Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Probabilistic Magnitude of Completeness of the Northern Californian Seismic Network

Diploma Thesis

Corinne Bachmann

March 2007

Advisors:
Prof. Dr. D. Schorlemmer
Prof. Dr. E. Kissling

Eidgenössiche Technische Hochschule Zürich
Department of Earth Sciences
Institute of Geophysics
ETH Hönggerberg, Schafmattstr. 30, 8093 Zürich

Author:

Corinne Bachmann

Grossalbis 22

8045 Zürich

E-Mail: corinneb@student.ethz.ch

TEL: +41 79 755 82 69

Supervisors

Prof. Dr. Danijel Schorlemmer

University of Southern California

Los Angeles, CA 90089

E-Mail: ds@usc.edu

TEL: +1 213 740 41 41

Prof. Dr. Eduard Kissling

Institute of Geophysics

ETH Zürich

Schaffmatstrasse 30

8093 Zürich

E-Mail: kissling@tomo.ig.erdw.ethz.ch

TEL: +41 44 633 26 23

Abstract

The probabilistic approach of detecting completeness of seismic networks by Schorlemmer and Woessner is based on event data, station information, and attenuation relations. An intermediate product are distributions of recording probabilities in the magnitude/distance space for each station. All completeness estimates are derived from them.

Many sources of data flaws exist, e.g., magnitude errors, event clusters, etc., which propagate into artifacts in these distributions. Here we present methods for improving the distributions by adding physical constraints to the data. We detected event clusters which disturb the homogeneity of the data distributions and therefore strongly alter the recording capabilities. We introduce a method for parameterizing the distributions of recording capabilities for easier station quality evaluation.

We present results for the Northern Californian Seismic Network (NCSN), including station evaluation, completeness maps over time, and detection probability maps.

Contents

Chapter 1

Introduction

Earthquake catalogues are the basis of many seismological studies. Every catalogue reflects the abilities of the seismic network that recorded it. Each of these networks covers a specific area, which can be local, regional or even global. The networks consist of numerous stations distributed heterogeneously over this area. The ability of a network to determine the size and location of an earthquake accurately depends mainly on this distribution. If the station density is too sparse, there is the possibility that there are not enough stations near an earthquake, and it will thus not be possible to determine its parameters accurately. There will always be a threshold magnitude, beneath which the network will not be able to record a signal on sufficient stations to accurately determine all parameteres of an earthquake. This has on one hand simple financial reasons; there is normally a limited amount of money that can be spend on stations, so the station density can not be infinitely high. On the other hand, this threshold magnitude also depends on the site conditions of the single stations; a station installed on soft underground will less likely record a clear signal of an event, than a station installed on bedrock. The same holds for a station in an urban area compared to a station in a quiet area; the noise level at the first station will be much higher, so the signal of a seismic event has to exceed this level to be recognised as an earthquake. There are other reasons that an earthquake will not be recorded, e.g. the signal is masked by the stronger signal of a larger event, which is often case in aftershocks sequences.
The threshold magnitude, above which the earthquake catalogue is supposed to be complete is defined as the magnitude of completeness M_c. Below this, several events will be missed for the aforementioned reasons. This magnitude is

always a function of space and time, and any change in the network will change this magnitude, e.g. a new station will increase the density and therefore raise the ability to record a earthquake in its vicinity.

Importance of Completeness There are various statistical analyses of earthquake catalogues that require a complete dataset. Therefore one must know the magnitude of completeness M_c, to only consider events above this magnitude. One example is seismic hazard analysis; M_c is one of the crucial input parameters here. Giardini et al. [2004], for example, used two different estimates for the completeness in their input for their hazard determination, the final models differed by 10%; similar examples can be found in other hazard studies [De Crook, 1989; D'Amicio and Albarello, 2003; Shanker and Sharma, 1998]. Different authors studied the influence of the completeness on the determination of b-values [Utsu, 1965; Aki, 1965; Weichert, 1980; Bender, 1983] of the Gutenberg-Richter law [Gutenberg and Richter, 1944]; changes in the completeness will also change b-value estimates, because the magnitude of completeness is an important input to determine the maximum likelihood of b:

$$\frac{1}{b * \ln(10)} = \overline{m} - m_0$$

where $\overline{m}$ is the is the average magnitude of the sample, and m_0 is the lowest magnitude at which event observations are complete [Aki, 1965]; m_0 is basically identical with M_c. A change in the completeness will lead to a significant change in b.

We mentioned above that events can be masked by a strong signal of a large event, which will change the completeness of aftershock sequences, therefore M_c will always be analysed for afthershock studies [Wiemer and Katsumata, 1999; Woessner et al., 2004; Gerstenberger et al., 2005]. M_c will normally be higher in the early part of the aftershock sequence, since small events will be masked by the coda of larger aftershocks and because the workload of the affected network will be immense. Later on, M_c will either return to the level as before the mainshock, or it will decrease as large events often lead to improvements in the network.

There are different books dealing with the problem of analyses with missing data, for example by Dodge [1985] or Little and Rubin [1987]. They emphasise

the importance of a good knowledge of the completeness of the catalogue, because analyses of incomplete parts will always lead to wrong results. There is always a trade-off of between a large data set and the estimate of M_c; overestimating M_c will reduce the amount of data significantly, underestimating M_c makes results of any study less reliable.

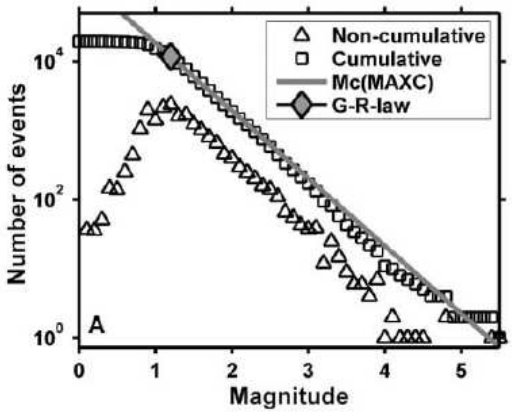


Figure 1.1: This figure shows how M_c is determined based on the deviation from the frequency magnitude distribution (FMD). The grey diamond shows where the cumulative number of events deviates from the FMD, this value is taken as M_c. (Figure taken from Woessner and Wiemer [2005]).

Approaches to determine the magnitude of completeness There are various approaches for the estimation of M_c. The most traditional and most commonly used approach defines the magnitude of completeness M_c as the deviation point from the Gutenberg-Richter line in the cumulative frequency-magnitude distribution (FMD) [Marsan, 2003; Wiemer and Wyss, 2000; Woessner and Wiemer, 2005; Cao and Gao, 2002]. Figure 1.1 illustrates this method; it shows the cumulative (squares) and the non-cumulative (triangles) number of events with a magnitude M versus the logarithmic number of events. The magnitude of completeness is taken as the point where the cumulative number deviates from the Gutenberg-Richter frequency magnitude distribution (grey diamond) [Woessner and Wiemer, 2005]. The method will be described in more detail later (cf. chapter 7).

Other approaches attempt to find the completeness of an earthquake catalogue by studying the noise-level at a station [Kværna et al., 2002a,b], or by com-

paring the day-to-night ratios [Rydelek and Sacks, 1989], these will also be discussed more extensively later (cf. chapter 7).

New approach The problem with the abovementioned methods is that they have to make too many assumptions. We try to overcome this problem by developing a new method for determination of magnitude of completeness which is only based on data recorded by a seismic network; this data will include the phase picks and the station properties. Every station in a network will have a different ability to record the signal of an event. As mentioned above, this ability depends strongly on the site conditions; a weak signal will be hard to identify on the record of a station which records high noise, for example. We therefore derive a probability distribution for each station reflecting the capability of this station to detect an earthquake at a specified magnitude-distance combination. We specify an earthquake sample for every magnitude-distance combination for a station and, from this data, compute the probability of detection as the ratio of the number of picked events to the number of all events within the sample. We will add the conditions that our probabilities will not decrease, if the magnitude increases at a constant distance. This reflects simple physical properties of the earthquake signal; an event with a higher magnitude at equal distance will generate a stronger signal than an event with a lower magnitude, making the probability higher or at least equal that this event will be recorded by a station. The same holds for an event within a shorter distance and an equal magnitude; a nearer earthquake will also produce a stronger signal. Therefore the probability will also not decrease for decreasing distances at constant magnitude. The distance here includes the station elevation and the depth of the earthquake.
With the probability distributions, we will be able to compute different maps. We can compute the probability that a given magnitude M can be detected at a specified point. Here we take into account how many stations are used for triggering in the studied seismic network; in our study area five stations are used. Thus we will compute the probability that M will be recorded at five or more stations. This will result in a map showing the probability that the magnitude M is detected by the studied seismic network. We can also compute maps showing the spatial distribution of the magnitude of completeness M_c at a specified probability level. For this, we use the same process as for the first

maps, but apply it iteratively, until we find the magnitude that will be detected at the chosen probability level. Both maps can be computed for different times; they will always depend on the station configuration on the specified date. By computing maps for different times, we can map M_c spatially and temporally.

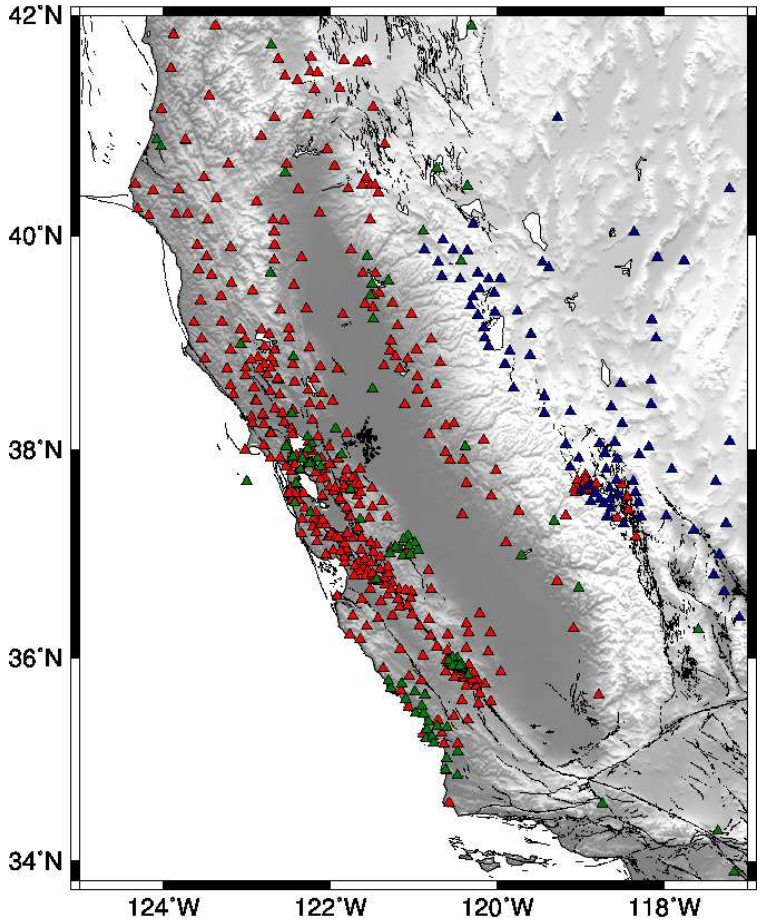


Figure 1.2: Map showing all stations used in this study. The blue triangles show the stations from the Northern Californian Seismic Network (NC), the red triangles show the UNR Broadband Network (NN) and the green triangles show the smaller networks (BP, BK, PG and WR). All stations have at least been partly active from 1^{st} January 2001 to 31^{st} December 2005.

Study Area Our study area is Northern California and we will analyse the Northern Californian Seismic Network (NCSN). This network consist of several subnetworks and maintains over 1,000 stations. However, not all of them are used in our approach. We only include stations used for triggering; these are

about 600 stations, depending on the date of our analysis. Figure 1.2 shows all stations that were at least partly active during the time of our analysis; different colours represent different networks (the Northern Californian Seismic Network NC is blue, the UNR broadband network NN is red and smaller networks are green). The NC network is maintained by the United States Geological Survey (USGS), the NN network by the University Nevada and the smaller networks are maintained by local organisations. However, all data from smaller networks are integrated in the Northern Californian Earthquake Data Centre (NCEDC, `http://www.ncedc.org`).

We investigate this region because it is a highly active and a well-studied region. There are many stations installed, especially along known faults, such as the San Andreas fault. The study region also includes geothermal and volcanically active fields. It also encloses a region that is essentially quiescent. Therefore we have a large diversity within this network, which makes it a good area for our study, as we want to show that the magnitude of completeness should not be averaged over a large area, but changes on a small spatial scale.

Will will study the data from five years, from 1^{st} January 2001 to 31^{st} December 2005. We will analyse the data of the catalogue provided by the NCEDC. We aim to find features for different stations and possibilities to improve the data, without excluding an excessive amount of data.

Chapter 2

Method

2.1 Analysis

Steps of the Analysis The analysis consists of five main parts (1) Defining a period, during which the data recording was homogeneous (2) Importing station data and creating a master station list (3) Importing event data with phase information (4) Assigning information about recorded (picked) and not recorded events to the stations. (5) Deriving probabilities of detection for each station and any given magnitude-distance combination.
(1) We will derive probabilities of detection for each station in the analysed network. The obtained probabilities can only be representative, if they have been derived during a period of homogeneous data recording. Particularly, this means that the triggering condition, the routine analysis condition and the magnitude definition must not change over the chosen time period. However, it is not possible to appoint every potential change within the routine processing workflow of a seismic network that may have a strong impact on the recording capability of the stations; slighter changes should not proceed into the probabilities of detection. By all means, one has to be careful at selecting this period; a longer period will lead to a larger amount of data but it will also contain more possible changes.
(2) For the chosen time period, we select all stations from the network having been in operation at least partially during the this period. It is important to have exact knowledge about on- and off- times of this stations. If it is not reported that a station is inactive, it will be interpreted as not recording, this will corrupt the probability of detection for this station. We also only import

station used to trigger the location procedure. In many networks, there are small subnetworks, for example temporal stations in boreholes, or local networks in high seismicity regions. These smaller networks are often not used for triggering, therefore they have to be excluded from the analysis.
(3) We will select all earthquakes from the catalogue, occurring during the chosen time period. From each event we have to select only phases, used for triggering, normally the P-wave picks on vertical components.
(4) The probabilities of detection are derived as the ratio of picked events to the total number of events. We select only events, occurring during the active time of a particular station, from this data we generate a triplet:

$$[\text{D M B}]$$

where D is the distance between the event and the station, M is the magnitude of the earthquake and B is a boolean number with the information if the station recorded the event or not. If $B = 1$, we call this plus-triplet, if $B = 0$ minus-triplet. These triplets repressent the raw data to compute the probability of detection for a station.
(5) To compute the probabilities of detection for different M/D combinations, $P_D(M, D)$, we sample the aforementioned data triplets. To define the bin, from which the triplets will be sampled, we translate the distances in magnitude units, defining a metric magnitude/distance space:

Defining a metric magnitude/distance space By comparing an event with the magnitude, M', and distance, D', we obtain two differences:

$$\begin{aligned} \Delta M &= |M - M'| \\ \Delta D &= |D - D'| \end{aligned}$$

Here the distance includes the depth of the event and the elevation of the station. Because magnitudes and distances are measured in magnitude units and meters, respectively, we translate distances into magnitude units. This translation uses the magnitude definition of the network. The Northern Californian network uses different magnitude definitions, depending on the size of the event. For events smaller than M3, they generally use the coda duration magnitude

equation, based on Eaton [1992]. This equation has the form:

$$\begin{aligned} MD(f-p) &= -0.81 + 2.22\log(f-p) + 0.0011 * D^* + Stacor + G + CC \\ &+ 0.006 * D^* \text{ if } D^* < 40 \text{ km} \\ &+ 0.006 * D^* \text{ if } D^* > 350 \text{ km} \\ &+ 0.014 * Z \text{ if } Z > 10 \text{ km} \end{aligned} \tag{2.1}$$

where $(f-p)$ is the end-of-coda (F) minus P-time, i.e. the duration, D^* is the epicentral distance of the station to the event, *Stacor* is the duration magnitude correction for the station, G is the gain correction, CC = Component Correction and Z is the (positive) depth of the event.
The coda duration magnitude has some restrictions, one of them is that it is almost independent of the distance to the earthquake. However, we want to translate the distance into magnitude units, therefore we encounter many problems using this magnitude definition. Thence, we chose to use the local magnitude definition, also by Eaton [1992]. The coda duration magnitude was derived, basing on this equation, therefore we will not make wrong assumptions. The local magnitude, M_L definition has the form

$$M_L = log(A_{WA}/(2 \times \text{CAL})) + F_1(s) + F_2(d) + \text{XCOR}_{\text{comp}} + \text{XCOR}_{\text{sta}} \tag{2.2}$$

where A_{WA} is the maximum peak-to-peak amplitude on the paper record for a Wood-Anderson seismometer, CAL is a dimensionless scaling factor assigned to the station and XCOR is a correction made for the component and the station, respectively. The two factors F_1 and F_2 are distance corrections [Eaton, 1992]:

$$\begin{aligned} F_1 &= 0.821 \times \log(S) + 0.00405 \times S + 0.955 \text{ for } S < 185.3 \text{ km} \\ F_1 &= 2.55 \times \log(S) \text{ for } S > 185.3 \text{ km} \\ F_2 &= -0.09 \times \sin(0.07 \times (D-25)) \text{ only if } D < 70 \text{ km} \end{aligned}$$

where S is the slant distance $S^2 = D^2 + Z^2$ and D is epicentral distance.
To determine the difference between the observations at the same station, we can make some assumptions: The amplitudes A_{WA1} and A_{WA2} will be the same for two similar events and the correction factors CAL and XCOR are the equal by definition for the same station. Therefore we can simplify the above

equation:

$$\begin{aligned} M_1 - M_2 &= F_1(s_1) + F_2(d_1) - F_1(s_2) - F_2(d_2) \text{ with} \\ F_1 &= 0.821 \times \log(S_1) + 0.00405 \times S_1 - 0.821 \times \log(S_2) - 0.00405 \times S_2 \\ &\text{for } S < 185.3 \text{ km} \\ F_1 &= 2.55 \times \log(S_1) - 2.55 \times \log(S_2) \\ &\text{for } S > 185.3 \text{ km} \\ F_2 &= -0.09 \times \sin(0.07 \times (D - 25)) \text{ only if } D < 70 \text{ km} \end{aligned}$$

We now translate the difference in distance ΔD into a magnitude difference $\Delta M^* = M_1 - M_2$. We define now our metric as the euclidian distance in this magnitude/magnitude space as:

$$D_M = \sqrt{\Delta M^2 + \Delta M^{*2}}$$

This magnitude definition describes only the attenuation for the NCSN, for any other network, there will be another definition. It is also essential that this definition is constant for the time period chosen above. For our case this is given, because this definition is in use since 1992 [Eaton, 1992].

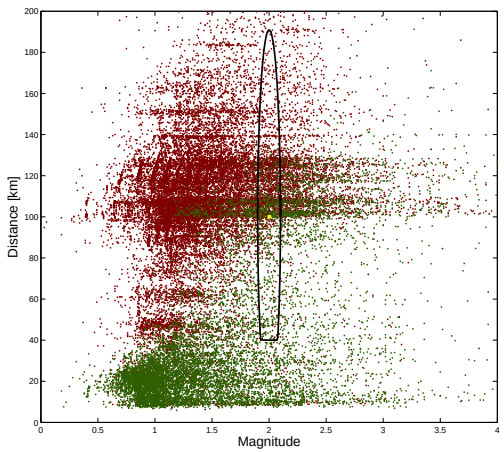


Figure 2.1: Magnitude against distance for events from 2001-2005. Green events have been recorded at the station BJO, red ones were missed. The ellipse outlines the events, being will be sampled to compute the probability at the point at $M'/D' = 2/100$, if the coda duration equation is used.

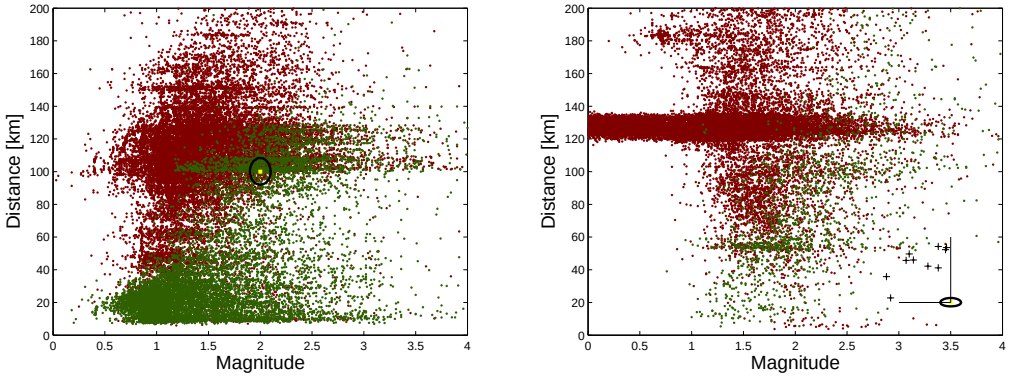
Derivation of probability of detection The probability of detection, $P_D(M, D)$, is derived from the aforementioned data triplets. We define $P_D(M', D')$ as:

$$P_D(M', D') = \frac{\text{number of picks}}{\text{total number of events}}$$

We are using a criterion to select the triplets used to determine the probability at detection at the point $M'/\ D'$. We select all events with:

$$D_M \leq 0.1$$

because 0.1 magnitude units can be considered as a usual magnitude error. This means that we normally will sample all events within an ellipse with the axes being ΔM and ΔM^*. We call N_p the number of triplets, obeing this criterion. If $N_p < 10$, we apply a second criterion to sample at least ten triplets.
From all triplets, not obeing the first criterion we select triplets with magnitudes $M' \leq M$ and distances $D' \geq D$. From this set of triplets, we select $10 - N_p$ triplets with the lowest D_M.
Figure 2.1 illustrates the independence of the coda duration magnitude equation to the distance; if we use this equation, we will sample all events within the displayed ellipse to compute $P_D(M', D')$ for $M' = 2$ and $D' = 100$ km. This basically means that using $D' = 180$ km leads to almost the same magnitude M' as using $D' = 50$ km, if we only consider the distance in equation 2.1. On the other hand, if we use equation 2.2 for our analysis, we obtain a completely different form of the ellipse. This is shown in figure 2.2 (a), this figure shows the same information as figure 2.1, but the ellipse outlines the events that will be sampled when the local magnitude equation 2.2 is used. All events within this ellipse have a $D_M \leq 0.1$ from the point with $M = 2'$ and $D' = 100$ km.
If the data is sparse, the second criteria will be applied, this is shown in figure 2.2 (b). We renounce to display an example for the coda duration magnitude for this case, as we already showed above that this equation can not be used for our approach. The ellipse in figure 2.2 outlines all events that would be sampled if the first criteria would be applied. It can be seen that there are no events within this ellipse, therefore the ten events with the smallest D_M and with $D \geq 20$ and $M \leq 3.5$ will be sampled; these are indicated with black crosses.
For both cases we sample $N \geq 10$ triplets, from these we will derive the prob-



(a) Station BJO, (M'/D')=(2/100)

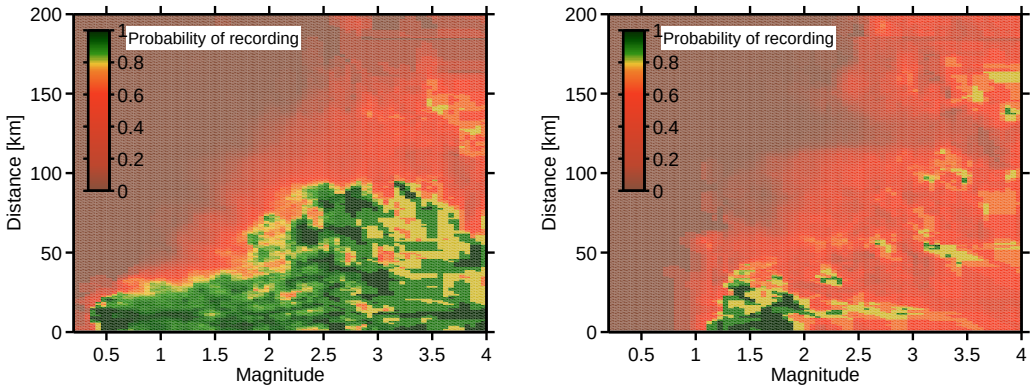
(b) Station KIP, (M'/D')=(3.5/20)

Figure 2.2: Magnitude against distance for events from 2001-2005, green events have been recorded red ones were missed at the given stations. The ellipse outlines the events meeting the first criteria, in (a) more than ten events lie within this ellipse. For (b) no events lie within the ellipse, therefore the second criteria will be applied. The black crosses indicate the sampled events, those are the events with the smallest D_M that have magnitudes $M' \leq M$ and distances $D' \geq D$ (indicated by the bars).

ability of detection as the number of plus-triplets divided by N. Therefore we estimate for any magnitude-distance combination the probability of detection, based only on the nearest data triplets, or in the second case, only triplets that represent weaker signals at the station.

Probability Matrices We are calculating the detection probabilities for each station for a distance range from 1 to 200 kilometres and a magnitude range for 0 to 4. Figure 2.3 shows two examples of matrices we obtain with this procedure. The left figure shows the station BJO, the raw data of this station was shown above in figure 2.2 (a). We can see that this station records many events and for many magnitude-distance combinations the probability of detection $P_D(M, D)$ is 0.8 or higher. On the right site of figure 2.3 we see the probability matrix of the station KIP. The raw data of this station (fig 2.2 (b)) shows that there were less events in the near vicinity of this station, this leads to less data and a lower probability of detection for many magnitude-distance combinations.

Adding physical constraints Until now, we only considered the raw data to calculate the probability matrices. We saw that the data is not everywhere

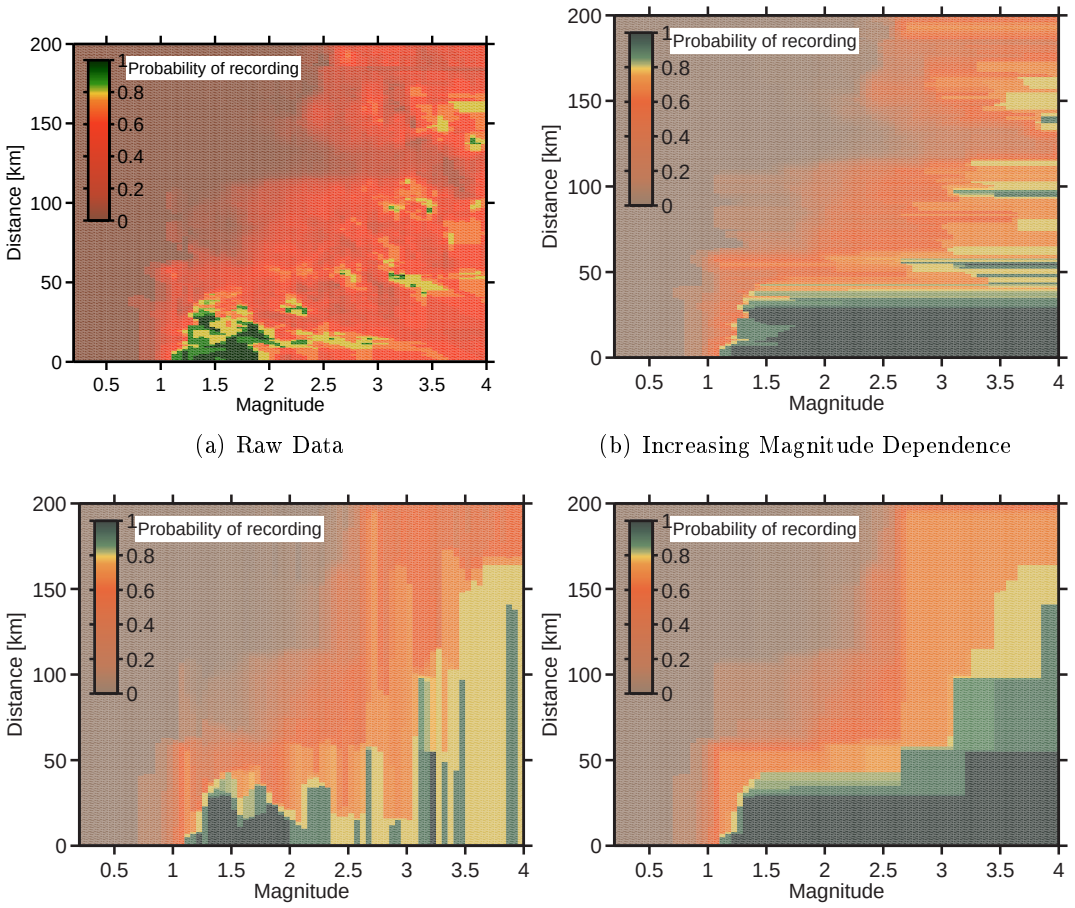


(a) Station BJO (b) Station KIP

Figure 2.3: Probability matrices for two arbitrary stations, the colour bar indicates the probability level.

dense enough so that we had to apply a second criteria to chose events that will be sampled for regions with too few data (see fig. 2.2 (b)). With this second criteria, we are able to determine a probability at these points, but figure 2.3 (b) shows that these probabilities are mostly lower than probabilities obtained where the data is denser. When the second criteria applies, we sample only ten events, therefore the effect of a missed event within the sample becomes much higher.

To overcome the effects of too few data, we add an assumption. We assume that the probability can not become smaller for the same distance but higher magnitudes. If a station is able to detect an earthquake with magnitude M1.5 at 20 kilometres distance, it will also detect events with magnitude M3 or M4 at this distance. The signal for an event with a higher magnitude will generally be stronger, therefore the station should record it, when it was also able to record the event with the weaker signal. It is possible that the station was not reported as being inactive during such an event, therefore the event will be registered as a missed event. Thence, it is possible that the probability decreases because the influence of such errors get too high. Figure 2.4 (a) and (b) show the comparison of the probability matrix based on the raw data only (a), and the matrix where we included our assumption (b). The probability increases at many places, especially for low distances. This can be explained if we look at the raw data of this station in figure 2.2 (b), there is only a few data for distances lower than 40 kilometers and magnitudes higher than 2.5. Therefore



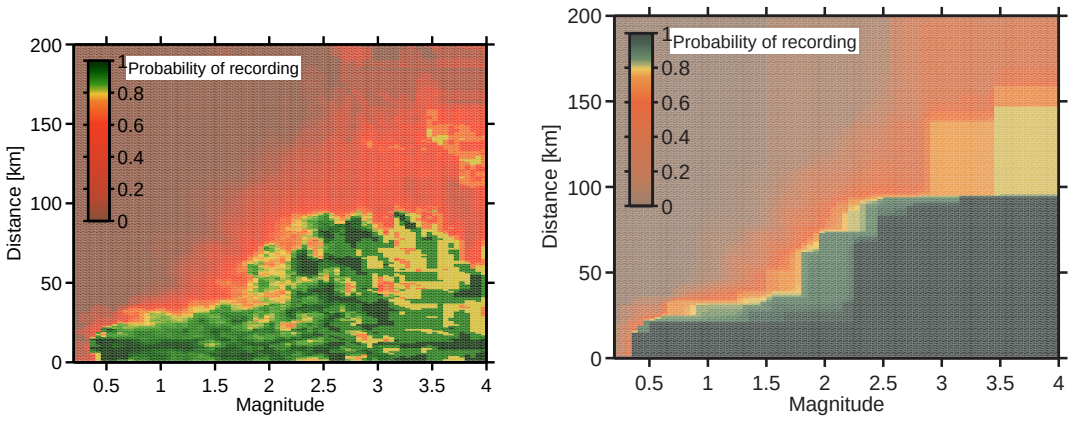
(a) Raw Data

(b) Increasing Magnitude Dependence

(c) Decreasing Distance Dependence

(d) Combination of (b) and (c)

Figure 2.4: Probability matrices for the station KIP, (a) is based on the raw data only, (b), (c) and (d) include assumption. In (b) the probabilities don't get smaller with increasing magnitudes for constant distances, in (c) the probabilities don't get smaller with constant magnitudes for decreasing distances. In (d), these both assumptions are combined.



(a) Raw Data (b) With assumptions

Figure 2.5: Probability matrices for the station BJO, (a) is based on the raw data only, (b) includes he assumptions that the probability does not decrease with constant magnitude and decreasing distance and constant distance and increasing magnitude.

the probabilities in this region will all be calculated with the second criteria, this can lead to the above mentioned effects.

In addition to the dependence on the magnitude, we also considered a dependence on the distance. The probabilities should not decrease if we have a constant magnitude and decreasing distance. This means that if a station is able to detect an event of magnitude M3.5 at 100 kilometres distance, it will not miss an event with the same magnitude at 20 kilometres distance. Figure 2.4 (c) shows this effect. Figure 2.4 (a) shows the probability matrix with based on the raw data, (c) shows the effect, if the probability does not decrease if the magnitude stays constant and the distance decreases.

The next step is now to combine both assumption, the result of this is shown in figure 2.4 (d). From the rather bad probability matrix of station KIP, we obtained now a probability matrix, where the probability is one for many magnitude-distance combinations. We applied this on all probability matrices, because even stations with a dense distribution of data had patches with lower probabilities. This is shown for the example of the station BJO in figure 2.5, we get the probability one for many magnitude-distance combinations, especially where we had regions with smaller probabilities before.

2.2 Synthesis

We are computing two different kinds of maps. (1) We compute the probability of detection of earthquakes of magnitude M at any location $\underline{x}$, resulting in probability map ($P_D(M, \underline{x})$ map). (2) We search the probabilities through the magnitude space to find the lowest magnitude with a probability of detection for a chosen probability level; this will results in a probabilistic completeness map (PMC map).
(1) To compute the probability of detection $P_D(M, \underline{x})$, of earthquakes of magnitude M at any location $\underline{x}$, we measure the distances to all stations in the network. We compute this probability of detection for one specified date, therefore we only consider stations having been active on this day

$$\begin{aligned} \underline{x} &= [\mathrm{Lon_x}/\mathrm{Lat_x}] \\ \mathrm{Station} &= [\mathrm{Lon_s}/\mathrm{Lat_s}/\mathrm{Elevation_s}] \\ D &= f(\mathrm{Lon_x}, \mathrm{Lat_x}, \mathrm{Lon_s}, \mathrm{Lat_s}, \mathrm{Elevation_s}) \end{aligned}$$

where Lon and Lat stand for longitude and latitude, respectively. The elevation of x is taken as constant. We then compute the probabilities of detection, $P_D(M, D)$ at every station for the given distances D to $\underline{x}$ and the corresponding probabilities of non-detection $P_N(M, D) = 1 - P_D(M, D)$ for later use. This probability is derived from the probability distribution of every station, for every distance D, we look up the probability that a magnitude M will be detected at the stations that lie in this distance. The probability of detection $P_D(M, \underline{x})$ is then defined as the joint probability that 5 or more stations have detected this event. This number depends on the triggering condition of the network, the Northern Californian network uses five stations. Most events will be detected at more than five stations, depending on the size of the earthquake on ten or even more stations. Therefore it is easier to compute the joint probability by computing the probabilities of detection at zero, one, two, three stations and four stations and subtracting these values from 1. This gives us the probabilities of detection at 5 or more stations within the network.

$$P_D = 1 - P_0 - P_1 - P_2 - P_3 - P_4$$

Each determination of P_D is based on the input magnitude M, by changing this, we will compute another map. With these calculations, we can map the distribution of the probability of detection for one specific magnitude M. Changing the date of the this computation will lead to another station configuration and therefore will change the probability of detection. By computing maps for different dates, we can sketch the temporal changes in the studied network.
(2) We will also calculate probabilistic magintude of completeness (PMC) maps. These are based on the above calculations; instead of computing P_D for only one magnitude, we apply the above calculation iteratively. For each map, we chose a threshold probability P_t, as soon as P_D exceed this threshold, we take the value of M^* as our probabilistic magnitude of completeness M_{pc}:

$$\begin{aligned} P_D &= f(M^*, D) \\ \text{if } P_D &\geq P_t \\ M^* &= M_{pc} \end{aligned}$$

Just as in (1), we can vary the date of this calculation to obtain the temporal changes in PMC. To investigate small spatial changes in PMC, we can do this calculation for a small region with a high resolution.

Chapter 3

Data

We are analysing the period from January 1^{st}, 2001 until December 31^{st}, 2005. For this time period, we need two kind of empirical data: (1) Station data describing location and active time of each station of the network and (2) event data. The event data consists of hypocentral parameter data (including location, origin time and magnitude of event), and series of arrival times (including station name, seismometer component, picked arrival time and phase identification.)

3.1 Seismic Network

Seismic Stations For this approach, we are using the data of the Northern Californian Seismic Network (NCSN). This networks contains of several smaller networks, not all of them are used for triggering though. Table 3.1 shows the networks being used for this study. The largest of these networks is the NC network, maintained by the US Geological Survey, Menlo Park (`http://www.ncedc.org/ncsn/`), the second largest is the NN network, maintained by the University of Nevada (`http://seismo.unr.edu/`). All networks are shown in figure 3.1, blue triangles belong to NC, red ones to NN and the green ones to some smaller network. The network, not used are not shown in this figure, these are some smaller local network, for example the BG Berkley Geysers Network that is only used locally.

The complete station list can be obtained online at `http://www.ncedc.org/ncedc/station.info.html` or `http://www.ncedc.org/ftp/pub/doc/ncsn/`

No	Shortcut	Network name
1	BP	Parkfield High Resolution Seismic Network
2	BK	Berkeley Digital Seismic Network
3	NC	Northern California Seismic Network
4	NN	UNR Broadband Network
5	PG	Pacific Gas and Electric Seismic Network
6	WR	California Water Resources

Table 3.1: Station Network codes from the Northern Californian Seismic Network, more information about the networks is available online at: http://www.ncedc.org/ncedc/station.info.html and at http://www.iris.edu/stations/networks.txt

`ncsn.stations`. We select stations, which have at least been partially active in the period from January 1^{st}, 2001 until December 31^{st}, 2005. It is important to have exact knowledge about on- and off- times of this stations. If it is not reported that a station is inactive, it will be interpreted as not recording, this will corrupt the probability of detection for this station. Unfortunately there is no exact knowledge about these inactive times in this network; it is possible that a station was not recording for some time, before it was removed from the field and marked as offline. It is also possible that is was inactive for some time and then was repaired and started to record again; such things are not noted in the station information file unfortunately. This can alter the probability of some stations, but it is not possible to remove this influence for us.
We only use stations and components used for triggering; this means that we only select the vertical components from the above mentioned list. As mentioned above, stations are often moved over a couple of meters over time. Most stations become renamed after this, adding a B and then a C to the station code is common for the NCSN.

3.2 Earthquake data

Phasedata Access to all events registered by the stations of the NCSN is provided online from the NCEDC at: `http://www.ncedc.org/ncedc/catalog-search.html`. To obtain information about the earthquake and the picked phases, one has to choose the option "Catalog + Phase in Hypoinverse format". This format provides information about the type and quality of picked phases, the station and station components and earthquake details. The for-

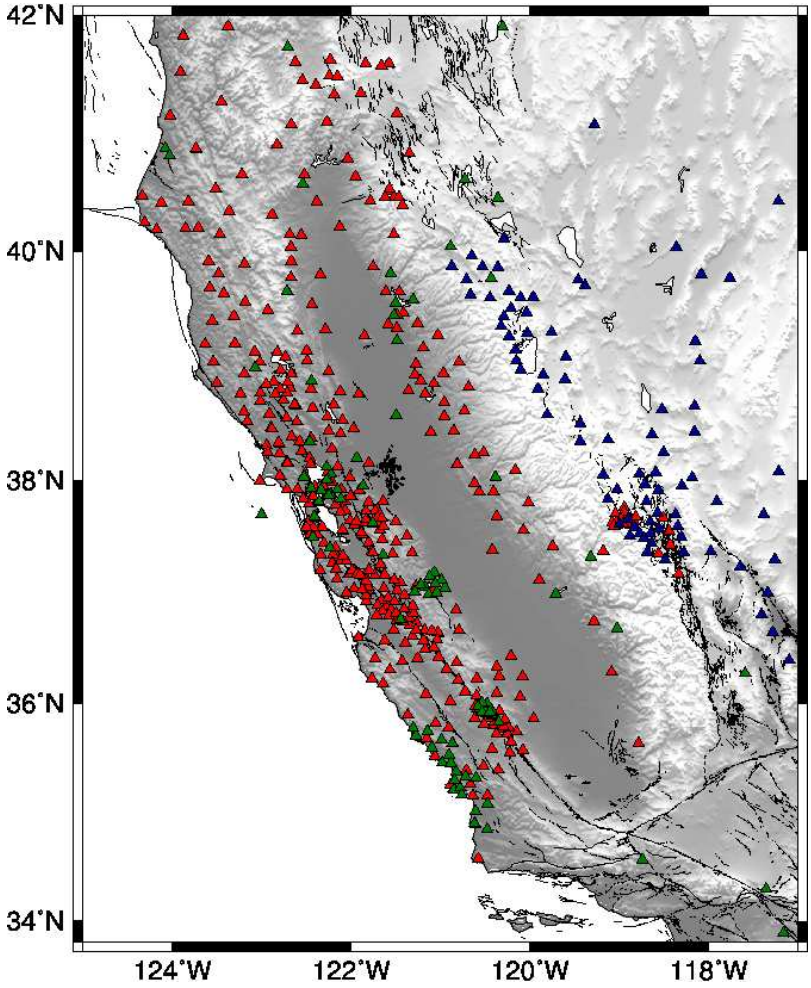


Figure 3.1: Map showing all stations used in this study. The blue triangles show the stations from the Northern Californian Seismic Network (NC), the red triangles show the UNR Broadband Network (NN) and the green triangles show the smaller networks (BP, BK, PG and WR). All stations have at least been partly active from 1^{st} January 2001 to 31^{st} December 2005.

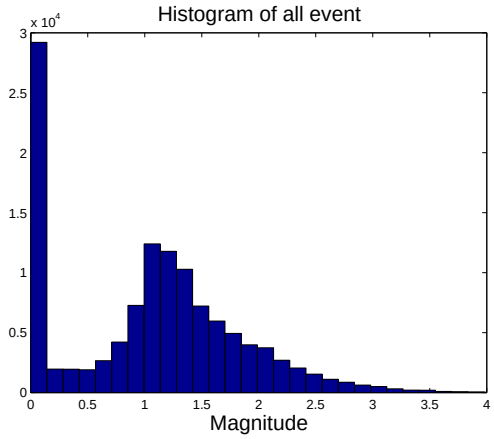


Figure 3.2: Histogram of the magnitude for all events from 2001-2005. It is obvious that there are too many events with magnitude M0, these events must have had a different magnitude that could not be determined.

mat was described by Klein [2006]. This file provides different magnitudes, depending on the type of data available, and the size of the earthquake. The coda duration magnitude is available for most events, but it saturates for events above 4.5. Therefore the NCSN also computes alternate magnitude from low-gain instruments that remain generally on-scale during large earthquakes. In addition to this, the catalogue provides local magnitudes, computed by UC Berkeley. The NCEDC provides a so-called preferred magnitude, which is the most reliable magnitude for each event [Klein, 2006], this magnitude generally follows this scheme:

if M $<$ 3 M = M_d (Coda Duration M.)

if M $>$ 3 M = M_L (Local M.)

if M $>$ 4.5 M = M_w (Moment M.) (only if good solution and only after 2000, otherwise M_L will be used)

If no magnitude can be calculated for an event, it is set to zero. This leads to a large amount of events with magnitude zero, this is shown in the histogram 3.2. It is obvious that there can not be that many events with this magnitude, these events must have had another magnitude that was not determined. The effect of this on the probability matrices will be shown in the next chapter. Excluding all events with magnitude zero lead to about 90'000 events for the time period from January 1^{st}, 2001 until December 31^{st}, 2005. All events are shown in figure

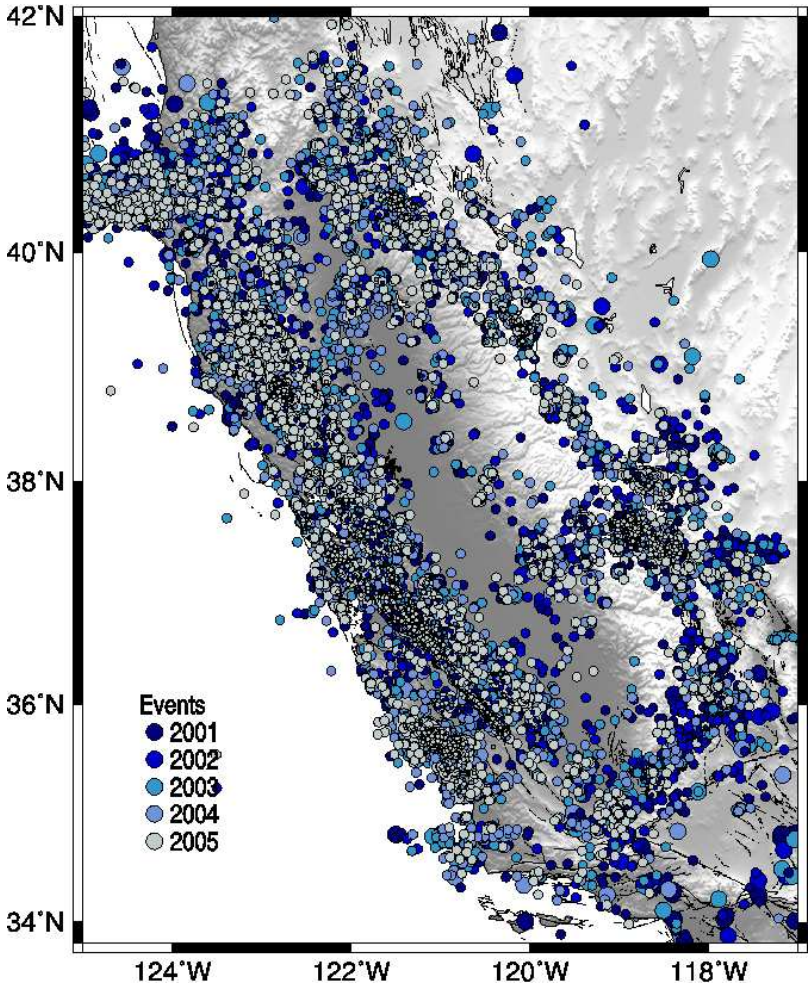


Figure 3.3: All events in Northern California from January 1^{st}, 2001 to December 31^{st}, 2005. Circle colours represent the year and circle sizes are scaled by magnitude.

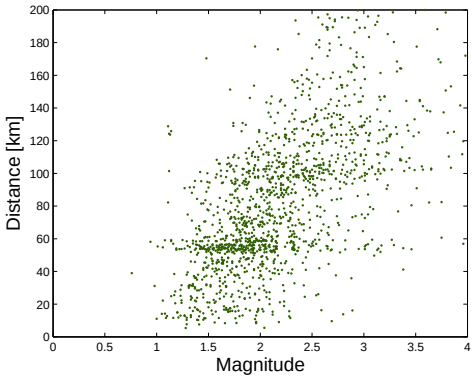
3.3; each colour represents one year and circle-sizes are scaled by magnitude. The activity does not change significantly over the years and it is mainly concentrated along on the faults, also indicated in figure 3.3. Most events occur along the San Andreas fault, but there is also volcanic and geothermal activity in this region.

Chapter 4

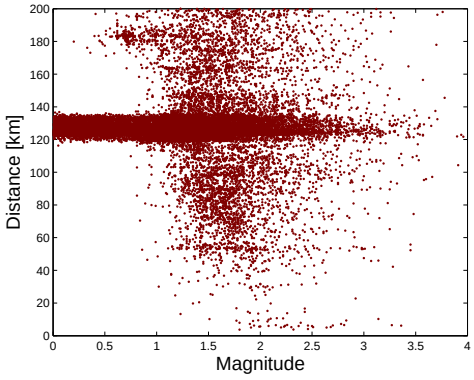
Station Analysis

4.1 Station Quality

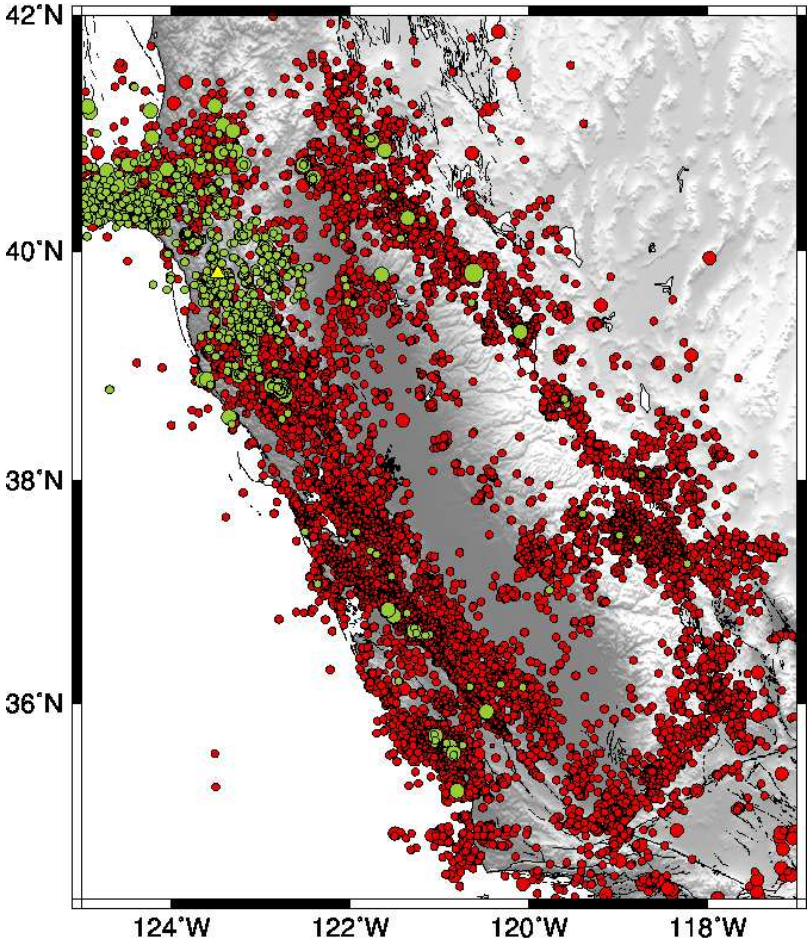
To analyse the features of one station, we extract the raw data of the picks. Two examples are shown in figures 4.1, 4.2 and 4.3, the first two are the same stations we used for examples above. In all three figures, (a) shows magnitude against distance for recorded events and (b) for missed events. In addition to this, (c) shows maps of the distribution of all events, red ones were missed by the station, green ones were recorded, larger squares indicate stronger events. The yellow triangle in this map shows the location of the analysed station.
Figures 4.1, 4.2 and 4.3 offer a convenient tool for the analysis of a single station. With the maps we can see which area is normally covered by a station. It is also easy to see which station can be considered as a good station or bad station.
For example, the station BJO is located near a fault, therefore there will be more events near this station and it is more likely that this station will record most of them. If we find a station, which is located in a high seismicity region that does not record as many events as expected, this is an indication that something is wrong with this station. For instance, it is possible that the noise level is to high at this station, or it is possible that the station is located on a different underground than a better recording station.
Figure 4.1 indicates that the station KIP is not as good as the station BJO. It can be seen that this station is located in a lower seismicity region, figure 4.1 (a) shows clearly that there are almost no events with magnitude higher than three within 40 kilometres distance from this station. Overall, the amount of events which this station was able to record is smaller and therefore this station



(a) Recorded Events



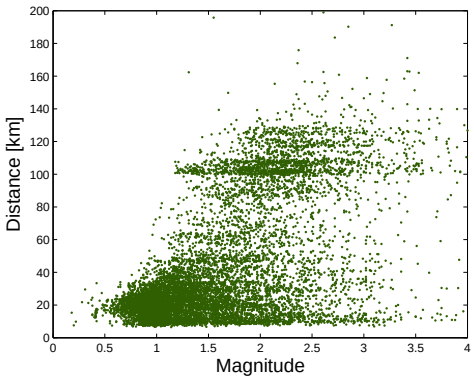
(b) Missed Events



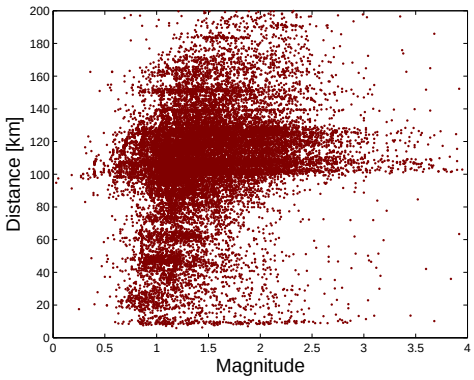
(c) Map of missed and recorded events

Figure 4.1: Analysis of the Station KIP. (top) Diagrams of magnitude against distance of recorded (a) and missed (b) events at the station. (bottom) Map of recorded (green) and missed (red) events, scaled by magnitude; station is indicated by a yellow triangle.

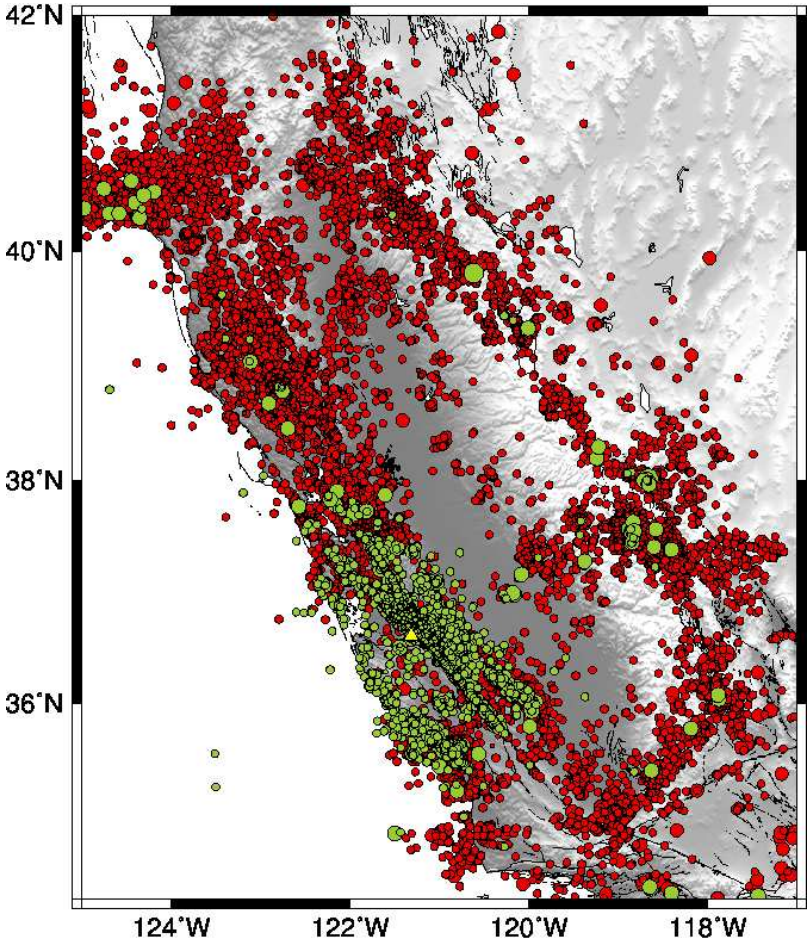
is considered to be a worse station than the station BJO.
Figure 4.3 shows an example of a reasonably well recording station. This station BCW is located about 40 kilometres away from the next high active region, this can be seen from fig. 4.3 (a) and (b), there is almost no data for any distance smaller than 40 kilometres. The map of the data distribution shows that this station is located near the coast, on a mountain range which is seismically quiet. Therefore this station records much fewer events than the station BJO, which is only about 30 kilometres away from this station.
By analysing single stations we can observe flaws in the data, which will be discussed in the next chapter.



(a) Recorded Events

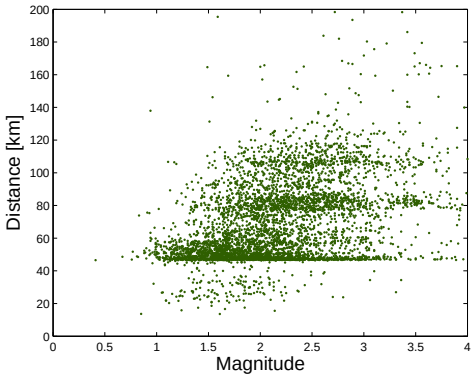


(b) Missed Events

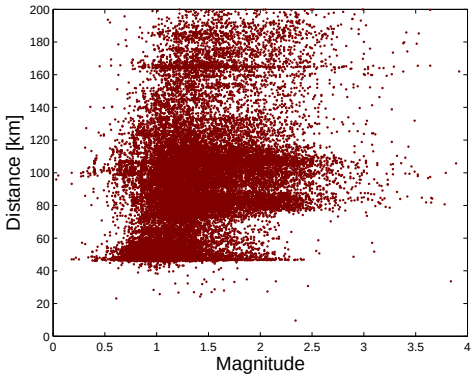


(c) Map of missed and recorded events

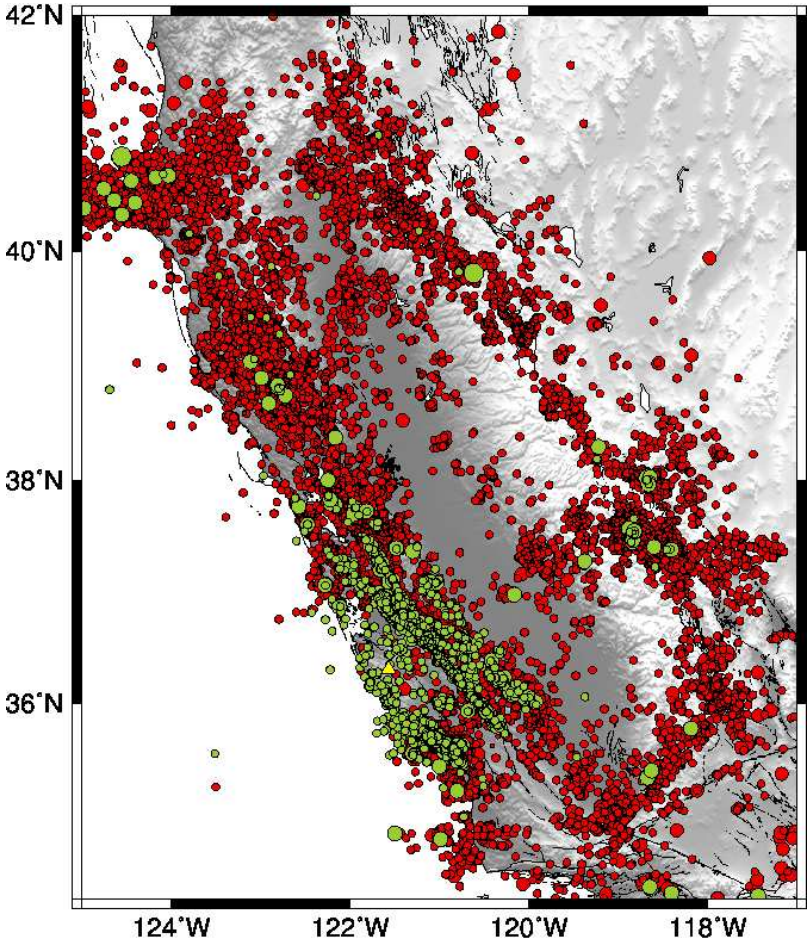
Figure 4.2: Analysis of the Station BJO. (top) Diagrams of magnitude against distance of recorded (a) and missed (b) events at the station. (bottom) Map of recorded (green) and missed (red) events, scaled by magnitude; station is indicated by a yellow triangle.



(a) Recorded Events



(b) Missed Events



(c) Map of missed and recorded events

Figure 4.3: Analysis of the Station BCW,(top) Diagrams of magnitude against distance of recorded (a) and missed (b) events at the station. (bottom) Map of recorded (green) and missed (red) events, scaled by magnitude; station is indicated by a yellow triangle.

Chapter 5

Data Flaws

5.1 Inspecting and reducing data flaws

5.1.1 Effect of events with magnitude zero

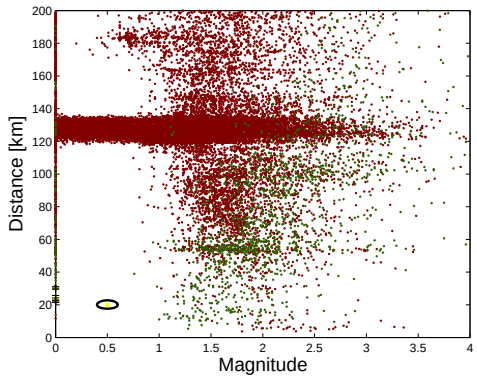
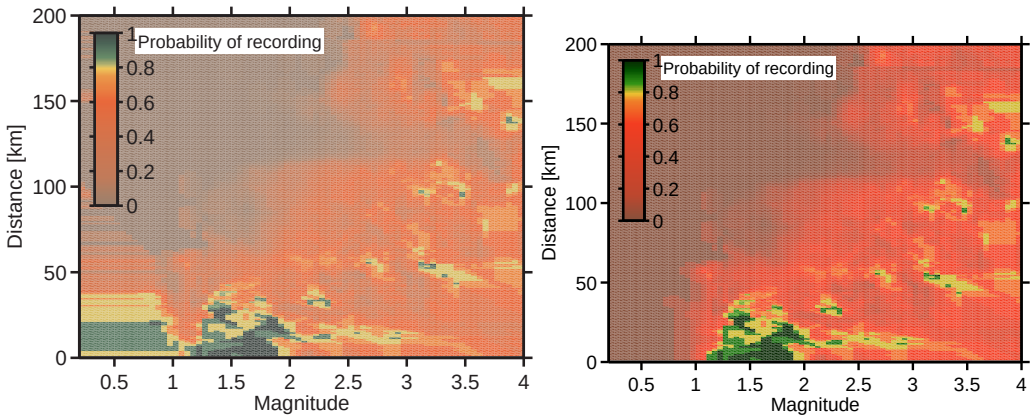


Figure 5.1: Magnitude against distance for recorded (green) and missed (red) events from 2001-2005 at the station KIP. Magnitude is set to zero if no magnitude could be determined, thus events with an artifical magnitude will be sampled to compute the probability at $M/D = 0.5/20$ (black crosses).

It was already mentioned above that events, for which no magnitude could be determined are assigned with magnitude zero. We observed this after the first computation of the probability matrices and the effect of these magnitudes was quite big. In all pictures that were shown above from the raw data, we already excluded the events with magnitude zero, figure 5.1 shows now the raw data with these events. Also shown on this figure are the events that will be sampled to compute the probability of detection for the point with magnitude M0.5 and



(a) Station KIP with M0

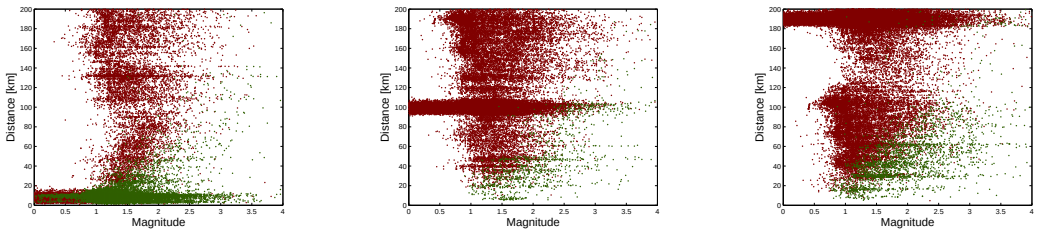
(b) Station KIP without M0

Figure 5.2: Two probability matrices of station KIP, for (a) events with magnitude M0 were included in the probability calculation, in (b) they were excluded.

distance 20 kilometres. Because there are no events within the ellipse, the second criteria will be applied. Because only events with magnitudes $M' \leq M$ and distances $D' \geq D$ will be sampled, we sample here events with magnitude zero. Some of these events were recorded and some of these were not, which leads to a complete artifical result for the probability at this point. Figure 5.2 shows the result of such a calculation, (a) shows high probabilities for low distances and small magnitudes. This result is highly unlikely. Figure 5.2 (b) shows the matrix that we already saw above, the only difference between these plots is that we did not include the events with magnitude zero for the calculation of the righthand figure.
As we can see, it is important to note that some events are assigned with magnitude zero, although the magnitude just could not be determined.

5.1.2 Earthquake Clusters

One of the most striking features of the raw data from the station KIP is the band of events in the distance range from about 120 to about 140 kilometres (see figures 5.1 or 4.1). There are over 30'000 events within this distance range. This station is not the only one where we observed such a feature, figure 5.3 shows more examples of such bands at different distances. The station GCR shows about 30'000 events within the first 20 kilometres, station CPM from 90 to about 110 kilometres and station JJO for a distance range from 180 to about



(a) Station GCR (b) Station CPM (c) Station JJO

Figure 5.3: Bands of events for different distances from the stations.

200 kilometres. We computed plots similar to figure 4.1 to 4.3 (c), but this time we only coloured events within the distance range of these bands. Figure 5.4 shows the example of the station KIP, where the distance band from 120 to 140 kilometres is highlighted. By inspecting many of such plots, we could find different sources of such bands with a dense event distribution, one origin are aftershock series, another one are earthquake clusters.

Aftershock Series During an aftershock series many events happen in the same regions and many of these events will be missed because their signal will be too small compared to the main shock. An example of such an aftershock series within our chosen time period, is the aftershock series of December 22^{nd}, 2003 San Simeon earthquake, with a magnitude M6.5. Figure 5.5 shows the events of this series. From the colour code, we can see that most events happened in 2004 and 2005, before that the region war relatively quiet. The yellow star shows the location of the main shock, most events happened to the south-east of this event.
Although this aftershock series caused some bands on some stations, its effect was not too large and was therefore not further inspected.

Earthquake Cluster An earthquake cluster is created by a region which is highly active, this can be a fault or a volcanic region for example. We found that the cluster from the Geysers Geothermal Field had the biggest influence in our region. This region is shown in figure 5.6, the colour code here shows that the amount of events does not change much over the years. Altough this region is small, there were over 30'000 events recorded over the period from January 1^{st}, 2001 until December 31^{st}, 2005. Most of the events happend at very shallow

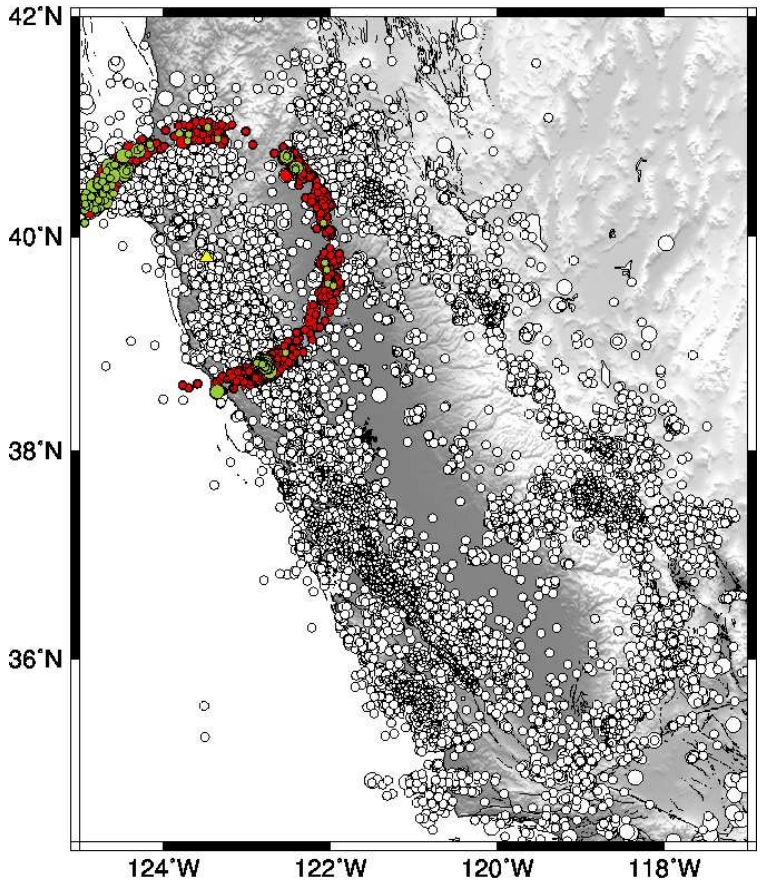


Figure 5.4: Map showing recorded (green) and missed (red) events within a distance range from 120 to 140 kilometres from station KIP; white circles indicate all events occuring during the active time of KIP. All circle sizes are scaled by magntidue.

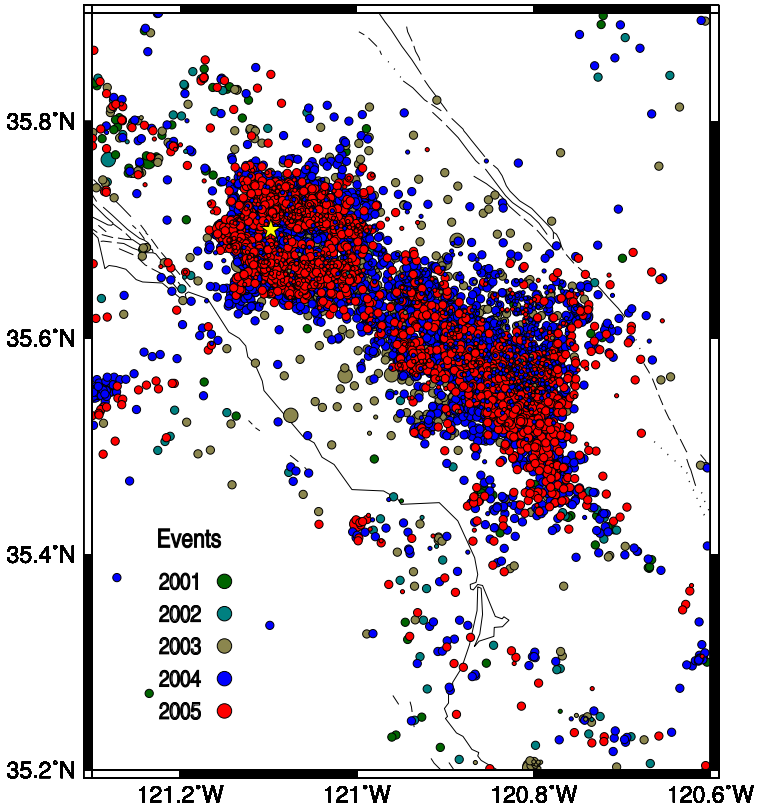


Figure 5.5: Aftershock series of the San Simeon 6.5 earthquake of December 22^{nd}, 2003; mainshock is indicated by a star, events are coloured by year and scaled by magnitude.

depth, which results in weak signals, which will not be recorded too far away. However, there is a local network, called Berkley Geysers Network, BG (yellow triangles in figure 5.6) installed to record all these events. Therefore they will be included in the catalogue, although they were missed by most stations. The stations from the BG network were not used in our analysis, because are only used to record events from the Geysers field, and mostly fail to detect any events outside this region.
However, as these events are in the catalogue, they reduce the probability that an event will be recorded at any station in the distance range in which the Geysers Geothermal Field is located. This influence is shown in figure 5.7, on the left hand is the probability matrices with all events, the right side shows the result, when the events within the Geysers field are removed. The black ellipse highlights the difference between the two figures, the probability increases for a distance range of 120 to 140 km when the Geysers events are removed.
Altough the Geyser events are limited to a small region (fig 5.6), their influence on the probabilities is large. Each probability matrix reflects the ability of a station to detect an event in a circle around the station location, therefore the Geysers events reduce the probabilities for a whole distance range for each station. Excluding the Geysers geothermal field from the earthquake catalogue leads to more reliable probability matrices. We could only overcome this influence by computing isotropic probabiliy matrices, this will be an subject for future studies.

5.1.3 Excluding picks not used in the location process

Another feature we found during the inspection of station qualities were events, which could not possibly be detected, but still were. An example of this is shown in figure 5.8, it shows on the left hand events that the station GFC detected and on the right side the ones station BMR detected. On both figures, events which would not be detected under normal circumstances are highlighted with a red circle, these events generally have a too low magnitude to be detected at such distances. Therefore, we decided to investigate these events further and found that the only thing they all had in common was that they were not used in the inversion process to determine the hypocenter of an earthquakes. Hence, we decided to exclude all picks from our analysis not used in the inversion. We only observed after that, that there are many picks not used in the inversion

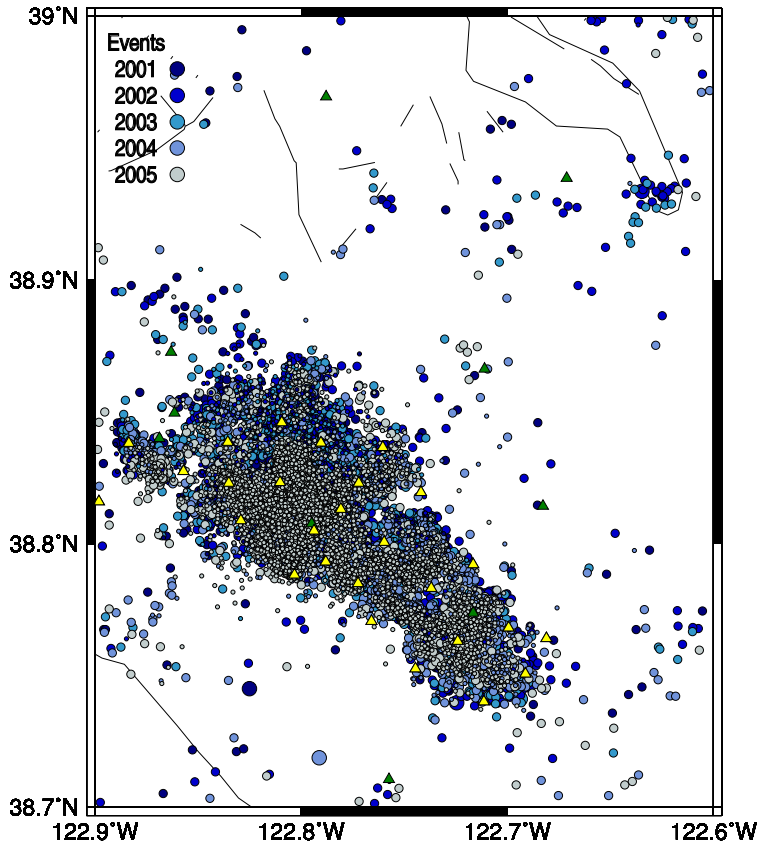
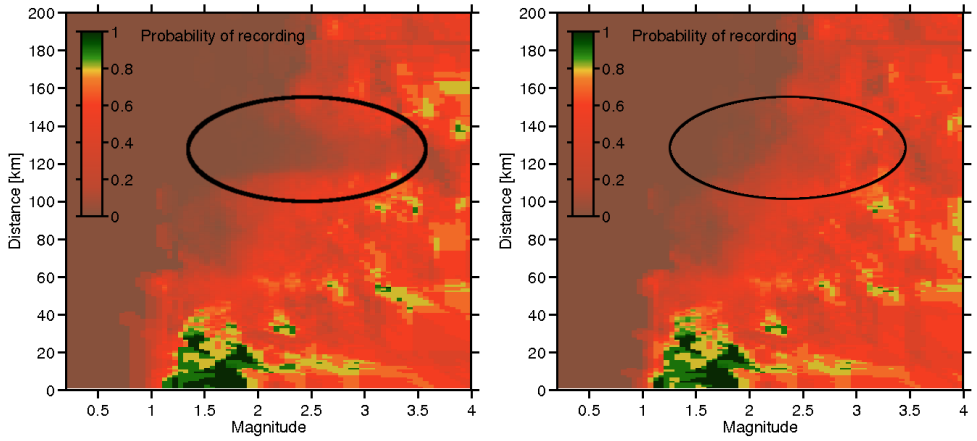
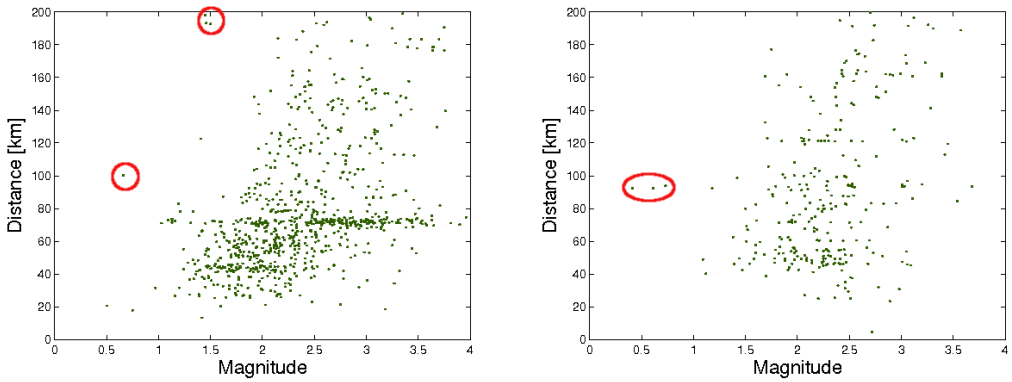


Figure 5.6: Earthquake cluster created by the Geysers Geothermal Field; events are coloured by year and scaled by magnitude. Triangles indicate location of seismic stations, green= NC network, yellow= BG network.



(a) KIP with data from Geyser (b) KIP without data from Geysers

Figure 5.7: Comparison of probability matrices from the station KIP with, and without the data of the Geysers region. The black ellipses highlight the differences between both matrices, eliminating the Geysers events leads to higher probabilities in the right figure.

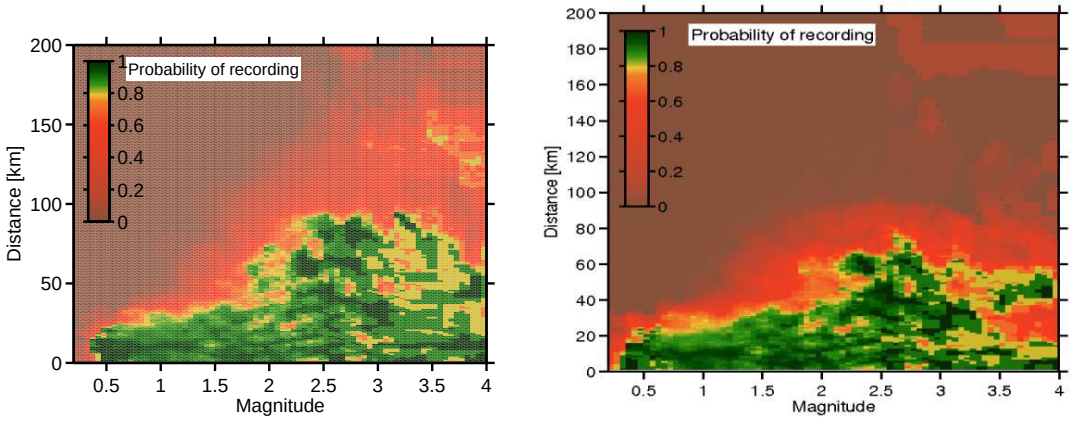


(a) Recorded events at station GFC (b) Recorded events at station BMR

Figure 5.8: Recorded events at two different stations, red circles show events which should not be recorded under normal circumstances.

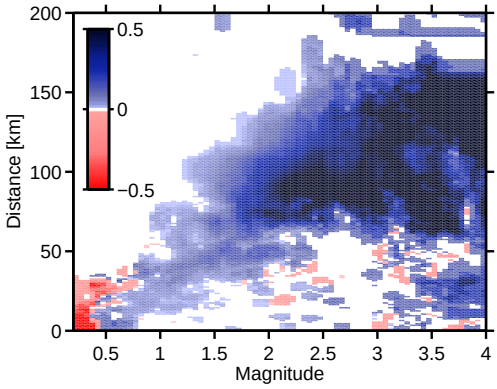
process, for some station we reduced the picks to 50 percent of the primary data.

So on one hand, excluding the picks, which are not used in the inversion process leads to more reliable data, as we are only using picks now, which were in fact used to determine the hypocenter of events. On the other hand, we exclude stations, which recorded an event in the first place. There are different reasons, why the pick of a station is not used, some times the station is just too far away, so there were better picks in the near vicinity of the event. So the station was just not used because it was too far away, while the signal was still good enough. It is also possible that the station was not used, because the signal was too weak or that there were other reasons, which made the station pick infeasible, but it is not possible for us to distinguish between these reasons.
Figure 5.9 shows the comparison between two probability matrices for the station BJO, (a) shows the probability matrix, which is based on the raw data only, (b) shows the result, if the picks not used in the inversion process, are excluded. Figure 5.9 (c) shows the difference between them, positive numbers (blue) mean that we obtained a higher probability in (a), negative numbers (red) mean that the probability computed in (b) is higher. We see that the probability based on the raw data is generally higher; only for distances smaller than 40 kilometres and magnitudes smaller than M0.5 the second data set leads to higher probabilities. The station BJO is not used in the inversion process at many magnitude-distance combinations, where it can be expected that the



(a) Raw data

(b) W/o picks not used in inversion



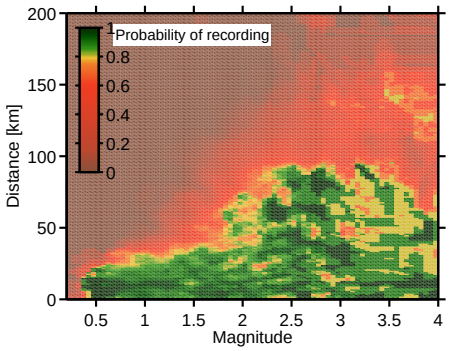
(c) Difference between (a) and (b)

Figure 5.9: Probability matrices of the station BJO, (a) is based on the raw data, in (b) no picks, which were not used in the inversion process were used. (c) shows the difference between the two plots, using the raw data leads to higher probabilities from about 60 kilometres on.

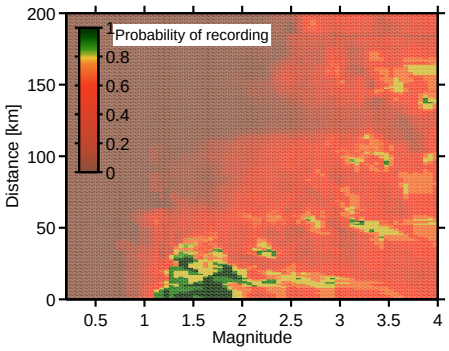
signal still should be good enough. As we saw in the previous chapter (see figure 4.2 (c)), the station BJO is located near an active fault. Therefore the station density is high and the signal of this station will not be used in the inversion process to locate an event, because there will be other stations nearer the event.
So, although it seemed like a good idea in the beginning to exclude all picks not used in the inversion process, to get more reliable data, we abandon this idea, because we excluded much to much data. This would not have been a problem, if we not also excluded data, which was in fact reliable and would have been used, if the station density was not that high. We are also mainly interested in stations used for triggering and as the stations recorded a signal, there will have been involved in this process, which is another reason for us to refrain from this idea.

5.1.4 Excluding automatic picks

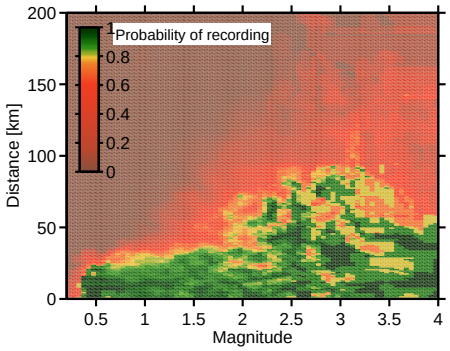
All triggering procedures include different steps, in the NCSN it is implemented that at least five station have to record a signal above a threshold to trigger an event. This signals are recorded automatically and are only reviewed manually if the signal is used. Therefore there will be many automatic picks, which can represent real picks, but they don't have to. It is possible that the noise was increased at a station, during an actual earthquake somewhere in the region of the network. Then, the signal that was recorded at this station will be interpreted as a pick information, altough it was just noise. For this reason, we decided to exclude the automatic picks and investigate the results of this. Using only manual picks will lead to more reliable data, but to much fewer pick information.
Figure 5.10 shows two examples of probability matrices we obtain, when we exclude the automatic picks. The top shows the probability matrices obtained with the raw data, the middle the ones without the automatic picks and the lower figures show the differences between the above plots. The difference plot in figure 5.10 (b) is not as significant as the difference plot in figure 5.9 (c), at some places, excluding the automatic picks can even lead to a higher probability, but over all the probability decreases. For the station KIP, we observe the same characteristics; excluding the automatic picks leads to a lower probability. For this station we also see that the difference is zero at many places, this means



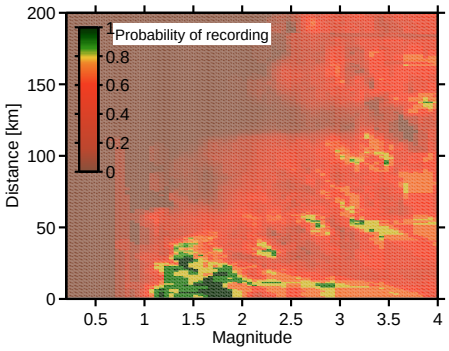
(a) BJO, raw data



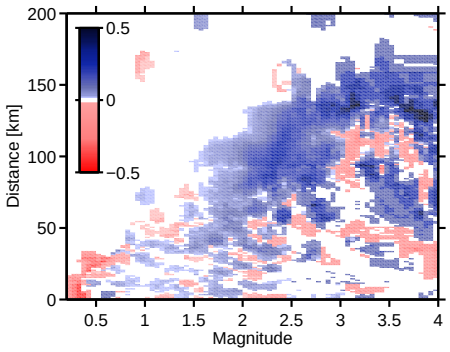
(b) KIP, raw data



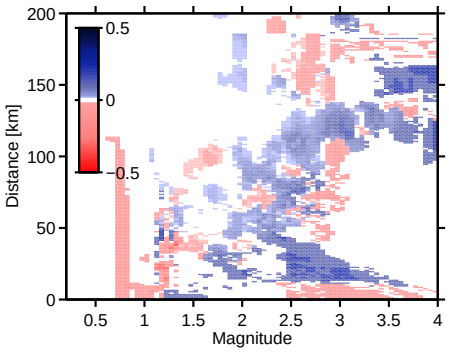
(c) BJO, w/o automatic



(d) KIP, w/o automatic



(e) BJO, difference plot



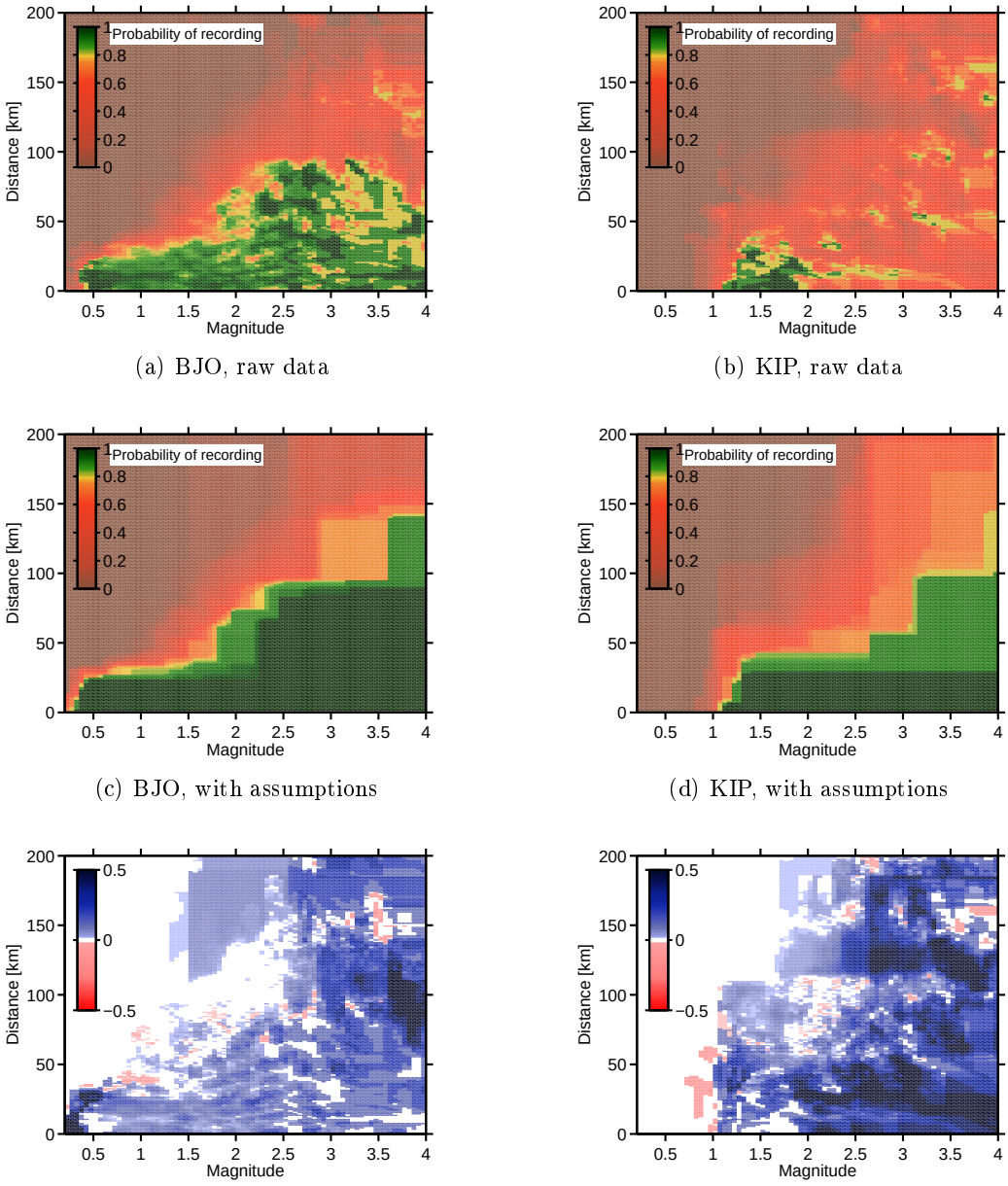
(f) KIP, difference plot

Figure 5.10: Left side = station BJO, right side = station KIP. (top) Probability matrices obtained with raw data, (middle) probability matrices obtained with the exclusion of automatic picks and (bottom) difference between above plots.

that there were not as many automatic picks as at the station BJO.

5.2 Summary

We showed here several possibilities, how the reliability of the pick data can be improved. Excluding picks not used in the inversion process does not improve our results, as we are excluding picks, which were good enough to be recorded, but they were just not used because there were better picks in the vicinity of the earthquake. Therefore we would reduce our data without a prober reason. Excluding automatic picks will lead to more reliable data on one hand, on the other hand, automatic picks do not have to be wrong. If they seem to be ok, they will not be picked manually again, so we are excluding data, which was good data in the first hand.
In the end, we chose to only exclude events with magnitude zero and events from Geyser Geothermal Field. In addition to this, we included that the probability can not decrease with increasing magnitude and constant distance and it can not decrease with decreasing distance and constant magnitude, as it was described in § 2.1. Figure 5.11 shows the final probability matrices we used for our calculation, again, figures on the left side show probability matrices and figures on the right side show the difference plots between the calculations based on the raw data and the calculation without the events from the Geysers region and with the probability dependence. We can clearly see that we increased the probabilities almost everywhere.



(a) BJO, raw data

(b) KIP, raw data

(c) BJO, with assumptions

(d) KIP, with assumptions

(e) BJO, difference plot

(f) KIP, difference plot

Figure 5.11: Left side = station BJO, right side = station KIP. (top) Probability matrices obtained with raw data, (middle) probability matrices obtained with the exclusion of events from the Geysers Geothermal field and with the addition of the constraints described in § 2.1 and (bottom) difference between above plots.

Chapter 6

Results

6.1 Maps

We will presents first the maps showing the probability of detection, $P_D(M, \underline{x})$, for a specified magnitude. Then we will present results of our calculation of the probabilistic magnitude of completeness. These maps show for every point the magnitude threshold, above which we can assume that the NCSN is complete. For both maps, we will present different maps, which represent the different steps in our analysis.

6.1.1 Probability of Detection Maps

Figure 6.1 shows the $P_D(M, \underline{x})$ map, where M is equally to one. This map results, if we are using the raw data, without excluding any picks or any events. It does also not include the dependence of the probability on the magnitude and the distance. We obtained this map without any assumptions. Figure 6.2 shows the result we obtain, if we exclude all events within the Geysers Geothermal Field. As we showed above, excluding these events leads to higher probabilities within the distance range of this cluster for every station. Therefore it not surprising that the probability to detect an event with magnitude M1 is higher in figure 6.2 than in figure 6.1. The effect is also not limited to the actual region of this cluster (compare fig. 5.6), as the probabilities were increased not only at this place, but for a whole distance range around every station.

Figure 6.3 shows our final result, in this map we excluded the events from the Geysers cluster and we implemented the dependence of the probability on magnitude and distance. This means that the probabilities do not decrease

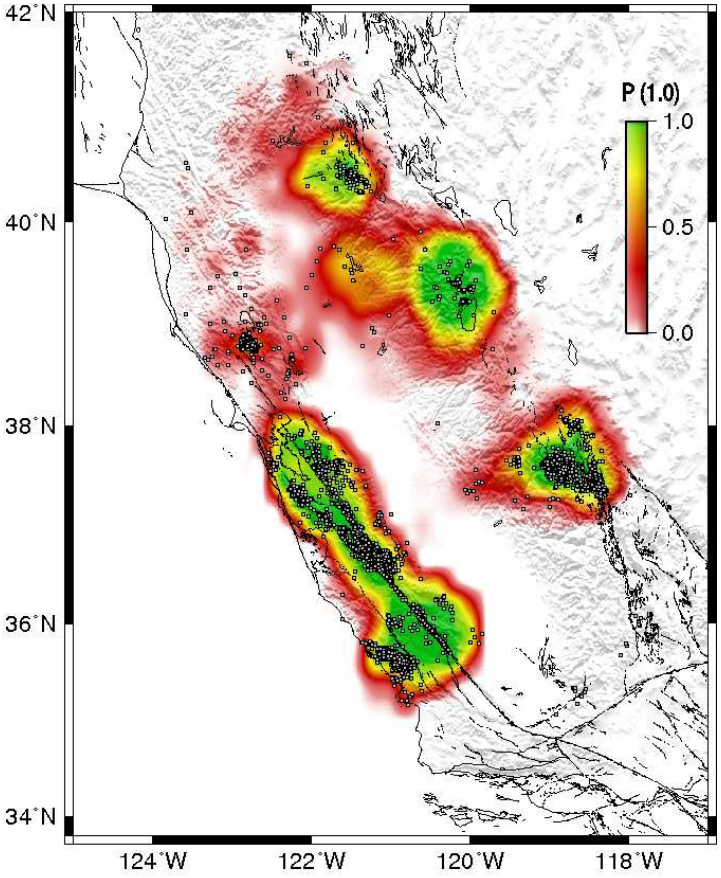


Figure 6.1: Probability of detection map, for magnitude M1. The colour bar indicates the probability level, with which a magnitude M1 event will be detected at the specified place. Squares indicate all events with magnitude M1 or lower, which were recorded from 2001 to 2005. This map is based on the raw data only.

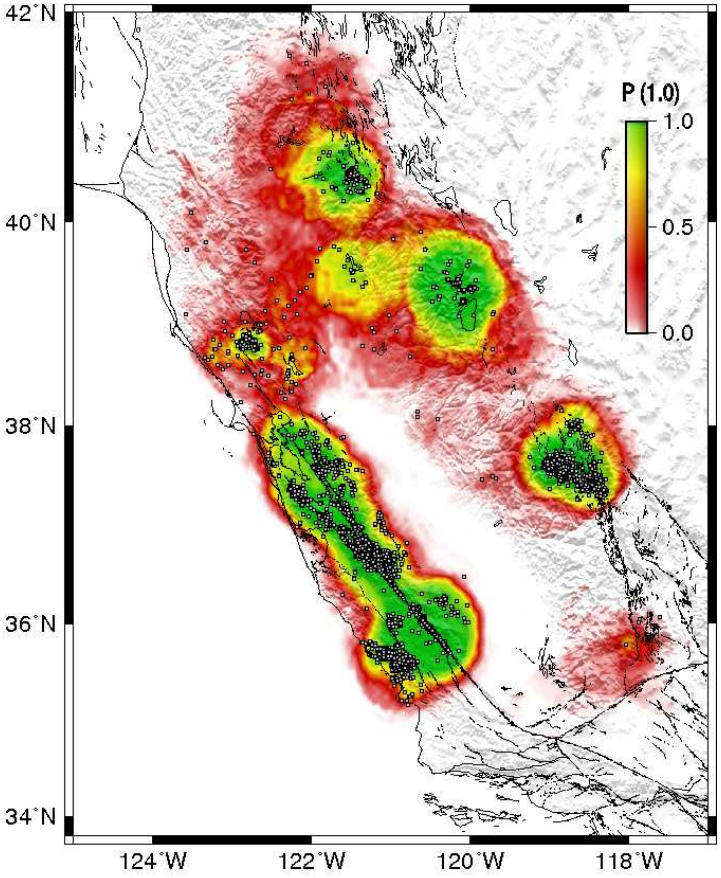


Figure 6.2: Probability of detection map, for magnitude M1. The colour bar indicates the probability level, with which a magnitude M1 event will be detected at the specified place. Squares indicate all events with magnitude M1 or lower, which were recorded from 2001 to 2005. For the calculation of this map, we excluded all events from the Geysers Geothermal Field.

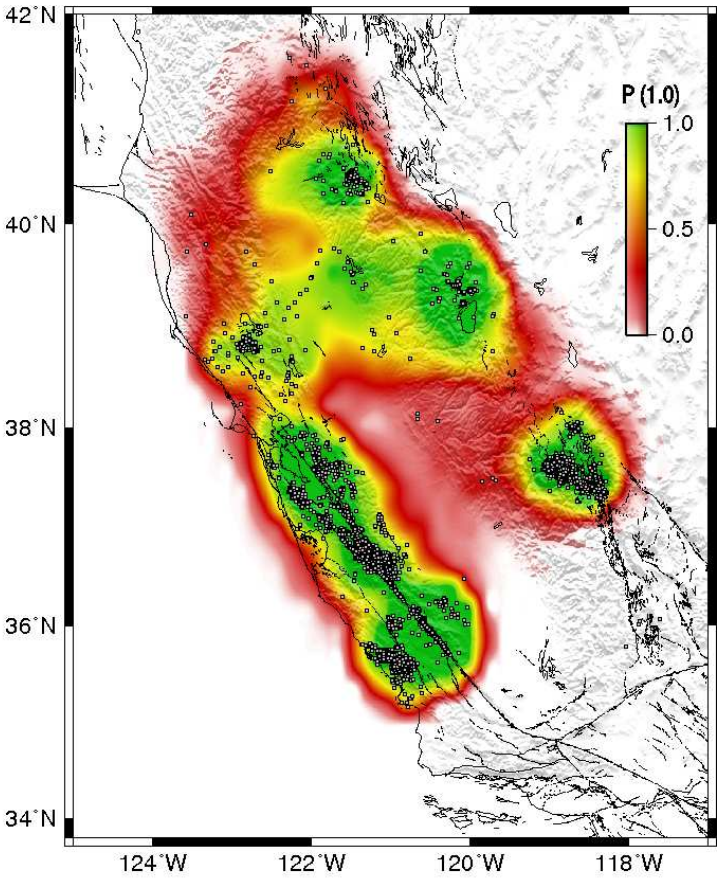


Figure 6.3: Probability of detection map, for magnitude M1. The colour bar indicates the probability level, with which a magnitude M1 event will be detected at the specified place. Squares indicate all events with magnitude M1 or lower, which were recorded from 2001 to 2005. For the calculation of this map, we excluded all the event from the Geysers Geothermal Field and we included a dependence of the probability on the magnitude and the distance.

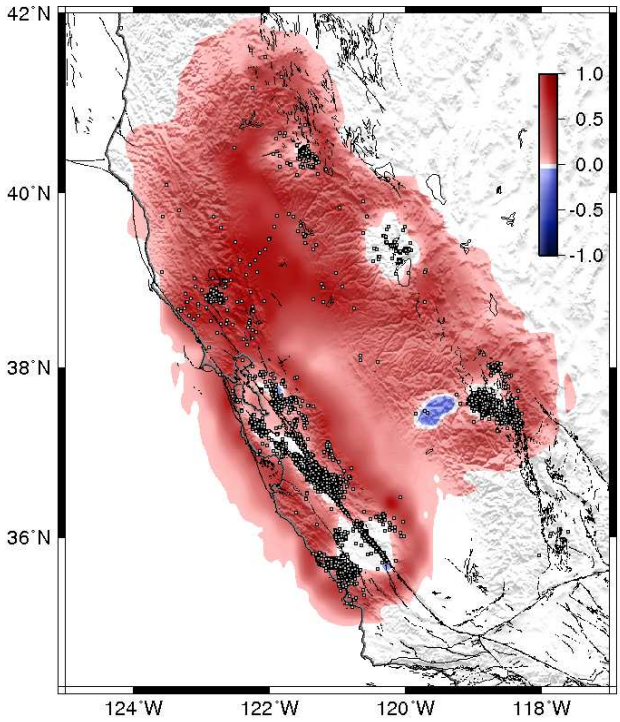


Figure 6.4: Difference in probability of detection between figure 6.1 and figure 6.3. The probability that an event with magnitude M1 will be detected is generally higher for the second case.

with increasing magnitude at constant distance and they do not decrease with decreasing distance at constant magnitude.

As we already saw in the probability matrices in figure 5.11, we obtain a higher probability if we include this assumption. The low probabilities, which are obtained due to a lack of data, are eliminated now. We see in figure 6.3 that we obtain an overall probability of detecting an event with magnitude M1, which is higher than the probability in the above figures. This is again visualised in figure 6.4, which shows the difference between figure 6.1 and figure 6.3. We already reached a probability of one at some places in figure 6.1, there we see no difference in both figures. In all other regions, we obtain a higher probability of detecting with our assumptions. The effect is highest in less active region, these are regions which lead to a lack of data. With our assumption that the

probabilities do not decrease for an increasing magnitude with constant distance and a decreasing distance with constant magnitude, we can correct for these lacks.

6.1.2 Probabilistic Magnitude of Completeness Maps

In this section, we will present maps of the probabilistic magnitude of completeness. We will present different maps, for the different correction steps we have made. All maps, show from which magnitude on the Northern Californian seismic catalogue can be expected to be complete. This magnitude varies in the following maps from M0.5 to about M2.5. Figure 6.5 shows the result, we obtain if we base our calculations on the raw data. In figure 6.6, we show the result, which we obtain if we exclude the Geysers cluster and if we include the above mentioned dependence of the probability on magnitude and distance.
Each map is based on the station configuration of one specified date, we only consider stations which were active during this day for the calculation. Figure 6.5 and 6.6 are based on the station configuration on January 1^{st}, 2006, figure 6.7 is based on the configuration on January 1^{st}, 2003 and figure 6.9 on January 1^{st}, 2001. On each of these maps, we show the station, which were used for the particular calculation. Figure 6.7 and 6.9 will show the influence of the different station configuration.

Regional differences in the probabilistic magnitude of completeness The first figure 6.5 shows the result we obtain, if we use the plain data, without any assumption. We obtain the same patterns in this map as in the above maps of the probability of detection. We are obtaining mainly three regions, where the PMC is significantly lower. This is especially the case if the region is of interest and the station density is therefore higher. One of the regions is along the San Andreas fault from around Parkfield up to the San Francisco Bay Area. This area is the most prominent feature on our maps. Based on the raw data, we obtain values for PMC from 1.2 down to 0.5. Figure 6.6 shows the result, we obtain if we exclude the Geysers cluster and include the assumptions. The small region, where we obtain a PMC of 0.5 with the raw data has now been greatly enlarged. In the southern part, the green pattern covers a broad region and more to the north, we find a PMC of 0.5 along the San Andreas Fault up to the San Francisco Bay Area.

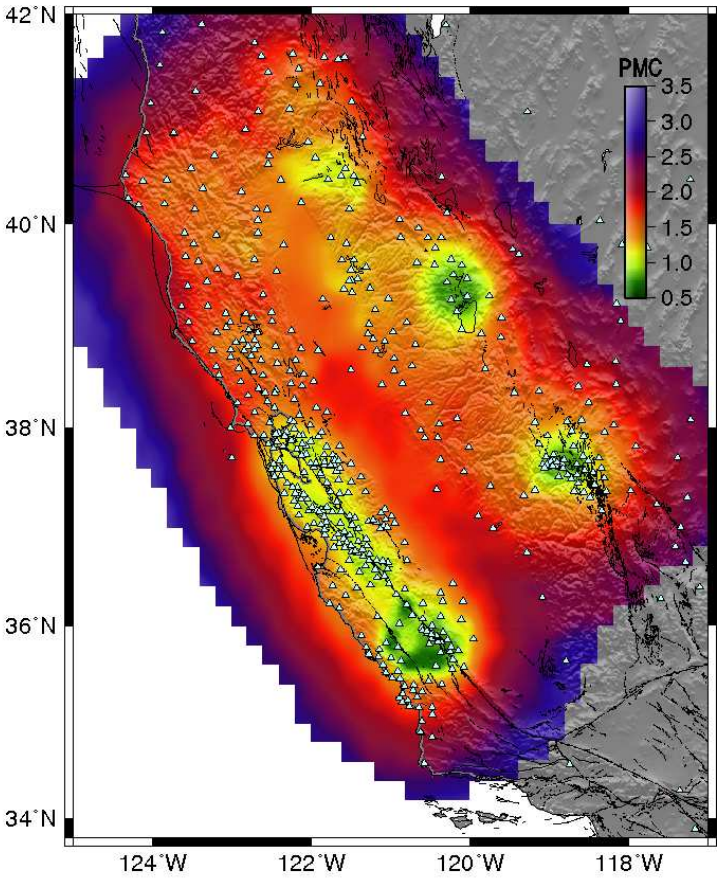


Figure 6.5: Probabilistic magnitude of completeness map, based on the raw data only. The colour bar indicates the probabilistic magnitude of completeness and triangles indicate the location of the stations, which were active on the 1/1/2006.

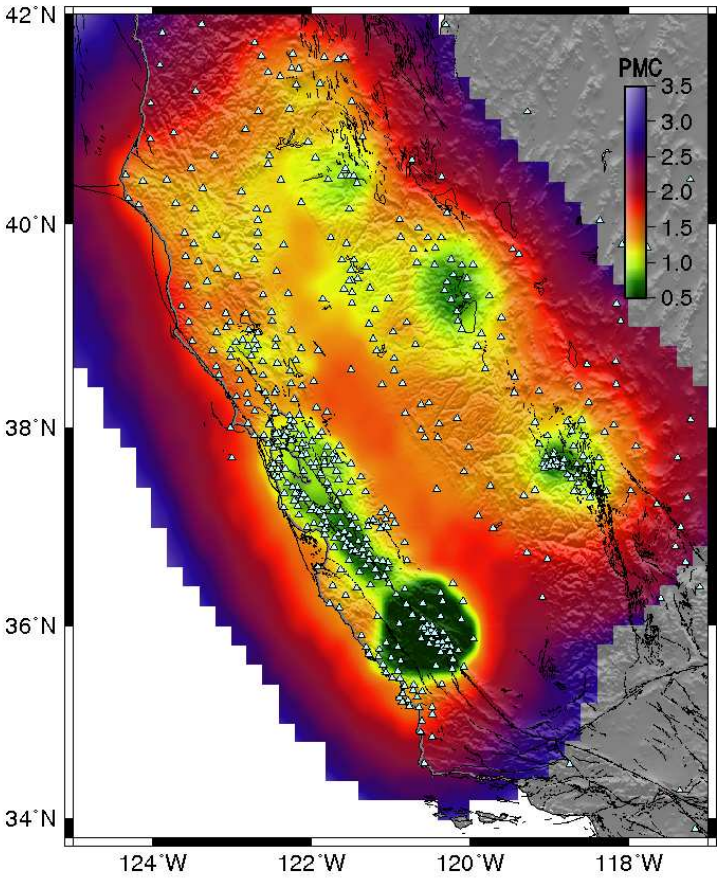


Figure 6.6: Probabilistic magnitude of completeness map, where the Geysers cluster is removed and the assumption about the dependence of the probability on magnitude and distance is added. The colour bar indicates the probabilistic magnitude of completeness and triangles indicate the location of the stations, which were active on the 1/1/2006.

A second region, where we find low probabilistic magnitude of completeness is the region north of Lake Tahoe, where a volcanic field is located. In figure 6.5, we see that we obtain values down to 0.5 in the middle of this pattern and values around one at the borders. In figure 6.6, the region where we obtain a value of 0.5 is broader. The pattern is also no more isolated, but it connects with the pattern more in the south. This region with a low probabilistic magnitude of completeness is located near Yosemite National Park. This region corresponds to the volcanic field at Mammoth Lakes in the Long Valley. Based on the raw data, we obtain values around one there, but in our final result we obtain values

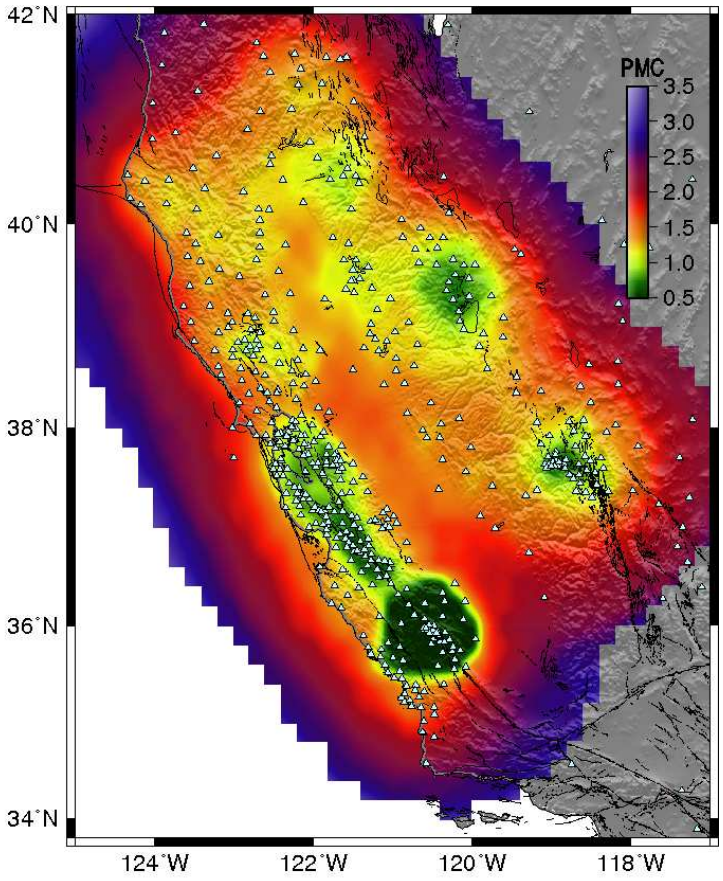


Figure 6.7: Probabilistic magnitude of completeness map, where the Geysers cluster is removed and the assumption about the dependence of the probability on magnitude and distance is added. The colour bar indicates the probabilistic magnitude of completeness and triangles indicate the location of the stations, which were active on the 1/1/2003.

of 0.5 there.

The highest PMC values are found in the Central Valley in both maps. This region is less active and there are fewer stations there (compare figures 3.1 and 3.3), based on the raw data we obtain values around 1.5 up to 2 there. The advantages of our assumptions are demonstrated here again. Because we can reduce the effect of too few data, we can obtain lower values at regions with a lower seismicity. Figure 6.6 shows values of about one for much broader regions and values of about 1.5 are only reached in the middle of the Central Valley.

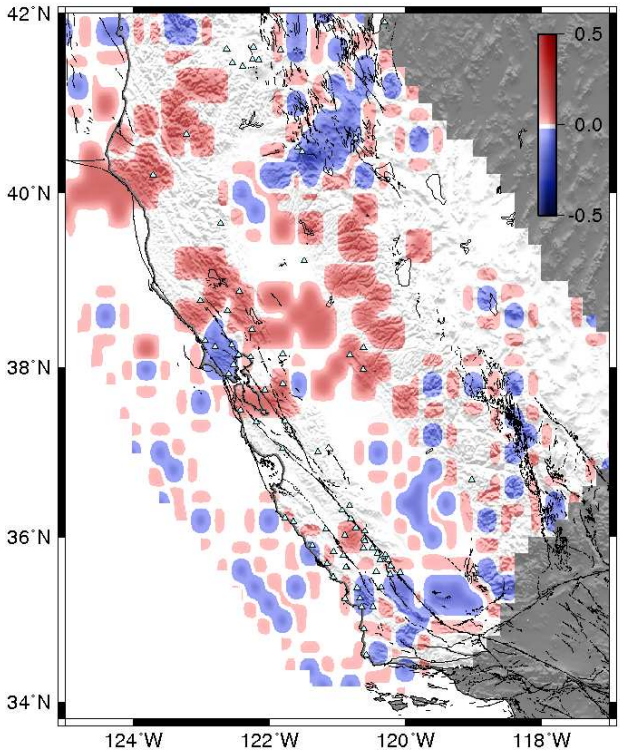


Figure 6.8: Difference between figures 6.6 and 6.7. There were not many changes in the network between the 1/1/2006 and the 1/1/2003, so the changes in M_c are minor. The triangles mark stations, which became active after the 1/1/2003.

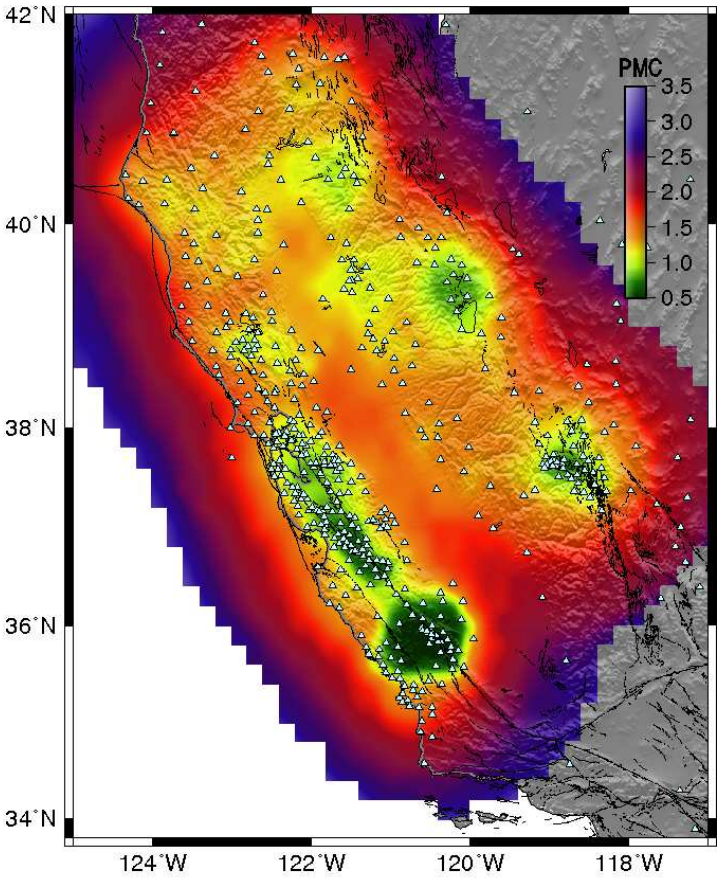


Figure 6.9: Probabilistic magnitude of completeness map, where the Geysers cluster is removed and the assumption about the dependence of the probability on magnitude and distance is added. The colour bar indicates the probabilistic magnitude of completeness and triangles indicate the location of the stations, which were active on the 1/1/2001.

Difference in the completeness over years. The maps in figures 6.5 and 6.6 are both calculated for the station configuration on the $1st$ January 2006. We also calculated the maps of the probabilistic magnitude of completeness for other dates, to investigate the influence of a altered station configuration. Figure 6.7 shows the result of a map for the January 1^st 2003 and figure 6.9 shows the result of a map for the January 1^st 2001. Both maps are calculated with the same data as figure 6.6. We can see, how the probabilistic magnitude of completeness changes, if we add more station or if a station is removed. Adding a station lowers the value, because this station will contribute to the new calculation and therefore increase the probability that an event will be recorded. This effect is higher, if a new station is added where the station distribution was sparse before. Adding a station somewhere, where the density of the station distribution is already high will not have such a big effect. It will still lower the obtained value, but not that significantly. Removing a station will of course have the opposite effect. It also makes a difference there, where this station is removed and how good the recording capability of this station was. Removing a station, which had a high probability to record an event in a region where the distribution is sparse will have the biggest impact on the obtained probabilistic magnitude of completeness values.

Figures 6.8 and 6.10 show the difference between the result with the station configuration on the January 1^{st} 2006 and the configuration on the January 2003 1^{st} and the January 2001 1^{st} , respectively. The changes in the configurations are indicated by the stations, which became only active after the specified date. There are not many stations which became active after the January1^{st} 2003, so there is not much change visible in figure 6.8. In contrast to this, there are more changes in figure 6.10, the most prominent feature are the changes around the southern part of the San Andreas fault; installing more station lowered the values for probabilistic magnitude completeness in 2006. In addition to this, there are more station in the San Francisco Bay area, which also lowers the PMC. The stations, which were removed are not shown on this map, but it is obvious that stations must have been removed in the northern part of the region, because the values in 2001 have been lower than the values in 2006.

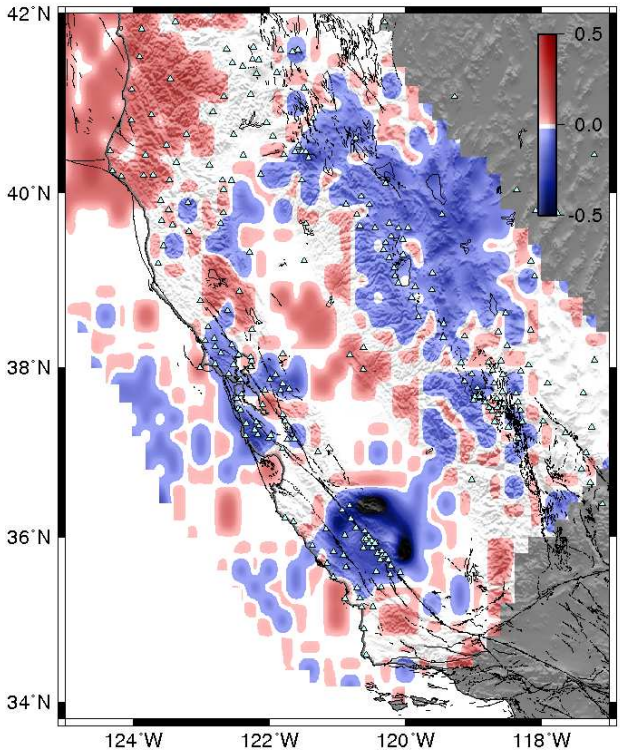
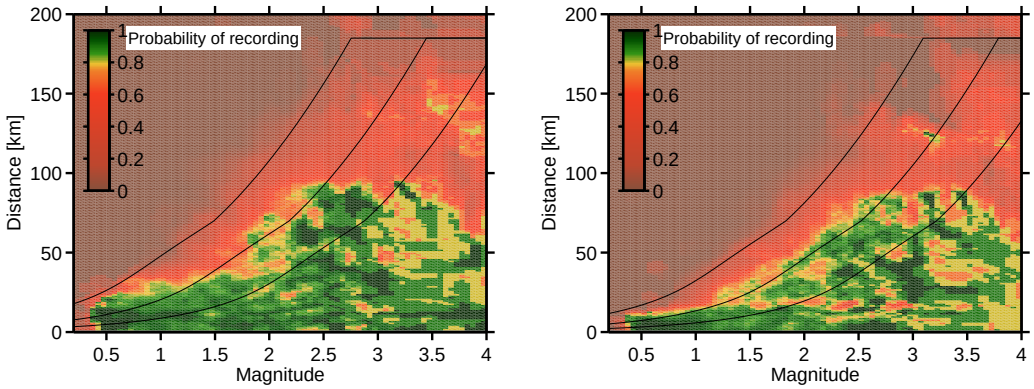


Figure 6.10: Difference between figure 6.6 and 6.9. The effect of the new station along the San Andreas fault can clearly be seen, the values for M_c decreased there. New stations in the San Francisco Bay area also lead to lower values in 2006. The triangles mark stations, which became active after the 1/1/2001.



(a) Station BJO

(b) Station BBG

Figure 6.11: Probability matrices of station BJO and station BBG. In both we included the attenuation curve, for three different amplitudes.

6.2 Attenuation Properties

The original magnitude from Richter is

$$M_L = \log\left(A_{WA}/2\right) - \log(A_0)$$

where A_{WA} is the peak-to-peak amplitude on a standard Wood Anderson torsion seismograph and $-\log(A_0)$ is an attenuation term and is a tabulated function of distance [Richter, 1935, 1958]. As mentioned above, in Northern California, the following formula is in use

$$M_L = \log\left(A_{WA}/(2 \times CAL)\right) + F_1(s) + F_2(d) + XCOR_{comp} + XCOR_{sta}$$

where CAL is a dimensionless scaling factor assigned to each station and XCOR are two corrections made for the component and the station. $F_1(s)$ and $F_2(d)$ are two distance correction terms.

We can now include this relation in our probability matrices, this is shown in figure 6.11 for three different amplitudes A_{WA}. We can clearly see that the slope of the probability and the equation fit reasonably good. The correlation is better for the station BBG, but for we see a match in both figures.

This means basically that the probability matrices, we compute represent for every station is attenuation properties.

Chapter 7

Discussion

We showed in the introduction the importance of a good knowledge of M_c. Here we will first discuss the advantages of our method over traditional methods that rely on various assumptions. Then we will briefly overview the steps we made to improve the quality of our data and show our motivation of excluding only the Geysers cluster and adding our physical constraints.

7.1 Comparison with traditional methods

Traditional methods to estimate the completeness of a seismic catalogue are based on several fundamentally different assumptions, the methods were already briefly mentioned in the introduction.

Methods, based on the Gutenberg-Richter FMD One approach is to investigate the deviation from the Gutenberg-Richter frequency magnitude distribution (FMD); methods, based on this, differ in the way how they determine this deviation point. The Gutenberg-Richter FMD [Gutenberg and Richter, 1944] describes the relationship between the frequency of occurrence and magnitude of earthquakes:

$$\log N(M) = a - bM$$

where N is the number of earthquakes with magnitude larger than M, and a and b are constants. Figure 7.1 shows this relation, the red line is an example of how the FMD can be fitted to the data. In this case, this was just made on a visual basis, but it should demonstrate the relation between the earthquake distribution and the power-law. There are different attempts on how to fit the

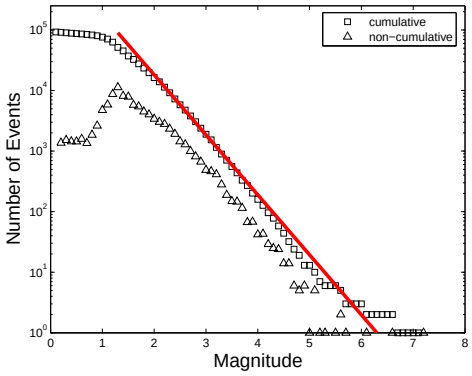


Figure 7.1: Figure, showing the cumulative and non-cumulative distribution of the magnitude against the logarithmic number of events. The red line shows an example of the Gutenberg-Richter power-law.

data to the power law; they all have to fit b-values to the earthquake sample.

In his study on Californian earthquakes, Marsan [2003] calculated the b-values for different magnitude bands above a cut-off magnitude, based on the model of Utsu [1966]. He defines the log likelihood of completeness as the logarithm of the probability that the best Gutenberg-Richter law fitted against all earthquakes with a magnitude above the cut-off can predict the number of earthquakes in the magnitude range just below this cut-off magnitude. He chose the minimum magnitude of completeness so that (1) the b value drops for magnitudes smaller than M_c and (2) the log likelihood drops for magnitudes equal M_c. This should then indicate that too few earthquakes occur with magnitudes smaller M_c as would be expected from the best Gutenberg-Richter law. Figure 7.2 visualises this method; the arrow in figure 7.2 (b) should indicate a drop in the b-value and in figure 7.2 (c) the arrow indicates the drop in the log likelihood. This picture shows already difficulties in this method, it is not certain, where the b-value drops. In addition to this, Woessner and Wiemer [2005] showed that the two criteria are difficult to combine if the calculation of M_c is done automatically. They also found instabilities, when they calculated the log likelihood for only one mangitude bin, as the frequencies of events within the magnitude bins vary strongly.

Wiemer and Wyss [2000] saw the necessarily to map the completeness spatially. They evaluate the goodness of fit by computing the difference between the observed FMD and a synthetic distribution; a simple power law will not be able to explain the observed FMD, if the data set is incomplete, making the

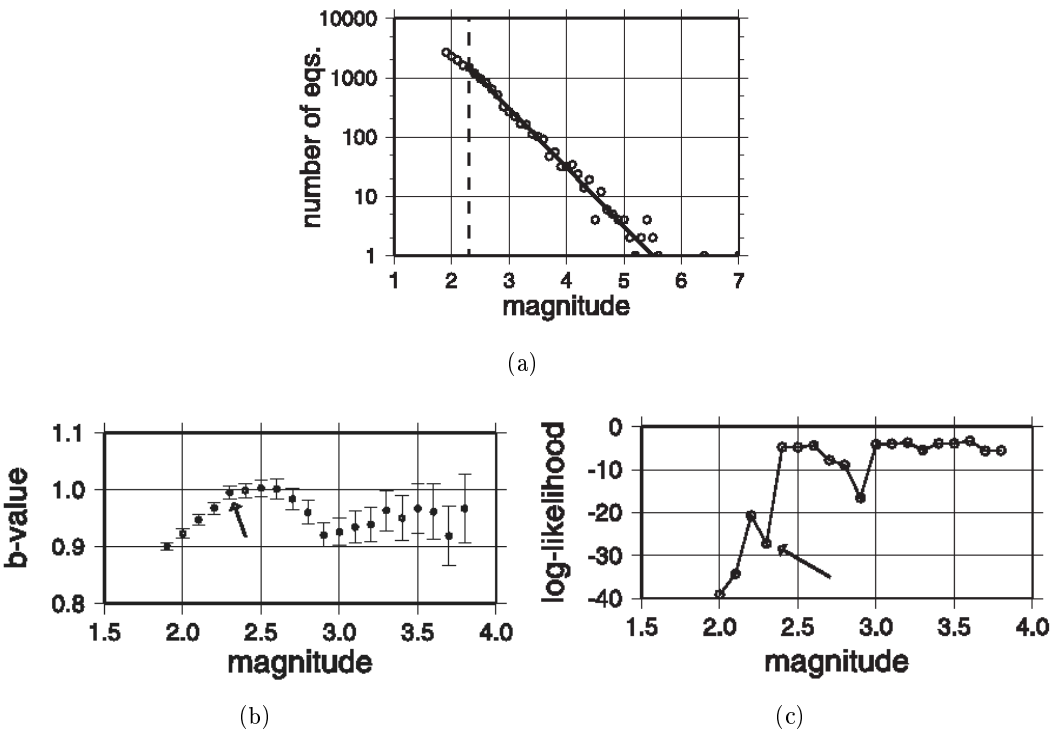


Figure 7.2: Figure, showing the method of Marsan [2003]. (a) shows the magnitude-frequency graph of their sample, the thick line is the best Gutenberg-Richter law. (b) shows variations in the b-value, the arrow indicates were the b-value drops. (c) shows the log likelihood, the arrow indicates again the drop.

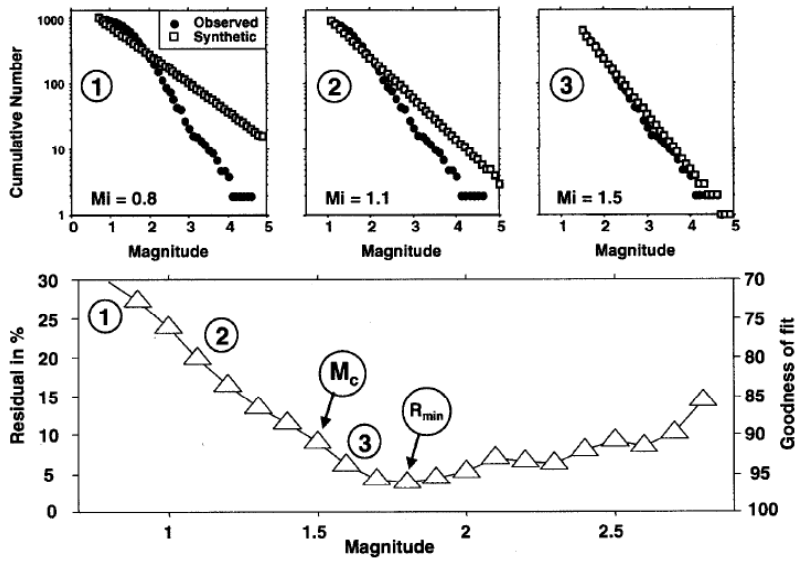


Figure 7.3: Figure, showing the method of Wiemer and Wyss [2000]. The difference between the observed and the synthetic distribution decreases from (1) to (3), the bottom shows the goodness of fit. M_c is taken where 90% of the observed data are fitted with a straight line.

difference high. Figure 7.3 shows how the data is fitted by a simple power law; the goodness of fit increases when the power law approaches the r̈ealm̈agnitude of completeness and decreases if it is exceeded. Wiemer and Wyss [2000] take M_c at the 90% level; 90% of the observed data will be modeled by the power law then. However, this level may not be reached if the FMDs are too curved to be fitted by a simple power law. They investigate regions, where their fit was poor and conclude that non-power law FMDs may have three different sources (1) artefacts in the catalogue, as a result of M_c changes over time, (2) mixing of heterogenous population of events, e.g. volcanic earthquake families and tectonic earthquakes and (3) spatially heterogenous M_c distributions. After eliminating such contaminations, one should be able to model each earthquake population with a power law for a wide range of magnitudes. However, this implies that every single earthquake population must be identified, to be able to fit a power law. This means that a lot of work must be invested in analysing the earthquake catalogue, as the completeness estimate will be wrong, if a earthquake population can not be identified. There are also drawbacks for quiescence regions, it wont be able to find a earthquake population there and therefore it wont be able to determine the completeness there. Although, the

method maps M_c spatially, temporal changes are ignored.
Woessner and Wiemer [2005] introduced the so called Entire Magnitude Range (EMR)-Method. They fit the complete part of the data with the FMD and search for models to fit the incomplete part. They tested different models and found that a normal cumulative distribution function fits the data best. The magnitude of completeness is then found, where the joint log-likelihood of the fit of the two models is the highest. This model does also not account for distributions not following the FMD.
Cao and Gao [2002] base their estimation of M_c on the stability of the b-value as a function of M_c. They use the maximum likelihood method of Aki [1965] to calculate the b-value

$$b = \frac{\log e}{\overline{M} - M_c}$$

where M_c is the cut-off magnitude (similar to our use of M_c) and $\overline{M}$ is the average of a group of earthquakes with $M \geq M_c$. They assume that b-values increase for $M* \leq M_c$, remain constant for $M* \geq M_c$ and increase again for $M* \gg M_c$. The authors defined the magnitude M_c as the value for $M*$, when the change in the b-value between two steps is smaller than 0.03. Woessner and Wiemer [2005] tested this approach and found that it is unstable, as the frequency of events in single magnitude bins can vary strongly.
All of the above mentioned methods base the estimation of the completeness of a catalogue on earthquake samples that should follow the Gutenberg-Richter frequency magnitude distribution. They do not take temporal changes in the completeness into account, altough some try to map spatial changes, they still rely on earthquake samples that have to be sampled over a certain space.

Comparing Day-To-Night Ratios Rydelek and Sacks [1989] investigated the day-to-night changes of magnitude bins. They assume that cultural activity and winds increase the noise level on seismogram during the day causing magnitudes to be missed during the day, while they are recorded at night; figure 7.4 show this schematically. They assume a Poisson distribution and say that every deviation from this distribution must come from cultural or solar thermal sources, if they investigate a time period of 24 hours. They then perform a random walk simulation to determine the changes from day to night ratios. Figure 7.5 shows three different steps of this procedure, for three different magnitude

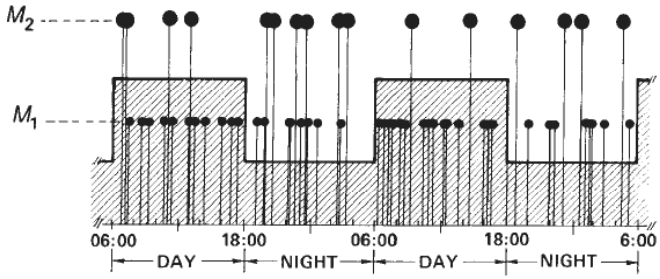


Figure 7.4: Figure, showing the method of Rydelek and Sacks [1989]. Due to the higher noise, the magnitudes M_1 are missed during the day, but recorded at night.

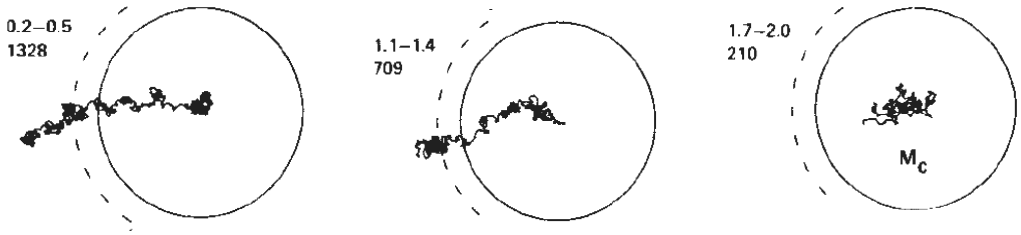
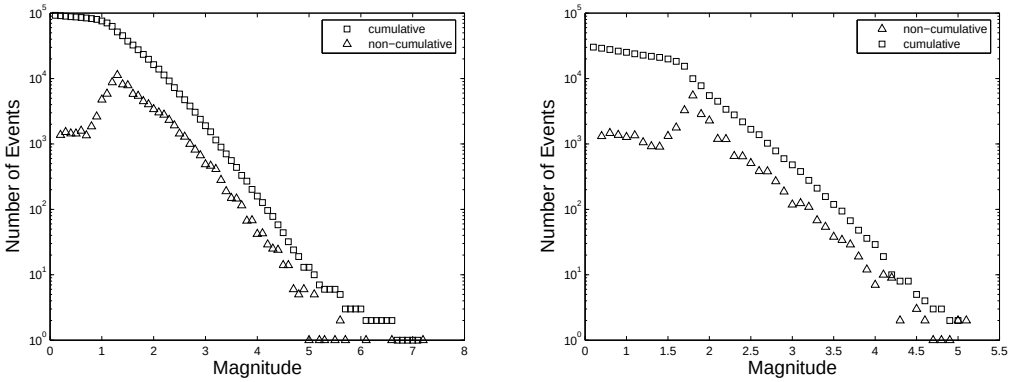


Figure 7.5: Three different steps of the random walk, numbers indicate the magnitude range and the numbers of events within this range. Different confidence levels are indicated: 95% (solid) and 99% (dashed) [Rydelek and Sacks, 1989].

intervals: (a) M0.2 to M0.5, (b) M1.1 to M1.4 and (c) M1.7 to M2.0. When the events stay within the 95% confidence level (solid circle), the catalogue is assumed to be complete. This method is able to determine M_c for regions where the FMD differs from linearity, but it has also drawbacks. It assumes that deviations from a Poisson distribution only come from day-to-night changes, but there are also other non-random features that can cause these, examples are quarry blasts, aftershock sequences or swarms. If there are features like this in the catalogue, the randow walk will lead to different results. This method will also not be applicable to regions that have low cultural noise, the 95% confidence level will there be achieved with an incomplete catalogue, examples of such regions are deserts.

Using Signal-To-Noise Ratio Kværna et al. [2002a,b] use the signal-to-noise ratio at a particular station to determine a threshold from which on a signal is



(a) All events of the catalogue

(b) Events within Geysers Cluster

Figure 7.6: Cumulative number of events against magnitude. The fluctuations in this distribution are much higher for the Geysers cluster.

supposed to be higher than the noise. This method, based on waveform calculation, is still too time-consuming to be practical for most regions. However, the method does not make the same assumptions as the methods above and, if improved, may be applicable to more regions.

Advantages of our method The biggest advantage of our method is that we do not base our estimation of the magnitude of completeness on earthquake samples, but all on events within the seismic catalogue and the station configuration. We do not assume that the events follow a Gutenberg-Richter based frequency-magnitude distribution. This distribution is violated at various places, for example in geothermal or volcanic regions; the Geysers Geothermal field and the volcanic field near Mammoth Lake are just two examples in our study regions, where it is highly likely that the event distribution deviates. Figure 7.6 shows this distribution for the events of the whole catalogue and only for the events within the Geysers cluster. The fluctuations for this field with geothermal activity in the event distribution are higher than the fluctuations for the whole catalogue, which would it make more difficult to fit a b-value to the data.

With our method, we also do not have to assume that the completeness is constant over space or time; we can compute an estimate of M_c for every date and this will be only based on the station configuration on this specified date. Once the probability distributions for each station are computed, we can compute the

probabilistic magnitude of completeness for every region within the network. We presented in this work maps for the whole region with a resolution of 0.2°, but it is possible to map a single region within Northern California with a much higher resolution. The computation-time depends on the size of the region and the resolution, maps that we produced take about 48 hours.
However, we have to make sure that the triggering condition was constant over the time period during which we collect our data. If this condition changes, we have to define a new period and base the completeness on this new period.
We also make one assumption: we assume that we can add the conditions that the probability of detection does not decrease with increasing magnitude at constant distance and that it does not decrease with decreasing distance at constant magnitude. This only means that we assume that a station follows some simple physical conditions. A station will record a stronger signal for an event with a high magnitude than for an event with a smaller magnitude, if they occurred at the same distance from this station. Therefore it should not be possible that the capability of a station to record an event with a lower magnitude is higher at the same distance. This can only be the case if there are no events of this magnitude within the specified distance of the station. The same hold for events a different distances and constant magnitudes. The signal should be stronger, if the event occurred within a shorter distance and therefore the capability of the station to record these events should be higher or at least equal. By adding our assumptions we are accounting for these effects. The only flaw in this assumption is that errors can occur if there is incorrect data in the event catalogue. If an event is wrongly reported as recorded at a certain distance/magnitude combination, this will also affect combinations with a lower distance and a higher magnitude. However, the effect is not very large, if just one event is reported wrongly, as the probability depends on the sample of at least ten events, but if there is a systematic error, it will affect the probability.

7.2 Steps of reducing data flaws

We introduced several steps of data correction. First, we excluded all events with a magnitude zero from our data, because they lead to artifical probabilities for small distances and low magnitudes. While analysing the stations we came

across dense bands of events within certain distance ranges at different stations; we detected that the origin of such bands are earthquake clusters or aftershock series. We decided to remove the earthquake cluster with the biggest influence in our region; the Geysers cluster. It was necessary to remove this cluster, because it influence the probabilities not only in the region where it was located. As the probability matrices we compute are isotropic, the Geysers cluster reduces the probability within a whole distance range of a station, not only at its origin. We did not remove other clusters or aftershock series, because we found that their influence is minor compared to the Geysers cluster. To overcome such an influence of earthquake clusters, a future approach should include the direction, when calculating the probability distributions of the stations.
We also investigated the effect of removing other data, e.g. all picks not used in the inversion process to locate an event, or all automatic picks. We believe that it is not reasonable to exclude all picks not used in the inversion process, because most of these were not bad picks, but not used simply because there were better picks nearer to the event. Excluding all automatic picks increases the reliability of the data; however this does not mean that all automatic picks are bad; there were just not checked manually because there was no need for it. Therefore we decided to neither exclude the picks not used in the inversion nor the automatic picks for our final results.

Bibliography

K. Aki. Maximum likelihood estimate of b in the formula $\log n = a - bm$ and its confidence limits. *Bull. Earthquake Re. Inst., Tokyo Univ.*, 43:237–239, 1965.

B. Bender. Maximum likelihood estimation of b values for magnitude grouped data. *Bull. Seismol. Soc. Am.*, 73(3):831–851, jun 1983.

A. M. Cao and S. S. Gao. Temporal variation of seismic b-values beneath northeastern Japan island arc. *Geophys. Res. Let.*, 29(9), 2002. doi: 10.1029/2001GL013775.

V. D'Amicio and D. Albarello. The role of data processing and uncertainty management in seismic hazard evaluations: Insights from estimates in the Garfagnana - Lunigiana area (Northern Italy). *Natural Hazards*, 29:77–95, 2003.

T. De Crook. Analysis of input data and seismic hazard assessment for the low seismicity area 'belgium, the netherlands and nw germany'. *Natural Hazards*, 2:349–362, 1989.

Y. Dodge. *Analysis of Experiments with Missing Data*. Wiley New York, 1985.

J. Eaton. Determination of amplitude and duration magnitudes and site residuals from short-period seismographs in Northern California. *Bull. Seismol. Soc. Am.*, 83(2):533–579, 1992.

M. C. Gerstenberger, S. Wiemer, L. M. Jones, and P. A. Reasenberg. Real-time forecasts of tomorrow's earthquakes in California. *Nature*, 435(7040):328–331, 2005.

D. Giardini, S. Wiemer, D.Fah, and N. Deichmann. Seismic hazard assessment of switzerland, 2004. Technical report, Swiss Seismological Service, ETH Zurich, 2004.

B. Gutenberg and C. F. Richter. Frequency of earthquakes in California. *Bull. Seismol. Soc. Am.*, 34:185–188, 1944.

F. Klein. Y2000 shadow format and NCSN data codes. Available online at: `http://www.ncedc.org/ftp/pub/doc/ncsn/shadow2000.pdf`, 2006.

T. Kværna, F. Ringdal, J. Schweitzer, and L. Taylor. Optimized seismic threshold monitoring - part 1: Regional processing. *Pure appl. geophys.*, 159(5): 969–987, apr 2002a.

T. Kværna, F. Ringdal, J. Schweitzer, and L. Taylor. Optimized seismic threshold monitoring - part 2: Teleseismic processing. *Pure appl. geophys.*, 159(5): 989–1004, apr 2002b.

R. J. A. Little and D. Rubin. *Statistical Analysis with Missing Data.* Wiley New York, 1987.

D. Marsan. Triggering of seismicity at short timescales following Californian earthquakes. *J. Geophy. Res.*, 108(B5):2266, 2003. doi: 10.1029/2002JB001946.

C. F. Richter. An instrumental earthquake magnitude scale. *Bull. Seismol. Soc. Am.*, 25:1–32, 1935.

C. F. Richter. *Elementary Seismology.* W. H. Freeman and Co., San Francisco, California, 1958.

P. A. Rydelek and I. S. Sacks. Testing the completeness of earthquake catalogues and the hypothesis of self-similarity. *Nature*, 337:251–253, jan 1989.

D. Shanker and M. L. Sharma. Estimation of seismic hazard parameters for the Himalayas and its vicinity from complete data files. *Pure appl. geophys.*, 152:267–279, 1998.

T. Utsu. A method for determining the value of b in a formula $\log n = a - bm$ showing the magnitude frequency for earthquakes. *Geophys. Bull. Hokkaido Univ.*, 13:99–103, 1965.

T. Utsu. A statistical significance test of the difference in b-value betweent wo earthquake groups. *J. Phys. Earth*, 14:37–40, 1966.

D. H. Weichert. Estimation of the earthquake recurrence parameters for unequal observation periods for different magnitudes. *Bull. Seismol. Soc. Am.*, 70(4): 1337–1346, 1980.

S. Wiemer and K. Katsumata. Spatial variability of seismicity parameters in aftershock zones. *J. Geophy. Res.*, 104:135–151, 1999.

S. Wiemer and M. Wyss. Minimum magnitude of complete reporting in earthquake catalogs: examples from Alaska, the Western United States, and Japan. *Bull. Seismol. Soc. Am.*, 90:859–869, 2000.

J. Woessner and S. Wiemer. Assessing the quality of earthquake catalogs: Estimating the magnitude of completeness and its uncertainties. *Bull. Seismol. Soc. Am.*, 95(2), April 2005. doi: 10.1785/0120040007.

J. Woessner, E. Hauksson, S. Wiemer, and S. Neukomm. The 1997 Kagoshima (Japan) earthquake doublet: A quantitative analysis of aftershock rate changes. *Geophys. Res. Let.*, 31:L03605, 2004. doi: 10.1029/2003GL018858.