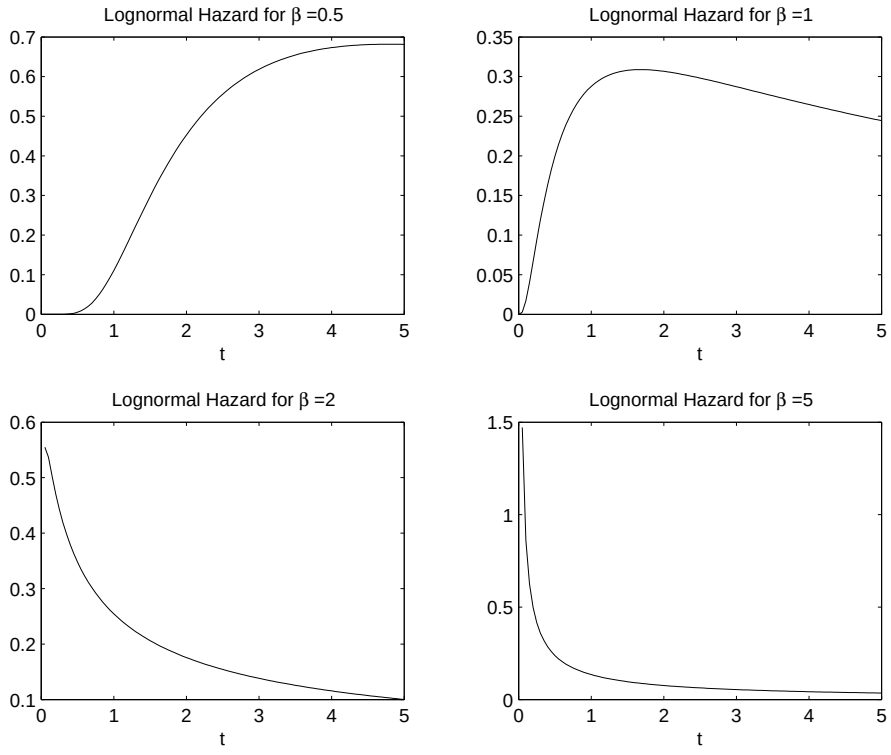

Figure B.5: The hazard function for a lognormal distribution with β equal to 0.5, 1, 2 and 5.

B.4 Brownian Passage Time Distribution

In the Brownian Passage Time model (BPT), the recurrence process is modelled by adding a Brownian perturbation to a steady tectonic loading. The BPT density function is defined as (Matthews *et al.*, 2002)

$$f(t;\mu,\alpha) = \Big(\frac{\mu}{2\pi\alpha^2 t^3}\Big)^{1/2} \exp\Big(-\frac{(t-\mu)^2}{2\mu\alpha^2 t}\Big) \tag{B.14}$$

where μ is the mean and α the shape parameter. The cumulative distribution function may be expressed in terms of normal cumulative Gaussian distribution (see equation B.10). Defining

$$\begin{aligned} u_1(t) &= \alpha^{-1}[t^{1/2}\mu^{-1/2} - t^{-1/2}\mu 1/2] \\ u_2(t) &= \alpha^{-1}[t^{1/2}\mu^{-1/2} + t^{-1/2}\mu^{1/2}] \end{aligned} \tag{B.15}$$

the cumulative is

$$F(t) = \Phi[u_1(y)] + \exp(2/\alpha^2)\Phi[-u_2(t)]. \tag{B.16}$$

The shape parameter α is the ratio of the standard deviation to the mean, since $\mathrm{var}(T) = \sigma_T^2 = (\mu\alpha)^2$, and it is also called the coefficient of variation of the distribution. The survivor and hazard functions can be expressed as a function of the pdf and cumulative, that is

$$S(t) = 1 - F(t) = 1 - \Phi[u_1(y)] + \exp(2/\alpha^2) \quad \Phi[-u_2(t)] \tag{B.17}$$

and

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{B.18}$$

Figure B.6 shows the shape of the BTP distribution for $\mu = 1$ and different values of α.

The BPT hazard function is always zero immediately after the rupture for $t = 0$; then it increases to a finite maximum value near the mean recurrence time and then decrease asymptotically to a quasi-stationary level, in which the conditional probability of event becomes time independent. In figure B.7 the trend of the hazard function versus time is reported for different value of α.

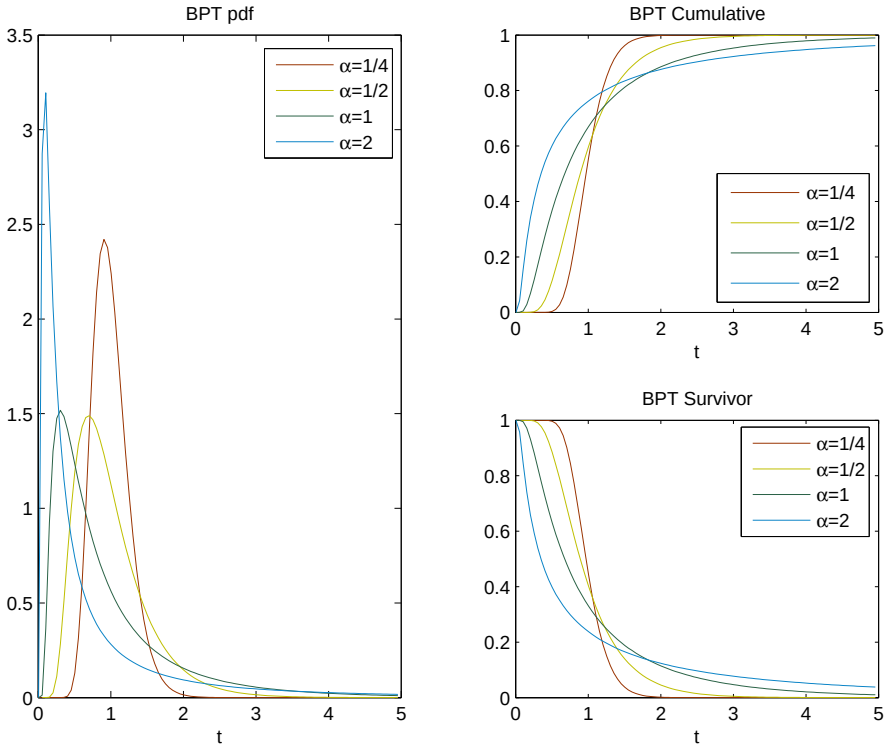

Figure B.6: The pdf, cumulative and survivor functions with α equal to 0.25, 0.5, 1 and 2.

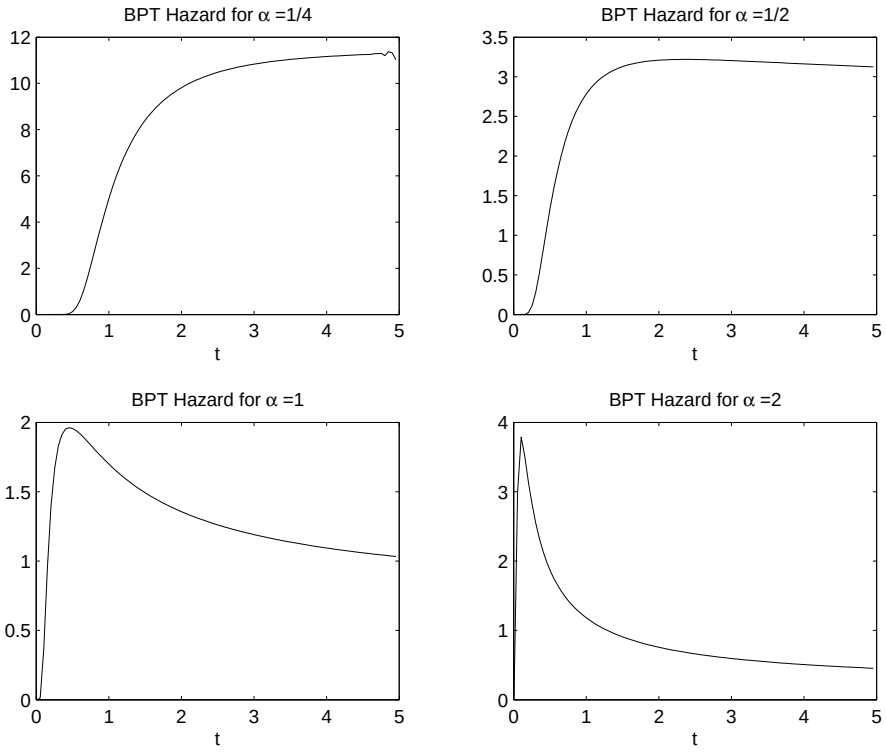

Figure B.7: The hazard function with α equal to 0.25, 0.5, 1 and 2.

Appendix C

The Likelihood of the Proportional Hazard Function

Let us suppose N IET $t_{(i)}$; $(i = 1, \ldots, N)$ with corresponding covariates $\mathbf{z}_{(i)}$ and no censoring time in the data. It is possible to define two vectors $\mathbf{o}(\mathbf{t}) = [t_{(i)}, \ldots, t_{(N)}]$ and $\mathbf{r}(\mathbf{t}) = [(1), \ldots, (N)]$. The order statistic $\mathbf{o}(\mathbf{t})$ refers to the $t_{(i)}$'s ordered from the smallest to the larger and the notation (i) in the rank statistic $\mathbf{r}(\mathbf{t})$ refers to the label attached to the i-th order statistic. For the Proportional Hazard Model (PHM) the rank statistic is sufficient for the estimation of $\boldsymbol{\beta}$. In fact, let's consider the PHM

$$\lambda(t; \mathbf{z}) = \lambda_0(t) \exp(\mathbf{z}\boldsymbol{\beta}) \tag{C.1}$$

in chapter 2 equation 2.1, and define a transformation $u = g^{-1}(t)$, where $g \in G$, the group of strictly increasing and differentiable transformation of $(0, \infty)$ onto $(0, \infty)$. The hazard function, given $\mathbf{z}$, can be written as:

$$\lambda_i(u) \exp(\mathbf{z}\boldsymbol{\beta}) \tag{C.2}$$

where $\lambda_i(u) = \lambda_0(g(u))g'(u)$. Thus if the data were presented in form of $u_{(i)}, \ldots, u_{(N)}$ and $\mathbf{z}_{(i)}, \ldots, \mathbf{z}_{(N)}$ where $g(u_{(i)}) = t_{(i)}$, the inference problem about $\boldsymbol{\beta}$ would be the same under the group G of transformation of the IETs. Moreover, if we consider the action of G on the sample space, the order statistic $\mathbf{o}(\mathbf{t})$ can be mapped to any specific order statistic while the rank statistic $\mathbf{r}(\mathbf{t})$ is left unchanged. Therefore, any specific order statistic can clearly be obtained for u by a judicious choice of $g \in G$. Since the estimation of $\boldsymbol{\beta}$ is the same under any such transformation, and since the

order statistic can be made arbitrary by such a transformation, only the rank statistic can carry information about $\boldsymbol{\beta}$ where $\lambda_0(t)$ is unknown; the rank statistic is said to be marginally sufficient for the estimation of $\boldsymbol{\beta}$ (Barnard, 1963). In this case, the likelihood is proportional to the probability that the rank vector $\mathbf{r}(\mathbf{t})$ should be the observed one, that is

$$\begin{aligned} P(\mathbf{r};\boldsymbol{\beta}) &= P\Big[\mathbf{r} = [(1),\cdots,(N)];\boldsymbol{\beta}\Big] \\ &= \int_0^\infty \int_{t_{(i)}}^\infty \cdots \int_{t_{(N-1)}}^\infty \prod_{i=1}^{N} f(t_{(i)};\mathbf{z}_{(i)}) dt_{(N)} \ldots dt_{(1)} \end{aligned} \tag{C.3}$$

where, $\mathbf{z}$ is a row vector of s measured covariates, $\boldsymbol{\beta}$ is a column vector of s regression parameters, and t is the associated IET. The pdf is

$$f(t;\mathbf{z}) = \lambda(t;\mathbf{z}) S(t;\mathbf{z}) \tag{C.4}$$

and for the PHM the hazard and the survivor functions are

$$\lambda(t;\mathbf{z}) = \lambda_0(t)\exp(\mathbf{z}\boldsymbol{\beta}) \tag{C.5}$$

$$S(t;\mathbf{z}) = \exp[-\int_0^t \lambda_0(u) e^{\mathbf{z}\boldsymbol{\beta}} du]. \tag{C.6}$$

Since all the variables are independent, let's consider the inner integral of equation C.3

$$\int_{t_{(N-1)}}^\infty f(t_{(N)};\mathbf{z}_{(N)}) dt_{(N)} =$$

$$\int_{t_{(N-1)}}^\infty \lambda_0(t_{(N)}) e^{\mathbf{z}_{(N)}\boldsymbol{\beta}} \exp\left[-\int_0^{t_{(N)}} \lambda_0(u) e^{\mathbf{z}_{(N)}\boldsymbol{\beta}} du\right] dt_{(N)}$$

Defining $\gamma_{(i)} = e^{\mathbf{z}_{(i)}\boldsymbol{\beta}}$ and $\Delta_0(t_{(i)}) = \int_0^{t_{(i)}} \lambda_0(u) du$, the equation becomes

$$\begin{aligned} \int_{t_{(N-1)}}^\infty \lambda_0(t_{(N)}) \gamma_{(N)} e^{-\Delta_0(t_{(N)})\gamma_{(N)}} dt_{(N)} &= \Big[-e^{\gamma_{(N)}\Delta_0(t_{(N)})}\Big]_{t_{(N-1)}}^\infty \\ &= e^{-\gamma_{(N)}\Delta_0(t_{(N-1)})} \end{aligned} \tag{C.7}$$

since $\Delta_0(\infty) = \infty$ for definition of hazard function. Let's now consider the next integral

$$\int_{t_{(N-2)}}^\infty \lambda_0(t_{(N-1)}) \gamma_{(N-1)} e^{-\Delta_0(t_{(N-1)})\gamma_{(N-1)}} e^{-\Delta_0(t_{(N-1)})\gamma_{(N)}} dt_{(N-1)} \tag{C.8}$$

where the last part of the integral derives from equation C.7. Let u be the integration variable

$$\int_{t_{(N-2)}}^{\infty} \lambda_0(u)\gamma_{(N-1)}e^{-\Delta_0(u)[\gamma_{(N-1)}+\gamma_{(N)}]}du = \tag{C.9}$$

$$= \Bigg[- e^{-\Delta_0(u)[\gamma_{(N-1)}+\gamma_{(N)}]} \Bigg]_{x_{(N-2)}}^{\infty} \frac{\gamma_{(N-1)}}{\gamma_{(N-1)} + \gamma_{(N)}} \tag{C.10}$$

that is the solution for the first and second integral of equation C.3. Equation C.10 can be easily generalized for a generic (i)

$$\Bigg[- e^{-\Delta_0(u)\sum_{j=(i)}^{(N)}\gamma_{(j)}} \Bigg]_{t_{(i-1)}}^{\infty} \prod_{k=(i)}^{(N)} \frac{\gamma_{(k)}}{\sum_{j=(k)}^{(N)}\gamma_{(j)}}, \tag{C.11}$$

that is the solution for the integral from the firth up to the i-th. The last integral, for $(i) = (1)$, will be

$$\Bigg[- e^{-\Delta_0(u)\sum_{j=(1)}^{(N)}\gamma_{(j)}} \Bigg]_{0}^{\infty} \prod_{k=(1)}^{(N)} \frac{\gamma_{(k)}}{\sum_{j=(k)}^{(N)}\gamma_{(j)}}. \tag{C.12}$$

The left part of the integral of equation C.12 is equal to 1, since $\Delta_0(\infty) = \infty$ and $\Delta_0(0) = 0$, thus, from equation C.12, we get

$$\prod_{i=(1)}^{(N)} \frac{\gamma_{(i)}}{\sum_{j\in R(t_{(i)})}\gamma_{(j)}} \tag{C.13}$$

that is equal to

$$\prod_{i=(1)}^{(N)} \frac{e^{\mathbf{z}_{(i)}\boldsymbol{\beta}}}{\sum_{j\in R(t_{(i)})} e^{\mathbf{z}_{(j)}\boldsymbol{\beta}}} \tag{C.14}$$

where $R(t_{(i)}) = (i), (i+1), \ldots, (N)$.

C.1 Inclusion of Censoring Time in the Data Set

Let us now consider the probability of having CTs in the data set. The inclusion of the censoring data sets carries some modification in the rank statistic, since censored samples bear partial information in the rank. N_1

are the IETs $(t_{(1)}, \ldots, t_{(N_1)})$, with corresponding covariates $\mathbf{z}_{(1)}, \ldots, \mathbf{z}_{(N_1)}$; N_2 are the CTs $\tilde{t}$, of which n_i are censored in the i-th interval $[t_{(i)}, t_{(i+1)})$, i.e. $t_{(i)} \leq \tilde{t}_{i1}, \ldots, \tilde{t}_{in_i} (i = 0, \ldots, N_1)$, where $t_{(0)} = 0$ and $t_{(N_1+1)} = \infty$ and with covariates $\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i}$. For sake of clarity we refer to subscript inside brackets for IET data (i.e. $t_{(i)}$) and the one without brackets for censoring data (i.e. t_i). The marginal likelihood of $\boldsymbol{\beta}$ is computed as the probability that the rank vector should be one of those possible in the sample and is, therefore, the sum of a large number of items like equation C.14. But the rank vector is characterised by

$$\begin{aligned} t_{(i)} &< \ldots, t_{N_1} \\ t_{(i)} < t_{i1}, \ldots, t_{im_i} &\qquad (i = 0, \ldots, N_1) \end{aligned} \tag{C.15}$$

where $t_{i1}, \ldots, t_{im_i}$ are the unobserved failure time associated with individual censored in $[t_{(i)}, t_{(i+1)})$. Censoring events like C.15 allow simple computation of the marginal likelihood since, given $t_{(i)}$, the event $t_{(i)} < t_{i1}, \ldots, t_{im1}$ has the conditional probability

$$g(t_{(i)}) = \exp\Big[- \sum_{j=i}^{n_i} \exp(\mathbf{z}_{ij}\boldsymbol{\beta}) \int_0^{t_{(i)}} \lambda_0(u) du \Big] \tag{C.16}$$

that is the contribution to the likelihood of an individual censored at $t_{(i)}$ is the survivor function. Here we are assuming that the observed censoring time $t_{ij} (j = 1, \ldots, N_i)$ tells us only that the unobserved failure time is greater than t_{ij}. Therefore the marginal likelihood can be written as

$$P(\mathbf{r}; \boldsymbol{\beta}) = \int_0^{\infty} \int_{t_{(1)}}^{\infty} \cdots \int_{t_{(N_1-1)}}^{\infty} \prod_{i=1}^{N_1} f(t_{(i)}; \mathbf{z}_{(i)}) g(t_{(i)}) dt_{(N_1)} \ldots dt_{(1)}. \tag{C.17}$$

To solve this equation, a procedure similar to the one used in case of IET only can be used, firstly solving the inner integral, and then proceeding with the next till the last one. The contribution of i-th integral's solution to the $(i+1)$-th integral contains the covariates of the ITEs and the CTs; similarly to equation C.12 a generic form for the i-th integral is:

$$\Big[- \exp\big[- \Delta_0(u) \sum_{k=i}^{N_1} (\gamma_{(k)} + \sum_{j=1}^{n_k} \gamma_{kj}) \big] \Big]_{x_{(i-1)}}^{\infty} \prod_{l=i}^{N_1} \frac{\gamma_{(l)}}{\sum_{k=l}^{N_1} (\gamma_{(k)} + \sum_{j=1}^{n_k} \gamma_{kj})}. \tag{C.18}$$

We have used the same definitions as before; the subscripts in rounds brackets (i.e. $\gamma_{(i)}; i = 1, \ldots, N_1$) are related to the IETs, the one not in brackets to the CTs (i.e. $\gamma_{kj}; j = 1, \ldots, n_k; k = 1, \ldots, N_1$). When $i = 1$, the final solution will be

$$\prod_{l=1}^{N_1} \frac{\gamma_{(l)}}{\sum_{k=l}^{N_1}(\gamma_{(k)} + \sum_{j=1}^{n_k} \gamma_{kj})}. \tag{C.19}$$

If we define $\Omega'(t_{(i)}) = \{[(j), j1, \ldots, jn_j], j = i, \ldots, N_1\}$ as the set of labels attached to the IETs and CTs with length $\geq t_{(i)}$, we can get the same solution reported in chapter 2 in equation 2.8

$$L = \prod_{i=1}^{N_1} \frac{\exp(\mathbf{z}_{(i)}\boldsymbol{\beta})}{\sum_{j \in \Omega'(t_{(i)})} \exp(\mathbf{z}_j\boldsymbol{\beta})}. \tag{C.20}$$

C.2 Inclusion of Ties in the Data Set

The same approach adopted to incorporate CTs can be used in case of ties. Once again N individuals are under test, individuals $i1, \ldots, id_i$ are observed to fail at $t_{(i)}, i = 1, \ldots, k$ where $t_{(i)} < \cdots < t_{(k)}$and $\sum d_i = N$ and k is the number of IETs in the data set. Like in the censoring data set, only partial information can be obtain from the rank vector, because the arrangement of the ranks of the d_i individuals failing at $t_{(i)}$ is otherwise unknown. Therefore, the probability that the rank vector should be one of those possible given the sample is the sum of $\prod_i^k d_i!$ terms like equation C.14. The calculation can be considerably simplified by noting that the ranks assigned to the d_i individuals who fail at $t_{(i)}$ is unaffected by the ranks assigned to the d_j individuals who fail at $t_{(j)}$. The sum then reduces the product of k sums, one for each failure time. Let Q_i be the set of permutations of the symbols $i1, \ldots, id_i$ and $\mathbf{P} = (p_1, \cdots, p_{d_{(i)}})$ be an element in Q_i. As before, $R(t_{(i)})$ is the risk set at $t_{(i)} - 0$ and let $R(t_{(i)}, p_r)$ be the set difference $R(t_{(i)}) - \{p_1, \cdots, p_{r-1}\}$. The marginal likelihood for $\boldsymbol{\beta}$ is

$$\prod_{i=1}^{k} \left\{ \exp(\mathbf{s}_i\boldsymbol{\beta}) \sum_{\mathbf{P} \in Q_i} \prod_{r=1}^{d_i} \left[\sum_{l \in R(t_{(i)}, p_r)} \exp(\mathbf{z}_l\boldsymbol{\beta}) \right]^{-1} \right\} \tag{C.21}$$

where $\mathbf{s}_i = \sum \mathbf{z}_{ij}$ is the sum of covariates of individuals observed to fail at $t_{(i)}$. Equations C.14 and C.20 are special cases of equation C.21. However, this formula is very complicated and under mild condition it is possible using an approximation, which is much simpler. In fact, if the number d_i of individuals failing at each failure points $t_{(i)}$ is small compared to the number of items in the risk set, the likelihood function can be well approximated by

$$L = \prod_{i=1}^{k} \frac{\exp(\mathbf{s}_i \boldsymbol{\beta})}{[\sum_{l \in R(t_{(i)})} \exp(\mathbf{z}_l \boldsymbol{\beta})]^{d_i}}. \tag{C.22}$$

For the purpose of this work this approximation is quite satisfactory because the number of ties for each IET is small. Several works and papers have been devoted to the study of the survival data analysis and the approximation C.22 is well discussed (see for example Cox, 1972; Peto, 1972, and Breslow, 1974, who proposed a maximum likelihood technique different from the one used in this work, but leading to the same results), therefore we are confident that this approximation will not influence our results.

Appendix D

Properties of the Score Function and the Information Matrix

Let us suppose that a set of data consists of N independent observations $t_1 \dots t_N$ of random variables T with probability density function $f(T;\theta)$ (hereinafter pdf) that involves a parameter θ. Let us suppose that the Maximum Likelihood Estimation $\hat{\theta}$ exists and it is a point of relative maximum of the log likelihood function. It is possible to define two functions: the score function $S(\theta)$ and the information matrix $I(\theta)$:

$$\begin{aligned} S(\theta) &= \frac{\partial}{\partial\theta}[\ln L(\theta)] \\ I(\theta) &= -\frac{\partial^2}{\partial\theta^2}[\ln L(\theta)] \end{aligned} \tag{D.1}$$

where $L(\theta)$ is the likelihood function $L(\theta) = \prod_{i=1}^{N} L_i(\theta)$, with $L_i(\theta)$ being proportional to the pdf of t_i. The score function is then the first derivative of the log likelihood function and the information matrix is defined as minus its second derivative.

The score function and the information matrix have important properties. We prove that, under mild condition of regularity, $S(\theta)$ has expected value 0 and variance equal to the information matrix, that is:

$$\begin{aligned} &E\{S(\theta)\} = 0 \\ &var(S(\theta)) = I(\theta). \end{aligned} \tag{D.2}$$

Since variances are non-negative, it then follows that $I(\theta) \geq 0$.

For the proof, consider the $S(\theta)$ and $I(\theta)$ for a particular outcome t_i been equal to

$$\begin{aligned} S_i(\theta) &= \frac{\partial}{\partial\theta}[\ln L_i(\theta)] \\ I_i(\theta) &= -\frac{\partial^2}{\partial\theta^2}[\ln L_i(\theta)] \end{aligned} \tag{D.3}$$

for $i = 1, \cdots, N$ and where $L_i(\theta) = c \cdot f(t_i;\theta)$, with c a positive constant that does not depend on θ. For any value of θ the total probability will be

$$\sum_{i=1}^{N} f(t_i;\theta) = 1 \tag{D.4}$$

as definition of pdf. By making its 1^{st} and 2^{nd} derivative with respect to θ, assuming that the order of differentiation and summation can be interchanged, one gets

$$\sum_{i=1}^{N} \frac{\partial}{\partial\theta}[f(t_i;\theta)] = 0 \qquad \sum_{i=1}^{N} \frac{\partial^2}{\partial\theta^2}[f(t_i;\theta)] = 0. \tag{D.5}$$

For proving the property of $E\{S(\theta)\}$, consider that $S_i(\theta) = \frac{\partial}{\partial\theta}[\ln f(t_i;\theta)]$ $= \frac{1}{f(t_i;\theta)}\frac{\partial}{\partial\theta}[f(t_i;\theta)]$ and therefore

$$\sum_{i=1}^{N} S_i(\theta) f(t_i;\theta) = \sum_{i=1}^{N} \frac{\partial}{\partial\theta}[f(t_i;\theta)] = 0. \tag{D.6}$$

This proves that the expected value of $S_i(\theta)$ is 0, i.e., $E\{S_i(\theta)\} = 0$.

Let's now consider that the information matrix can be written as:

$$\begin{aligned} I_i(\theta) = -\frac{\partial}{\partial\theta}[S_i(\theta)] = -\frac{\partial}{\partial\theta}\left(\frac{1}{f(t_i;\theta)}\frac{\partial}{\partial\theta}[f(t_i;\theta)]\right) = \\ = \frac{1}{f(t_i;\theta)^2}\left(\frac{\partial}{\partial\theta}[f(t_i;\theta)]\right)^2 - \frac{1}{f(t_i;\theta)}\frac{\partial^2}{\partial\theta^2}[f(t_i;\theta)], \end{aligned} \tag{D.7}$$

from which one obtains

$$\frac{\partial^2}{\partial\theta^2}[f(t_i;\theta)] = S_i(\theta)^2 f(t_i;\theta) - I_i(\theta) f(t_i;\theta) \tag{D.8}$$

and

$$\sum_{i=1}^{N} \frac{\partial^2}{\partial\theta^2}[f(t_i;\theta)] = \sum_{i=1}^{N} S_i(\theta)^2 f(t_i;\theta) - \sum_{i=1}^{N} I_i(\theta) f(t_i;\theta) = 0. \tag{D.9}$$

The first sum in right part is $E\{S_i(\theta)^2\}$ and the second one is the expectation of the information matrix $\mathcal{I}_i(\theta) = E\{I_i(\theta)\}$ that is

$$E\{S_i(\theta)^2\} = E\{I_i(\theta)\}. \tag{D.10}$$

In general the variance of a variable x, with mean μ, can be written as

$$var(x) = E\{(x-\mu)^2\} = E\{x^2\} - E\{x\}^2, \tag{D.11}$$

in our case $E\{S_i(\theta)\} = 0$, therefore

$$var(S_i(\theta)) = E\{S_i(\theta)^2\} = E\{I_i(\theta)\} \tag{D.12}$$

$E\{I_i(\theta)\}$ is also called the Fisher information. Since the $t_1, \ldots, t_N$ are independent, $S_1(\theta), \ldots, S_N(\theta)$ are also independent. Then, under mild condition (Kalbfleisch & Prentice, 1980), a central limit theorem applies to the total score statistic $S(\theta) = \sum_{i=1}^{N} S_i(\theta)$. As a consequence, $S(\theta)$ has mean 0 and covariance matrix $\mathcal{I}(\theta) = \sum_{i}^{N} \mathcal{I}_i(\theta)$. In general, in the result of equation D.12 the Fisher information can be replaced with the observed information $I(\theta)$ or by $I(\hat{\theta})$ without affecting the asymptotic distribution (Kalbfleisch, 1980; Kalbfleisch & Prentice, 1980).

The property of the score statistics can be used for inference about the distribution of θ. It is possible to demonstrate that variance of θ is $var(\theta) = I(\hat{\theta})^{-1}$. In fact, the Taylor's series expansion at $\theta = \hat{\theta}$ of the score function $S(\theta)$ is

$$S(\theta) = S(\hat{\theta}) + \frac{\partial S(\theta)}{\partial \theta}|_{\theta=\hat{\theta}}(\theta - \hat{\theta}) \tag{D.13}$$

but

$$\begin{aligned} \frac{\partial S(\theta)}{\partial \theta}|_{\theta=\hat{\theta}} &= -I(\hat{\theta}) \\ S(\hat{\theta}) &= 0 \end{aligned} \tag{D.14}$$

then

$$S(\theta) = -I(\hat{\theta})(\theta - \hat{\theta}) \tag{D.15}$$

In general, for a variable x, $var(ax) = a^2 var(x)$ therefore the variance of $S(\theta)$ is

$$var(S(\theta)) = I(\hat{\theta})^2 var(\theta). \tag{D.16}$$

Using the property of the $var(S(\theta))$, the variance of θ can be written as

$$var(\theta) = I(\hat{\theta})^{-1}. \tag{D.17}$$

The MLE of $\hat{\theta}$ describe the location of the likelihood function, while $I(\hat{\theta})$ measures its spread, and it is therefore a measure of the informativeness or precision of the experiment with respect to θ.

As a final step, a rapid demonstration of equations 2.10 and 2.11 in chapter 2 is supplied. Starting with the likelihood of equation 2.9 in chapter 2 that allows censoring and ties into the data set, the explicit form of the score vector is

$$\mathbf{U}_j(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_j} \ln(L) =$$

$$\sum_{i=1}^{N_1} \left[s_{ji} - d_i \frac{\sum_{l \in \Omega'(t_{(i)})} z_{jl} \exp(\mathbf{z}_l \boldsymbol{\beta})}{\sum_{l \in \Omega'(t_{(i)})} \exp(\mathbf{z}_l \boldsymbol{\beta})} \right] = 0 \quad (j = 1, \cdots, s) \tag{D.18}$$

where s is the dimension of $\boldsymbol{\beta}$, $\mathbf{s}_i = \sum_{j=1}^{d_i} \mathbf{z}_{ij}$ is the sum of the covariates of the d_i individuals observed to fail at $t_{(i)}$, and s_{ji} is the jth element in the vector $\mathbf{s}_i$. The information matrix is then

$$\mathbf{I}_{jk} = -\frac{\partial^2}{\partial \beta_j \partial \beta_k} \ln(L) \tag{D.19}$$

$$= \sum_{i=1}^{N_1} \left[d_i \frac{\sum_{l \in \Omega'(t_{(i)})} z_{kl} z_{jl} \exp(\mathbf{z}_l \boldsymbol{\beta})}{\sum_{l \in \Omega'(t_{(i)})} \exp(\mathbf{z}_l \boldsymbol{\beta})} - \frac{\sum_{l \in \Omega'(t_{(i)})} z_{kl} \exp(\mathbf{z}_l \boldsymbol{\beta}) \sum_{l \in \Omega'(t_{(i)})} z_{jl} \exp(\mathbf{z}_l \boldsymbol{\beta})}{\sum_{l \in \Omega'(t_{(i)})} \exp(\mathbf{z}_l \boldsymbol{\beta}) \sum_{l \in \Omega'(t_{(i)})} \exp(\mathbf{z}_l \boldsymbol{\beta})} \right].$$

Appendix E

Consider the natural logarithm of equation 2.17 in chapter 2:

$$\ln(L) = \sum_{i=1}^{N_1} \Big[\sum_{j \in P_i} \ln(1 - \alpha_i^{\exp(\mathbf{z}_j \boldsymbol{\beta})}) + \sum_{k \in \Omega'(t_{(i)}) - P_i} \exp(\mathbf{z}_k \boldsymbol{\beta}) \ln(\alpha_i) \Big] \quad \text{(E.1)}$$

its derivative is

$$\frac{\partial \ln(L)}{\partial \alpha_i} = \sum_{j \in P_i} \frac{-\exp(\mathbf{z}_j \boldsymbol{\beta}) \alpha_i^{(\exp(\mathbf{z}_j \boldsymbol{\beta}) - 1)}}{1 - \alpha_i^{(\exp(\mathbf{z}_j \boldsymbol{\beta}))}} + \sum_{k \in \Omega'(t_{(i)}) - P_i} \frac{\exp(\mathbf{z}_k \boldsymbol{\beta})}{\alpha_i}. \quad \text{(E.2)}$$

Adding and subtracting the same quantity $\sum_{j \in P_i} \frac{\exp(\mathbf{z}_j \boldsymbol{\beta})}{\alpha_i}$, and rearranging terms, we get

$$\sum_{j \in P_i} \Big[\frac{-\exp(\mathbf{z}_j \boldsymbol{\beta}) \alpha_i^{(\exp(\mathbf{z}_j \boldsymbol{\beta}) - 1)}}{1 - \alpha_i^{\exp(\mathbf{z}_j \boldsymbol{\beta})}} - \frac{\exp(\mathbf{z}_j \boldsymbol{\beta}) \alpha_i^{(\exp(\mathbf{z}_j \boldsymbol{\beta}) - 1)}}{\alpha_i^{\exp(\mathbf{z}_j \boldsymbol{\beta})}} \Big] + \sum_{k \in \Omega'(t_{(i)})} \frac{\exp(\mathbf{z}_k \boldsymbol{\beta})}{\alpha_i} = 0 \quad \text{(E.3)}$$

that easily becomes equal to:

$$\sum_{j \in P_i} \frac{\exp(\mathbf{z}_j \boldsymbol{\beta})}{1 - \alpha_i^{\exp(\mathbf{z}_j \boldsymbol{\beta})}} = \sum_{k \in \Omega'_{(t_{(i)})}} \exp(\mathbf{z}_k \boldsymbol{\beta}). \quad \text{(E.4)}$$

To solve equation E.4 for $\alpha_i = \hat{\alpha}_i$ an iteration solution is required. Using the approximation $\hat{\alpha}_i^{\exp(\mathbf{z}_j \boldsymbol{\beta})} \simeq 1 + \exp(\mathbf{z}_j \boldsymbol{\beta}) \ln(\hat{\alpha}_i)$ into equation E.4, we get a suitable initial value equal to:

$$\sum_{j \in P_i} \frac{\exp(\mathbf{z}_j \boldsymbol{\beta})}{1 - 1 - \exp(\mathbf{z}_j \boldsymbol{\beta}) \ln(\hat{\alpha}_i)} = \sum_{k \in \Omega'(t_{(i)})} \exp(\mathbf{z}_k \boldsymbol{\beta})$$

$$\sum_{j \in P_i} \frac{1}{\ln(\hat{\alpha}_i)} = \sum_{k \in \Omega'(t_{(i)})} \exp(\mathbf{z}_k \boldsymbol{\beta}) \quad \text{(E.5)}$$

therefore the initial value α_{i0} for such iteration is then:

$$\ln(\alpha_{i0}) = \frac{-d_i}{\sum_{k \in \Omega'(t_{(i)})} \exp(\mathbf{z}_k \boldsymbol{\beta})}. \tag{E.6}$$

Appendix F

Residual Analysis

The residual analysis is a well known technique used by statisticians and it is useful for rescaling point processes and dealing them with a Poisson process with intensity 1. The main theorem asserts that a continuous random transformation of a change point onto itself transforms the non-negative points of the given process into a Poisson process with rate 1. A proof and a detailed explanation can be found in Papangelou (1972). The survivor function of the Poisson process for a variable x with rate $\lambda = 1$ is:

$$S(x) = \exp(-x) \tag{F.1}$$

while the survivor function in term of hazard function is

$$S(t) = \exp - \int_0^t \lambda(u)du \tag{F.2}$$

(see equation A.5 in Appendix A). The basic idea is that if the "cumulative hazard" $\int_0^t \lambda(u)du$ is a random variable, it can be only for a Poisson process with rate 1.

In our case, we transform the IET into residuals through the equation:

$$\hat{e}_i = \hat{\Lambda}_0(t_i) \exp(\mathbf{z}_i \hat{\boldsymbol{\beta}}) \tag{F.3}$$

where

$$\hat{\Lambda}_0(t) = \int_0^t \lambda_0(u)du \tag{F.4}$$

which is a monotonically increasing function because $\lambda_0(t)$ is nonnegative (see Ogata 1988). Since we adopt the nonparametric maximum likelihood

approach, $\hat{\Lambda}_0(t)$ can be written as:

$$\hat{\Lambda}_0(t) = \sum_{i|t_{(i)}<t} \Big(\sum_{k \in \Omega'(t_{(i)})} \exp(\mathbf{z}_k \hat{\boldsymbol{\beta}}) \Big)^{-1}. \tag{F.5}$$

In fact, the estimator $\hat{S}_0(\cdot)$ is equal to $\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t))$, the "cumulative hazard" F.5 arises directly from the equation of the nonparametric maximum likelihood $\hat{S}_0(t) = \prod_{i|t_i<t} \hat{\alpha}_i$ (equation 2.19 in chapter 2) following the first order expansions of $\ln \hat{\alpha}_i$. Therefore, $\{t_i\}$ is transformed one-to-one into $\{e_i\}$. Roughly speaking if the model is a good approximation of reality, then the transformed data $\{e_i\}$ are expected to have a Poissonian behaviour. Thus, a deviation from a property of $\{e_i\}$ from the expected Poissonian process implies the existence of a corresponding feature in the data $\{t_i\}$ that is not captured by PHM.

Appendix G

The Kolmogorov-Smirnov test

G.1 The one-sample Kolmogorov-Smirnov test

The one-sample Kolmogorov-Smirnov test is used to decide if a sample comes from a population with a specific continuous distribution.
A random sample $t_1, t_2, \ldots, t_n$ is drawn from a population T with unknown cumulative distribution function $F_T(t)$. For any value of t, the empirical cumulative distribution of the sample is $S_n(t)$, that is a step function that increases by $\frac{1}{n}$ at the value of everyone ordered data point, and it represents the statistical image of the true statistic $F_T(t)$.
The null hypothesis to be tested is

$$H_0: \quad F_T(t) = F_0(t) \qquad \text{for all } t. \tag{G.1}$$

where $F_0(t)$ is a completely specified continuous cumulative function.
Since $S_n(t)$ is the statistical image of $F_T(t)$, the null hypothesis is true if the difference between $S_n(T)$ and $F_0(t)$ is small for all t.

The statistic $KS1$ used in the one-sample Kolmogorov-Smirnov test is defined as (e.g.,Gibbons, 1971):

$$KS1 = \max_{-\infty<t<\infty} \big|F_0(t) - S_n(t)\big|. \tag{G.2}$$

The critical value of the exact distribution of $KS1$ under hypothesis H_0 has been tabulated for various values of n; see for example Siegel (1956, Table E, p.251) or Owen (1962, Table 15.1, pp.423–425).

G.2 The two-sample Kolmogorov-Smirnov test

The purpose of the two-sample Kolmogorov-Smirnov test is to decide if two samples come from the same population; it compares two empirical cumulative distribution functions.

Let us consider two samples of size m and n, from continuous populations T and X with population cumulative functions $F_T(t)$ and $G_X(t)$. The empirical cumulative distributions are $S_m(t)$ and $G_n(t)$, respectively. The null hypothesis is then

$$H_0: \quad F_T(t) = G_X(t) \qquad \text{for all } t, \tag{G.3}$$

and the two sides criterion ($KS2$) is the maximum absolute difference between the empirical distributions (e.g., Gibbons, 1971)

$$KS2 = \frac{mn}{d} \max_{-\infty<t<\infty} \left|S_m(t) - G_n(t)\right| \tag{G.4}$$

where d is the maximum common divisor of m and n.

The $KS2$ statistic is related to the significance level α at which the samples 1 and 2 have a different distribution function. The critical values $KS2\ (\alpha, m, n)$ for m, $n \geq 30$, rewritten as

$$KS2' = KS2 \frac{d}{[(mn)(m+n)]^{\frac{1}{2}}} \tag{G.5}$$

$$KS2' = \left(\frac{mn}{m+n}\right)^{\frac{1}{2}} \max_{-\infty<t<\infty} \left|S_m(t) - G_n(t)\right| \tag{G.6}$$

are well approximated by the distribution

$$P(KS2' < \lambda) = 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2} \quad ; \quad \lambda > 0 \tag{G.7}$$

Also critical values for small n, m can be found in the literature (see e.g. Hollander and Wolfe, 1973).

References

Akinci A., Mueller, A.C., Malagnini, L. & Lombardi A.M., 2004. Seismic hazard estimate in the Alps and Apennines (Italy) using smoothed historical seismicity and regionalized predictive ground-motion relationships. *Boll. Geof. Teor. Appl.*, **45**, 285-304.

Albarello D., Bosi, D., Bramerini, F., Lucantoni, A., Naso, G., Peruzza, L., Rebez, F., Sabetta, F. & Slejko, D., 1999. New seismic hazard maps of the Italian territory. *Serv. Sismico Naz.*, Rome. (Available at http://www.serviziosismico.it/PROG/2000/carte_pericolosita/)

Amato A. & Cimini, G.B., 2001. Deep structures from seismic tomography, in *Anatomy of an orogen: the Apennines and adjacent Mediterranean basins*, edited by Vai, G.B. & Martini, I.P., pp. 33-46, Kluwer Acad., Norwell, Mass.

Anderson H. & Jackson, J., 1987. Active tectonics of the Adriatic region. *Geophys. J. R. Astron. Soc.*, **91**, 937-987.

Barnard G.A., 1963. Some aspects of the fiducial argument. *J. R. Stat. Soc. B.*, **25**, 111-114

Bigi G., *et al.*, 1991. Synthetic structural-kinematic map of Italy, scale 1:2.000.000. *CNR-PFG Quad. Ric. Sci.*, **114**, 3, Consiglio Naz. delle Ric., Rome.

Boschi E., Gasperini, P. & Mulargia, F., 1995. Forecasting where larger crustal earthquakes are likely to occur in Italy in the near future. *Bull.*

Seismol. Soc. Am., **85**, 1475-1482.

Breslow N.E., 1974. Covariance analysis of censored survival data. *Biometrics*, **30**, 89-99.

CMT Catalog, available at http://www.seismology.harvard.edu/projects/CMT/.

Console R. & Murru, M., 2001. A simple and testable model for earthquake clustering. *J. Geophys. Res.*, **106**, 8699-8711.

Console R., Murru M. & Lombardi, A.M., 2003. Refining earthquake clustering models. *J. Geophys. Res.*, **108(B10)**, 2468, doi:10.1029/2002JB002130.

Cornell C.A., 1968. Engineering seismic risk analysis. *Bull. Seismol. Soc. Am.*, **58**, 1503-1606.

Cox D.R., 1972. Regression models and life tables (with discussion). *J. R. Stat. Soc. B*, **34**, 187-220.

CPTI Working Group, 1999. Catalogo parametrico dei terremoti italiani: GNDT-ING-SGA-SSN, pp. 88, Ist. Naz. di Geofis. e Vulcanol., Bologna, July. (Also available at http://emidius.mi.ingv.it/CPTI)

CPTI Working Group, 2004. Catalogo parametrico dei terremoti italiani, version 2004 (CPTI04), Ist. Naz. di Geofis. e Vulcanol., Bologna. (Available at http://emidius.mi.ingv.it/CPTI)

Davis P.M., Jackson, D.D. & Kagan, Y.Y., 1989. The longer it has been since the last earthquake, the longer the expected time till the next? *Bull. Seismol. Soc. Am.*, **79**, 1439-1456.

Dietrich J., 1994. A constitutive law for rate of earthquake production and its application to earthquake clustering. *J. Geophys. Res.*, **99**, 2001-2618.

Dziewonski A.M. & Woodhouse, J.H., 1983. An experiment in systematic study of global seismicity: Centroid-moment tensor solutions for 201 moderate and large earthquakes of 1981. *J. Geophys. Res.*, **88**, 3247-3271.

Dziewonski A.M., Ekström, G., Franzen, J.E. & Woodhous, J.H., 1987. Centoid-moment tensor solutions for January-March 1986. *Phys. Earth Plan. Int.* **45**, 1-10.

Ellsworth W.L., Matthews, M.V., Nadeau, R.M., Nishenko, S.P., Reasenberg, P.A. & Simpson, R.W., 1998. A physically-based earthquake recurrence model for estimation of long-term earthquake probabilities. *Proceedings of the Second Joint Meeting of the UJNR panel on Earthquake research*, 135-149. Geographical Survey Institute.

Freed A.M. & Lin, J., 2001. Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress tranfer. *Nature*, **411**, 180-183.

Galadini F., Meletti, C. & Vittori, E., 2001. Major active faults in Italy: Available superficial data. *Netherlands J. Geosc./ Geologie en Mijnbouw*, **80 (3-4)**, 273-296.

Galli P. & Bosi, V., 2002. Paleoseismology along the Cittanova Fault; implications for seismotectonics and earthquake recurrence in Calabria (southern Italy). *J. Geophys. Res.*, **107(B3)**, 10.1029/2001JB000234.

Galli P. & Bosi, V., 2003. Catastrophic 1638 earthquakes in Calabria (southern Italy): New insights from paleoseismological investigation. *J. Geophys. Res.*, **108(B1)**, 10.1029/2001JB001713.

Gasperini P., & Mulargia, F., 1989. A statistical analysis of seismicity in Italy: the clustering properties. *Bull. Seismol. Soc. Am.*, **79**, 973-988.

Geller R.J., Jackson, D.D., Kagan, Y.Y. & Mulargia, F., 1997. Earthquakes cannot be predicted. *Science*, **275**, 1616-1619.

Gibbons J.D., 1971. *Non–parametric Statistical Inference*. McGraw–Hill, New York, 306 pp.

Helmstetter A. & Sornette, D., 2003. Foreshocks and cascades of triggered seismicity. Submitted to *J. Geophys. Res.*.

Hollamder M. & Wolfe, D.A., 1973 Nonparametric statistical methods. *J. Wiley & Sons*, New York.

Hunstad I., Selvaggi, G., D'Agostino, N., England, P., Clarke, P. & Pierozzi, M., 2003. Geodetic strain in peninsular Italy between 1875 and 2001. *Geophys. Res. Lett.*, **30**, 4, 1181, doi: 10.1029/2002GL016447.

Jackson D.D. & Kagan, Y.Y., 1993. Reply. *J. Geophys. Res.*, **98**, 9917-9920.

Kagan Y.Y., 1991. Likelihood analysis of earthquake catalogues. *Geophys. J. Int.*, **106**, 135-148.

Kagan Y.Y. & Jackson, D.D., 1991. Long-term earthquake clustering. *Geophys. J. Int.*, **104**, 117-133.

Kagan Y.Y. & Jackson, D.D., 1994. Long-term probabilistic forecasting of earthquakes. *J. Geophys. Res.*, **99**, 13685-13700.

Kagan Y.Y. & Jackson, D.D., 1999. Worldwide doublets of large shallow earthquakes. *Bull. Seismol. Soc. Am.*, **85**, 1147-1155.

Kagan Y.Y. & D.D. Jackson, 2000. Probabilistic forecasting of earthquakes. Geophys. J. Int., **143**, 438-453.

Kalbfleisch J.D., 1985. *Probability and Statistical Inference*, 2nd edn, Springer-Verlag, New York.

Kalbfleisch J.D. & Prentice, R.L., 1980. *The statistical analysis of failure time data*, John Wiley & Sons, New York.

Kaplan E.L. & Meier, P., 1958. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457-481.

Kostrov B.V., 1974. Seismic moment and energy of earthquakes and seismic flow of rock. *Izv. Acad. Sci. USSR Phys. Solid Earth*, **135**, 23-40.

Lucente F.P., Chiarabba, C., Cimini, G.B. & Giardini, D., 1999. Tomographic constraints on the geodynamic evolution of the Italian region. *J. Geophys. Res.*, **104**, 20307-20327.

Malinverno A. & Ryan, W.B.F., 1986. Extension in the Tyrrhenian Sea and shortening in the Apennines as a result of arc migration driven by sinking of the lithosphere. *Tectonics*, **5**, 227-245.

Marzocchi W., Sandri, L. & Boschi, E., 2003A. On the validation of earthquake-forecasting models: the case of pattern recognition algorithms. *Bull. Seismol. Soc. Am.*, **93**, 1994-2004.

Marzocchi W., Selva, J., Piersanti, A. & Boschi, E., 2003B. On the long time interaction among earthquakes: some insight from a model simulation. *J. Geophys. Res.*, **108**(B11), 2538, doi:10.1029/2003JB002390.

Matthews M.V., Ellsworth, W.L., Reasenberg, P.A., 2002. A Brownian Model for Recurrent Earthquakes. *Bull, Seismol Soc. Am.*, **92**, 2233-2250.

McCann W.R., Nishenko, S.P., Sykes, L.R. & Krause, J., 1979. Seismic gaps and plate tectonics: seismic potential for major boundaries. *Pageoph*, **117**, 1082-1147.

Meletti C., Patacca, E. & Scandone, P., 2000. Construction of a seismotectonic model: the case of Italy. *Pure Appl. Geophys.*, **157**, 11-35.

Michael A.J. & Jones, L.M., 1998. Seismicity alert probabilities at Park-

field, California, revisited. *Bull. Seismol. Soc. Am.*, **88**, 117-130.

Montone P., Amato, A. & Pondrelli, S., 1999. Active stress map of Italy. *J. Geophys. Res.*, **104**, 25595-25610.

Montone P., Mariucci, M.T., Pondrelli, S.,& Amato, A., 2004. An improved stress map for Italy and surrounding regions (Central Mediterranean). *J. Geophys. Res.*, **109**, B10410, doi:10.1024/2003JB002703.

Mulargia F. & Gasperini P., 1995 Evaluation of the applicability of the time- and slip- predictable earthquake recurrence models to Italian seismicity. *Geophys. J. Int.*, **120**, 453-573.

Mulargia F. & Tinti, S., 1985 Seismic sample areas define from incomplete catalogues: an application to the Italian territory. *Phys. Earth planet. Inter.*,**40**, 273-300.

Nishenko S.P., 1985. Seismic potential for large and great interplate earthquakes along the Chilean and Southern Peruvian margins of South America: a quantitative reappraisal. *J. Geophys. Res.*, **90**, 3589-3615.

Nishenko S.P. & Buland, R.A., 1987. A generic recurrence interval distribution for earthquake forecasting. *Bull. Seismol. Soc. Am.*, **77**, 1382-1399.

Ogata Y., 1981. On Lewis' Simulation Method for Point Processes, *IEEE Trans. Inform. Theory*, IT-30,23-31.

Ogata Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. Ass.*, **83**, 9-27.

Ogata Y., 1998. Space-time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.*, **50**, 379-402.

Owen, D.B., 1962. Handbook of Statistical Tables. *Addison-Wesley Publishing Company*, Inc., Reading, Mass.

Pace B., Peruzza, L., Lavecchia, G., & Boncio, P., 2002. Seismogenic sources in Central Italy: from causes to effects. *Mem. Soc. Geol. It.*, **57**, 419-429.

Pacheco J.F. & Sykes, L.R., 1992. Seismic moment catalogue of large shallow earthquakes, 1900-1989. *Bull. Seismol. Soc. Am.* **82**, 1306-1349.

Pantosti D., Schwartz, D.P. & Valensise, G., 1993. Paleoseismology along the 1980 surface rupture of the Irpinia fault; implications for earthquake recurrence in Southern Apennines, Italy. *J. Geophys. Res.*, **98**, 6561-6577.

Papangelou F., 1972. Integrability of the Expected increments of point processes and a related random change of scale. *Transactions of the American Mathematical Society*, **165**, 483-506.

Papazachos B.C., 1989. A time predicable model for the earthquake generation in Greece. *Bull. Seismol. Soc. Am.*, **79**, 77-84.

Papazachos B.C., 1992. A time and magnitude predictable model for generation of shallow earthquakes in the Aegean area. *Pure Appl. Geophys.*, **138**, 287-308.

Papazachos B.C. & Papadimitriou, E.E., 1997. Evaluation of the global applicability of the regional time- and magnitude-predictable seismicity model. *Bull. Seismol. Soc. Am.*, **87**, 799-808.

Papazachos B.C. & Papaioannou, Ch. A., 1993. Long-term earthquake prediction in the Aegean area based on a time and magnitude predictable model. *Pageoph*, **140**, 595-615.

Patacca E. & Scandone, P., 1989. Post-Tortonian mountain building in the Apennines. The role of passive sinking of a relic lithospheric slab, in *The Lithosphere in Italy: Advances in Earth Science Research*, edited by

A. Boriani et al., pp. 157-176, Accad. Naz. dei Lincei, Rome, Italy.

Peto R., 1972. Contribution to the discussion of paper by D.R. Cox. *J. R. Stat. Soc. B*, **34**, 205-207.

Piersanti A., Spada, G. & Sabadini, R., 1997. Global post-seismic rebound of a viscoelastic Earth: Theory for the finite faults and application to the Alpine-Mediterranean area. *J. Geophys. Res.*, **102**, 477-492.

Piromallo C. & Morelli, A., 2003. P-wave tomography of the mantle under the Alpine-Mediterranean area. *J. Geophys. Res*, **108**(B2), 2065, doi:10.1029/2002JB001757.

Polliz F.F., B*ü*rgmann, R. & Romanowicz, B., 1998. Viscosity of oceanic asthenosphere inferred from remote triggering of earthquakes. *Science*, **280**, 1245-1249.

Posadas A., Hirata, T. & Vidal F., 2002. Information theory to characterize spatiotemporal patterns of seismicity in the Kanto region. *Bull. Seismol. Soc. Am.*, **92**, 600-610.

Rong Y. & Jackson, D.D., 2002. Earthquake potential in and around China: estimated from past earthquakes. *Geophys. Res. Lett.*, **29**, 10.1029/2002GL015297.

Scandone P. & Stucchi, M., 2000. La zonazione sismogenetica ZS4 come strumento per la valutazione della pericolositâ sismica, in *Le ricerche del GNDT nel campo della pericolositâ sismica (1996-1999)*,edited by F. Galadini *et al.*, pp. 3-14, Gruppo Naz. per la Difesa dai Terremoti, Roma. (Available at http://emidius.mi.ingv.it/ GNDT / ZONE /zone_sismo.html)

Schwartz D.P. & Coppersmith, K.J., 1984. Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas fault zones. *J. Geophys. Res.*, **89**, 5681-5698.

Selvaggi G. & Amato, A., 1992. Subcrustal earthquakes in the northern Apennines (Italy): Evidence for a still active subduction? *Geophys. Res. Lett.*, **19**, 2127-2130.

Shimazaki K. & Nakata, T., 1980. Time-predictable recurrence model for large earthquakes. *Geophys. Res. Lett.*, **7**, 279-282.

Siegel S., 1956. Non-parametric Statistic for the Behavioral Science. *McGraw-Hill Book Company*, New York.

Sornette D. & Knopoff, L., 1997. The paradox of the expected time until the next earthquake. *Bull. Seismol. Soc. Am.*, **87**, 789-798.

Stock C. & Smith, E.G.C., 2002. Adaptative kernel estimation and continuous probability representation of historical earthquake catalogs. *Bull. Seismol. Soc. Am.*, **92**, 904-912.

Valensise G. & Pantosti, D., (Eds.), 2001. Database of Potential Sources for Earthquakes Larger than M 5.5 in Italy. *Ann Geofis*, **44**(4) Suppl., pp 797-964, with CD-ROM.

Vere-Jones D., 1970. Stochastic models for earthquake occurrence (with discussion). *J. R. Stat. Soc.*, **B32**, 1-62.

Westaway R., 1992. Seismic moment summation for historical earthquakes in Italy: tectonic implications. *J. Geophys.Res.*, **97**, 15437-15464.

Wyss M, 1997. Cannot earthquakes be predicted? *Science*, **278**, 487-488.

Working Group on California Earthquake Probability, 1999. Earthquake probabilities in the San Francisco Bay region: 2000-2030 - A summary of findings, Open-File Report 99-517, USGS.

Working Group CSTI, 2001. Catalogo strumentale dei terremoti "italiani" dal 1981 al 1996. Version 1.0, CDROM.

Zoback M.L., 1992. First- and second-order patterns of stress in the lithosphere: The World Stress Map Project. *J. Geophys. Res.*, **97**, 11703-11728.