Diss. ETH No.

Automatic reconstruction of fault networks from seismicity catalogs including location uncertainty

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Sciences

presented by

Yaming Wang

Master of Opto-electrical Engineering,
Beijing University of Aeronautics and Astronautics (China)
born October 06, 1981
citizen of China

accepted on the recommendation of
Prof. Dr. Didier Sornette, examiner
Dr. Stephan Husen, co-examiner
Dr. Jochen Woessner, co-examiner
Dr. Guy Ouillon, co-examiner
Prof. Dr. David Marsan, co-examiner

2013

Contents

Résumé

Dans le cadre de la tectonique des plaques, la déformation induite par le mouvement relatif de deux plaques se produit le long de discontinuités de déplacement dans la croûte terrestre appelées zones de failles. Les zones de failles actives ont un lien causal direct avec les tremblements de terre, qui relâchent soudainement les contraintes tectoniques dans un intervalle de temps très court. Réciproquement, les zones de failles grandissent lentement par accumulation de déplacement dû aux séismes, par endommagement croissant à leurs extrémités, ainsi que par des processus de branchement ou de connexion entre failles préexistantes de diverses tailles. Dans les dernières décennies, la connaissance de la phénoménologie et de la mécanique des failles et séismes individuels a énormément progressé, mais il manque encore une compréhension profonde de leurs liens et interactions. Un des principaux problèmes est notre incapacité à attribuer avec certitude un séisme donné à sa faille causale. Utilisant une approche de reconnaissance de forme, mon but est d'examiner la relation entre séismes et failles en développant une méthode de reconstruction automatique d'un réseau de failles, en utilisant des catalogues de données de haute résolution à des échelles très différentes et tenant compte des incertitudes de localisation propres à chaque événement.

Dans cette thèse, j'introduiso une méthode, baptisée *Anisotropic Clustering of Location Uncertainty Distributions (ACLUD)*, afin de reconstruire les réseaux de failles actifs en utilisant les localisations de séismes et leurs incertitudes individuelles. Cette méthode consiste à ajuster un ensemble donné d'hypocentres avec un nombre croissant de segments de plans jusqu'à ce que l'écart résiduel soit comparable aux incertitudes de localisation. Après une recherche massive dans l'espace des solutions possibles, j'applique six procédures différentes de validation afin de sélectionner le meilleur réseau correspondant. Deux des étapes de validation (par validation croisée et critère d'information Bayésien (BIC)) traitent les résidus de l'ajustement, alors que les quatre autres cherchent les solutions les plus en adéquation avec les mécanismes au foyer observés indépendamment. Les méthodes d'ajustement et de validation sont testées avec succès sur des exemples synthétiques. La méthode ACLUD fournit des solutions proches de celles attendues, spécialement dans le cas de la validation par BIC ou par les mécanismes au foyer. Dans le cas de la présence d'un taux significatif de sismicité non corrélée, de bonnes solutions sont obtenues en utilisant une validation basée sur les mécanismes au foyer.

Cette nouvelle méthode de reconnaissance de forme étant capable d'intégrer la plus grande partie de l'information contenue dans les catalogues de sismicité modernes, j'évalue de quelle manière la géométrie du réseau de stations sismographiques local améliore, ou altère, la reconstruction du réseau de failles sous-jacent. Je montre cela en utilisant les données les plus fiables (selon des critères relatifs au réseau de stations), qui permettent d'obtenir une reconstruction des failles de meilleur qualité et plus précise. Utiliser des données de qualité plus médiocre peut conduire à des reconstructions instables et non fiables, en particulier dans les régions où le réseau de failles possède une structure complexe. Nos résultats mettent en lumière la nécessité d'une évaluation méticuleuse de la qualité et de la fiabilité des réseaux de failles reconstruits pour des applications sur des données réelles qui, inévitablement, impliquent l'ajustement d’ensembles de données de qualités hétérogènes. A partir de tests réalistes sur des réseaux de failles synthétiques, les résultats montrent également la nécessité de prendre en compte les structures à petite échelle qui sont mal échantillonnées par la sismicité, ainsi que l’hétérogénéité spatiale des incertitudes de localisation des événements.

J’applique cette méthode de reconstruction à deux exemples naturels concernant deux échelles spatiales très différentes : la séquence de séismes suivant le choc de Landers (1992, Californie du Sud, M=7) et la sismicité induite à Bâle (Suisse). Les deux cas, j’obtiens des réseaux de failles raisonnablement comparables à des données indépendantes de géologie structurale. Ceci suggère l’existence de structures faillées complexes dans chaque cas, à l’échelle de Landers (couvrant un volume d’environ 70,000km^3) et à celle de Bâle (pour un volume d’environ 1km^3). Cette complexité des réseaux de failles reconstruits implique que les méthodes de reconstruction automatique de réseaux de failles pourraient, dans le futur, être utilisées afin d’obtenir de meilleures prévisions de la distribution spatiale des événements au sein des séquences sismiques.

Abstract

Within the framework of plate tectonics, the deformation that arises from the relative movement of two plates occurs across discontinuities in the earth's crust, labeled as faults zones. Active fault zones are the causal locations of most earthquake, which suddenly release tectonic stresses within a very short time. In return, fault zones slowly grow by accumulating slip due to such earthquakes by cumulated damage at their tips, and by branching or linking between pre-existing faults of various sizes. Over the last decades, a large amount of knowledge has been acquired about the overall phenomenology and mechanics of individual faults and earthquakes, but a deep physical and mechanical understanding of the links and interactions between and among them is still missing. One of the main issues lies in our failure to always succeed in assigning an earthquake to its causative fault. Using approaches based in pattern recognition theory, I aim to gain more insight in the relationship between earthquakes and fault structure by developing an automatic fault network reconstruction approach using high resolution earthquake data sets at largely different scales and considering individual event uncertainties.

In this thesis, I introduce the Anisotropic Clustering of Location Uncertainty Distributions (ACLUD) method to reconstruct active fault networks on the basis of both earthquake locations and their estimated individual uncertainties. This method consists in fitting a given set of hypocenters with an increasing amount of finite planes until the residuals of the fit compare with location uncertainties. After a massive search through the large solution space of possible reconstructed fault networks, I apply six different validation procedures in order to select the corresponding best fault network. Two of the validation steps (cross-validation and Bayesian Information Criterion (BIC) process the fit residuals, while the four others look for solutions that provide the best agreement with independently observed focal mechanisms. Tests on synthetic catalogs allow us to qualify the performance of the fitting method and of the various validation procedures. The ACLUD method is able to provide solutions that are close to the expected ones, especially for the BIC and focal mechanism-based techniques. The clustering method complemented by the validation step based on focal mechanisms provides good solutions even in the presence of a significant spatial background seismicity rate.

As the new clustering method is able to deal with most of the information contained in modern earthquake catalogs, I assess how the geometry of the local station network may improve or alter the reconstruction of the underlying fault system. I illustrate this by using the highest-quality data selected by station network criteria which results in reconstructed fault planes of higher quality and accuracy. Using lower-quality data can lead to unstable and unreliable fault networks and may introduce artifacts, in particular in regions of a complex fault structure. The results highlight the need to carefully assess the quality and reliability of reconstructed fault networks from real data that unavoidably involve clustering of data of heterogeneous qualities. Based on realistic tests with synthetic fault network structures, the results also stress the importance of accounting for under-sampled sub-fault structures as well as for the spatially inhomogeneous location uncertainties.

I apply the fault reconstruction method to two real datasets at two very different spatial scales, i.e. the 1992 Landers M7

earthquake sequence in Southern California, and the Basel (Switzerland) induced seismicity sequence. In both case studies, I find reasonable fault network results compared to independent structural analysis data, suggesting highly complex fault structures on both, at the scale of the Landers earthquake covering a volume of about 70,000km^3 and in the volume of the Basel induced seismicity sequence contained in a 1km^3 cube. This complexity of reconstructed fault network implies that the application of automatic network reconstruction methods may be added in the future to better forecast the spatial distribution of earthquakes within such sequences.

Chapter 1

Introduction

Within the framework of plate tectonics, the deformation that arises from the relative movement of two plates occurs across displacement discontinuities labeled as faults. Different style of faulting develop as a function of boundary conditions: divergence is dominated by normal faults (the most spectacular examples being the mid-ocean ridges of the Baïkal Lake in a continental setting); convergence is dominated by thrust faults (as in subduction zone or within the India-Asian collision zone); strike-slip setting is dominated mostly by vertical fault with slip vector along the horizontal plane (well-known and studied examples are the San Andreas fault system, California, USA, and the North Anatolian Fault system, Turkey). Tectonic deformation encompasses a very wide spectrum of scales, both in spatial and temporal dimensions. Fault-like structures are observed at scales ranging from a few centimeters (in the field or laboratory experiments), to hundreds of meters (as for example within mines or domains hosting induced seismicity experiments), to several hundreds of kilometers. Active fault zones are the causal locations of most earthquakes, which stand as a brutal process of releasing tectonic stresses. In return, faults slowly grow by accumulation of slip due to such earthquakes (which also induce damage processes such as gouge formation within the fault zone), by cumulated damage at their tips, and by branching or linking between pre-existing faults of various sizes. Over the last decades, a large amount of knowledge has been acquired about the overall phenomenology and mechanics of individual faults and earthquakes (Passchier and Trouw 2005; Scholz 2002; Stirling et al. 1996), but a deep physical and mechanical understanding of the links and interactions between and among them is still missing.

One of the main issues lies in our failure to always succeed in assigning a given earthquakes to its causative fault. A recent effort for such an assignment in the San Francisco Bay area showed significant discrepancies that arose from the simplified geometry of fault zones at depth and the amount and direction of systematic biases in the calculation of earthquake hypocenters (Wesson et al. 2003). Plotting the fraction of earthquakes inside a swath around each of the fault segments in the Southern California Community Fault Model as a function of distance and magnitude suggests that events with larger magnitudes tend to occur closer to the mapped fault. This observation may however be tempered by the fact that hypocenter locations of events with smaller magnitudes are less well constrained due to the lower number of stations detecting them. In the other hand, smaller events could also occur on subsidiary or buried faults, which are not accounted for by assuming simple large scale fault system geometries.

The detection of linear and planar structures in seismotectonics has a long history. From the early years of instrumental seismology, the main method to identify faults from earthquakes has been simply visual inspection. The geometry of an active fault zone is often constrained by mapping the surface trace; the dip angle at depth and depth extent are either constrained by results of controlled source seismology (if available), the distribution of hypocenter locations, or they are just extrapolated using geometrical constraints, if seismological ones are not available. For example, one of the most sophisticated fault models available, the Community Fault Model (CFM) of the Southern California Earthquake Center (SCEC), combines all available information on observed surface traces, seismicity, seismic reflection profiles, borehole data, and other subsurface imaging techniques to provide three-dimensional representations of major strike-slip, blind-thrust, and oblique-reverse faults of southern California (Plesch et al. 2007). Each fault is then represented by a triangulated surface in a precise geographic reference frame. However, the representation of a fault by such a simple surface cannot reflect the fine-detailed structure observed in the field within extinct fault zones or in drilling experiments across active faults (Faulkner et al. 2003; Scholz 2002). These results suggest that fault zones actually consist of narrow earthquake-generating cores, possibly accompanied by a complex set of small subsidiary faults.

Very few efforts have been devoted to the automatic digital detection of linear or planar spatial features in earthquake catalogs. Ouillon et al. (2008) introduced the three-dimensional optimal anisotropic dynamic clustering method (OADC) in order to quantitatively estimate the geometrical properties of brittle structures incorporating the uncertainties of the earthquake locations. In a nutshell, OADC is an iterative method that progressively fits a hypocenter data set by introducing an increasing number of finite planes whose positions, sizes and orientations are optimized by minimizing event-to-fault distances. It logically stops when the standard deviation of the events' location across each associated plane is smaller than the assumed location uncertainty. The way the algorithm adds new planes in the system follows a stochastic scheme so that different solutions for the same dataset are obtained for different runs. This ensures to explore more or less randomly the solution space.

Despite providing encouraging results when applied to the Landers 1992 earthquake sequence (Ouillon et al. 2008), or more recently to the Shoreline Fault, Central California (Hardebeck 2013), the main flaws of OADC are:

1. Earthquake location uncertainties are assumed to be isotropic and identical for all events. This value in return totally controls the overall resolution below which the fitting process is stopped.
2. There is no automated validation procedure of the obtained solutions. Ouillon et al. (2008) simply choose the most frequent solution and notice that it is also the most similar to the independent fault traces maps provided by the CFM in the Landers area. It ignores other prior available information, such as the focal

mechanisms of the events that stand as natural candidates to validate or reject a solution.

However, Bondár et al. (2004) show that earthquake location uncertainty in the direction of focal depth is generally several times larger than in the epicentral plane. Moreover, even within a single catalog, location uncertainties vary significantly with space and time due to differences in the station coverage, phase picking quality, velocity model quality, and so on. The strong assumption underpinning OADC can thus hardly be met in real catalogs.

We shall then first introduce a new fault network reconstruction method (see Chapter 2) that improves OADC by incorporating the full set of uncertainties as given by the PDF of the location problem. The new method will consist of two main steps:

1) A training phase, which performs the fit of a given data set or subset;

2) A validation phase, which quantifies the ability of the fit solution to explain another set or subset of independent data.

In the training phase, the new method will take account of the detailed and individual location uncertainties of each event, which control both the fitting process, through the use of the Expected Square Distance (ESD) between an event and a fault, and the space-varying resolution at which the fit stops. As the training process is strongly nonlinear, so that different runs generally converge towards different local minima of the residuals, the new method will dynamically generate many different solutions. In order to find the "optimal" one, we shall submit them to a series of different validation processes, each coming with its own specific criterion: two of them are based on the residuals of the fit, and four others are based on the compatibility of the fault networks with known focal mechanisms. The latter consist in checking the agreement between the reconstructed fault planes and the observed potential failure planes deduced from double-couple source solutions. Numerous synthetic examples will be presented in this chapter as well as in the related Appendix.

As the new clustering method will be able to deal with most of the information contained within modern earthquake catalogs, we may question, in return, the influence of the quality of the catalogs on the performance of the clustering scheme, and on the significance of its outputs. Following Bondár et al. (2004) who showed that the seismic network geometry information can be used to qualify earthquake epicenter location accuracy, in Chapter 3, we will revisit seismic network criteria to assess earthquake location quality for local networks based on a nonlinear earthquake location scheme. We shall then get one step further by assessing how the geometry of the local station network may improve or alter the reconstruction of the underlying fault system. Those investigations will be presented in Chapter 3 and will outline the rules that one should follow to select a high quality dataset. An application to the case of the Landers area will demonstrate the power of such an approach on a synthetic example with a realistic station network geometry.

This integrated approach will also be applied to two real datasets concerning two very different spatial scales. The first one (see Chapter 2) is part of the sequence that followed the 1992 Landers M7 earthquake in Southern California. It appears as a natural candidate, as a lot of literature has been published about it and it can allow for comparison with the work of Ouillon et al (2008). The second one (see Chapter 4) concerns a much smaller scale of a 1km^3 volume featuring seismic events induced by fluid injection during the Basel Enhanced Geothermal System (EGS) Project. It will allow us to check if the fault network complexity we observe in our reconstruction strongly depends on scale or not.

Chapter 2

Automatic reconstruction of fault networks from seismicity catalogs including location uncertainty

Y. Wang, G. Ouillon, J. Woessner, D. Sornette, S. Husen

Currently under revision by the Journal of Geophysics Research

2.1 Abstract

We introduce the Anisotropic Clustering of Location Uncertainty Distributions (ACLUD) method to reconstruct active fault networks on the basis of both earthquake locations and their estimated individual uncertainties. After a massive search through the large solution space of possible reconstructed fault networks, we apply six different validation procedures in order to select the corresponding best fault network. Two of the validation steps (cross-validation and Bayesian Information Criterion (BIC) process the fit residuals, while the four others look for solutions that provide the best agreement with independently observed focal mechanisms. Tests on synthetic catalogs allow us to qualify the performance of the fitting method and of the various validation procedures. The ACLUD method is able to provide solutions that are close to the expected ones, especially for the BIC and focal mechanism-based techniques. The clustering method complemented by the validation step based on focal mechanisms provides good solutions even in the presence of a significant spatial background seismicity rate. Our new fault reconstruction method is then applied to the Landers area in Southern California and compared with previous clustering methods. The results stress the importance of taking into account undersampled sub-fault structures as well as of the spatially inhomogeneous location uncertainties.

2.2 Introduction

Earthquake forecasts should ultimately be founded on the premise that seismicity and faulting are intimately interwoven:

earthquakes occur on faults and faults grow and organize in complex networks through accumulation of earthquakes. The obvious character and the power of this well-established fact are obfuscated by serious difficulties in exploiting it for a better science of earthquakes and their prediction. Indeed, an intrinsic limitation of present efforts to forecast earthquakes lies in the fact that only a limited part of the full fault network has been revealed, notwithstanding the best efforts combining geological, geodetic and geophysical methods (see Mace and Keranen (2012), for instance) together with past seismicity to illuminate fault structures (Plesch et al. 2007; Basili R. et al. 2013). Nevertheless, these studies suggest that fault networks display multiscaling hierarchical properties (Cowie et al. 1995), which are intimately associated with the modes of tensorial deformations accommodating large scale tectonic driving forces (Sornette 1991; Sornette and Virieux 1992). Neglecting the information from fault networks constitutes a major gap in the understanding of the spatial-temporal organization of earthquakes (see however early attempts by Cowie et al. (1995); Cowie et al. (1993); and Sornette et al. (1994)), thus limiting the quality and efficiency of most current earthquake forecasting methods. Including more realistic geometries and tensorial strain information associated with the underlying reconstructed fault networks will in the long-term improve present attempts to develop better space-time models of earthquake triggering, which still lack information on fault localization by assuming diffuse seismicity unrelated to faults or assume very simplified structures (Woessner et al. 2010; Ogata and Zhuang 2006; Gerstenberger et al. 2005). A reliable association of earthquakes and faults is an important constraint to determine the spatial decay of earthquakes in aftershock sequences, which provides insights into the triggering mechanisms of earthquakes (Stein 1999) and improves estimates of where aftershock hypocenters are located in comparison to the main shock properties (Woessner et al. 2006; Hauksson 2010; Powers and Jordan 2010).

Earthquake forecasting must issue statements about the likely spatial location of upcoming events. In an ideal case, we would like to forecast the set of faults or fault segments about to break in the near future. This would help predicting the expected ground motions due to radiated seismic waves, as well as anticipating problems due to surface faulting prone to cause damage on infrastructures. This goal is addressed with current fault-based approaches that use catalogs of mapped faults such as the Community Fault Model (CFM) in Southern California (Basili R. et al. 2013; Plesch et al. 2007); which however lack the small-scale structures that may contribute significantly to short and intermediate-term hazards. Moreover, as illustrated by the M_w 6.7 Northridge, 1994 earthquake, a significant number of large earthquakes continue to occur on faults that were not yet mapped and were only revealed by the earthquake itself. In the case of Northern California, most of the seismicity remains unexplained by the set of mapped faults as shown for example in Wesson et al. (2003), where most events are labeled under 'BKGD', for 'background', whereas they seem to occur on well-defined fault structures. Moreover, such extensive fault catalogs do not necessarily exist in other parts of the world exposed to intense seismic hazard.

Using the magnitude of recorded events to determine empirically their contribution to the amount of slip over each fault patch, an improved knowledge of the underlying fault network may allow one to infer average slip rates on each fault at geological time scales and convert them into long-term average seismicity rates (Gabrielov et al. 1996), possibly considering the information given by paleoseismological studies (see for example the National Seismic Hazard Mapping Program; Frankel et al. (2002); the Uniform California Earthquake Rupture Forecast model; Petersen et al. (2007); Field (2008)). This approach is used to provide long-term time-dependent or time-independent forecasts.

The usual, and necessary, trick used in existing earthquake forecasting methods thus consists in smoothing the spatial structure of the earthquake catalog, in order to approximate the geological complexity of the local fault network. Only recently, forecast models were proposed that attempt to combine both seismicity and fault data sets in a common approach, yet blurring the knowledge of the fault structure by smoothing techniques (Hiemer et al. 2013; Rhoades and Stirling 2012). Smoothing is performed using only the 2D set of epicenters (and not the 3D set of hypocenters), and this process always involves a set of arbitrary choices or parameters. The simplest smoothing consists in superimposing a regular grid onto the target area, thus coarse-graining the fault network at a homogeneous (and arbitrary) spatial resolution. A softer method consists in smoothing the set of declustered events with Gaussian kernels, whose bandwidths are adapted to optimize the quality of smoothing according to some metric (Zechar and Jordan 2010), by using adaptive kernels or those respecting the distance to their closest neighbors (Hiemer et al. 2013). In general, events are simply replaced by kernels that are added up over the whole space and normalized so that the integral of the spatial density of events is equal to the number of events in the catalog. In many implementations, a smoothing is considered as optimal when it maximizes the score of the forecasts on an independent dataset. It follows that the smoothing parameters do not stem from independent geological or physical knowledge. They thus look more like hidden parameters of the forecasting technique as a whole. Moreover, the use of square cells or isotropic kernels is totally opposite to what could be expected to best approximate a set of plane segments, whose orientations vary in space (see for example Gaillot et al. (2002) and Courjault-Radé et al. (2009), for the spatial analysis of sets of epicenters using anisotropic wavelets, inspired by a methodology initially developed by Ouillon et al. (1996); Ouillon et al. (1995) for maps of fault or joint traces). In some cases, the bandwidth of the kernels may also depend on the size of the local events or on their spatial density: the larger the latter, the finer the resolution.

The well-documented multiscale organization of earthquakes and faults precludes any objective choice of the most appropriate spatial resolution to study their dynamics. The only characteristic scales in such systems are the size of the system itself (at large scales), and the scale below which scale invariance breaks down without producing bonus information; typically, this is the smallest distance between pairs of events, or the size of the smallest fault, or the width of geological and rheologically different layers (Ouillon et al. 1996). From a statistical physics point of view, one may argue that taking account of the numerous 'microscopic' spatial details of the seismicity process may only deteriorate our ability to model their dynamics and provide efficient forecasts, which is then a good reason to perform a smoothing. Another obvious reason is that events are always spatially located up to some finite uncertainties. However, Werner et al. (2011) brought into the debate new interesting elements by noticing that accounting for small magnitude earthquakes (down to M = 2) in the input data set increased the likelihood of the forecasts. As increasing the number of small-scale earthquakes allows one to take account of smaller-scale details of the fault network, it follows that the smoothed seismicity rate of Werner et al. (2011) closely reflects the best possible approximation of the fault network they could hope to get. This result echoes the conclusion of Zechar and Jordan (2008) and (Woessner et al. 2011) who suggest that future seismicity-based techniques should also use this set of faults as a data input.

No independent and accurate geophysical technique exists that provides a detailed and complete 3D map of active fault networks. As a consequence, we rely in this approach on seismicity itself as the best proxy to image the current fault network. Continuous and recent progress in earthquake location techniques now allow the manipulation of rather precise spatial data. For example, as absolute locations used to feature uncertainties of the order of a few kilometers in Southern California are now re-estimated using relative location algorithms, the (relative) uncertainties are now shrinking down to only a few tens of meters (Waldhauser and Schaff 2008; Hauksson et al. 2012). Nonlinear location algorithms (Lomax et al. 2009; Husen et al. 2007; Husen et al. 2003) even allow the direct sampling of the full probability density function (hereafter pdf) of the location of each event. It follows that seismologists now have the opportunity to access to the detailed topology of the active part of the fault network, provided they have the tools to estimate the position, size and orientation of fault segments from the precise location of events listed in earthquake catalogs, i.e. to extract the full value from these golden data.

Ouillon et al. (2008) recently proposed a new method of pattern recognition that reconstructs the active part of a fault network from the spatial location of earthquake hypocenters. It is inspired from the seminal k-means method (MacQueen 1967), which partitions a given dataset into a set of (*a priori* isotropic) clusters by minimizing the global variance of the partition. Ouillon et al. (2008) generalized this method to the anisotropic case with a new algorithm, which, in a nutshell, fits the spatial structure of the set of events with a set of finite-size plane segments. The number of segments used is increased until the residuals of the fit become comparable to the average hypocenters location uncertainty. One can then estimate the position, size and orientation of each plane segment. Ouillon et al. (2008) applied this algorithm to synthetic datasets as well as to the aftershock cloud of the Landers, 1992 event, in Southern California, for which they showed that 16 planes were necessary to provide a fit compatible with the average location errors. Moreover, extrapolating the set of plane segments to the free surface, the predicted fault traces showed a good agreement with observed fault traces of the Southern California Community Fault Model (CFM) and also allowed to map faults of significant size that are not reported in the CFM.

The main shortcoming of the Ouillon et al. (2008) clustering method is its rough account of location uncertainties, assumed to be constant for the whole catalog. In this paper, we improve on this method by taking account of the detailed and individual location uncertainties of each event, which control both the fit through the use of the Expected Squared Distance (ESD) between an event and a plane and the resolution at which the latter is performed. As the fitting method is still strongly nonlinear, different runs generally converge towards different local minima of the residuals. We thus introduce new methodologies to validate the obtained solutions, as systematic and automatic comparison with existing fault maps, if existent, is a very difficult exercise, in particular because it lacks a precise metric. We thus present six validation schemes: two of them based on the residuals of the fit, and four others based on the compatibility of the fault networks with known focal mechanisms. The new method is then tested on simple and more complex synthetic fault networks, as well as on a new catalog of the Landers area.

2.3 The optimal anisotropic data clustering (OADC) method

The new clustering method proposed here is based on a pattern recognition technique called k-means, shortly described in Ouillon et al. (2008) and in more details in Bishop (2006), Duda et al. (2001) and MacQueen (1967). This technique makes no assumption about the shape of the individual clusters. In that sense, it can be viewed as an 'isotropic' processing of data. When dealing with earthquakes, it is desirable to cluster data within structures that can be identified as faults. In that case, the minimum *a priori* information that may help to constrain the pattern recognition process is that the clusters we look for should be highly anisotropic, i.e. that their thicknesses should be very small compared to their other dimensions.

The OADC method of Ouillon et al. (2008) provides an attempt to reconstruct fault networks using solely the information contained within seismicity catalogs. Compared with other strategies, e.g. the Community Fault Model (CFM) of the Southern California Earthquake Center (SCEC), it defines a general method that can identify active fault segments without taking into account direct observations such as maps of fault surface traces and/or subsurface borehole data, nor indirect observations like seismic reflection profiles to map deeper structures. Ouillon et al. (2008) also provide a discussion of other seismicity clustering techniques.

The OADC method is directly inspired from the original definition of the k-means method, yet generalizes it to strongly anisotropic clusters, whose thicknesses are assumed to be very small. Each fault segment is thus approximated by a finite rectangular plane, characterized by its dimension (length and width), orientation (strike and dip) and position of its center. Earthquakes are handled as pure data points, while a uniform and isotropic location uncertainty ε is assumed to hold for all events.

The general algorithm of the method is the following:

1. Initialize N_0 planes with randomly chosen center positions, orientations and dimensions.

2. For each earthquake *O* in the catalog, compute the distance from it to each plane *C*, determine the closest plane, and associate the former to the latter. Earthquake locations are treated as points, and Euclidean distances to the finite planes are computed. This first partition provides us a set of N_0 clusters of events.
3. For each cluster, perform a spatial principal component analysis (PCA), and use the eigenvalues and

eigenvectors to define their new dimensions, orientations, and center positions. The thickness of each cluster is given by the square root of the smallest eigenvalue. The two other eigenvalues provide the length and width of the cluster (see Ouillon et al. (2008) for details).

4. Assuming a uniform catalog spatial location uncertainty ε, the computation stops if the thickness of each cluster is smaller than ε, as the dispersion of events across each plane can be fully explained by location errors. If there is at least one cluster for which the thickness is larger than ε, then proceed to step 5.

5. Split randomly the plane associated to the thickest cluster into m sub-planes, increase N_0 accordingly by m-1, and go back to step 2.

This procedure, which is nothing but a nonlinear fitting technique, ensures that events will be partitioned into clusters with negligible thickness (up to location uncertainties), i.e. plane-like structures, which are the assumed *a priori* model for faults.

Similarly to the classical k-means method, the OADC method may converge to a local minimum of the global clusters fit residual. One can solve this problem by running the clustering procedure several times, with different initial conditions, in order to explore the solution space and select the fault network model that achieves a genuine global minimum. However, as the method itself ensures that all fit residuals are smaller than location uncertainties, all solutions are therefore statistically equivalent. Picking one of them as the best one thus requires an independent validation process. Due to computational limitations, Ouillon et al. (2008) provided only ten runs on the Landers aftershocks dataset, yet noticed that the method converged more often to one of the solutions than to any other (thus suggesting a validation based on the most frequently selected solution). For each solution, extrapolating all the planes they obtained to the free surface, thus generating the corresponding predicted surface fault traces, they noticed that the most frequent solution was also the one that fits best the observed natural fault traces in this area. While offering a validation procedure on an independent dataset, this approach would prove cumbersome when dealing with much larger areas, or with zones where no such fault traces maps or incomplete ones are available. Another drawback is the subjectivity of the comparison, which is not based on any quantitative metric. The systematic validation of the obtained solutions is thus still an open problem.

Another obvious limitation of the OADC method is the assumption made about location uncertainties, which are considered to be uniform and isotropic. This hypothesis is unrealistic since focal depth is often less well constrained than the epicentral location. Moreover, location uncertainty is strongly influenced by the velocity model error, the quality of waveform pickings, the station network geometry, etc., and is thus very heterogeneous in space and time (e.g. Husen and Hardebeck (2010)). It thus follows that the clustering process should be more detailed in some areas and sparser in some others. The clustering method should take this heterogeneity into account.

2.4 Anisotropic clustering of location uncertainty distributions (ACLUD)

The original k-means method assumes that the uncertainty of the spatial location of data points is negligible. In the case of real physical systems, the story is different. For earthquakes, location uncertainty is an inherent property due to wave arrival time inaccuracy, velocity model errors, station network geometry, or outdated data sources like historical seismicity catalogs. When taking uncertainty into account, data can no longer be described as a point-process, but as a more or less complex probability density function (hereafter pdf). *Chau et al. (2006)* claim that uncertainties can significantly affect the results provided by clustering techniques such as k-means. They thus introduce the *uk-means* algorithm (where ‘u’ stands for ‘uncertain’, see electronic supplement), which incorporates uncertainty information and provides, when considering synthetic samples, more satisfying results than the standard algorithm.

We now show how to extend the uk-means method of Chau et al. (2006) to the case where the cluster model $\vec{C}$ is a plane, in the spirit of Ouillon et al. (2008), and the object to cluster $\vec{O}$ is the pdf of an earthquake location. We term the new method the “anisotropic clustering of location uncertainty distributions” (ACLUD).

Chau et al. (2006) suggest using the expected squared distance (hereafter ESD), which, in our case, is defined as:

$$d^2(\vec{O},\vec{C}) = \int_{\vec{x}\in\vec{O}} \left\|\vec{x}-\vec{C}\right\|^2 f(\vec{x})d\vec{x} = \int_{\vec{x}\in\vec{O}} \inf_{\vec{c}\in\vec{C}} \left\|\vec{x}-\vec{c}\right\|^2 f(\vec{x})d\vec{x} \qquad (2.1)$$

where f(x) is the pdf of the earthquake location. While this distance is easily estimated in the case of an infinite plane $\vec{C}$, we also propose computationally efficient approximations in the case of a finite-size plane.

2.4.1 Expected square distance (ESD) between a probability density function and an infinite plane

We consider an infinite plane within a Euclidean three-dimensional space. The coordinate system is chosen such that its origin is located on the plane, whose orientation is given by two of the basis vectors, the third one being normal to it. Then, Eq. (2.1) can be rewritten as:

$$d^2(\vec{O},\vec{C}) = \int_{\vec{x}\in\vec{O}} \left\|x_3\right\|^2 f(\vec{x})d\vec{x} \qquad (2.2)$$

where x_3 is the third component of point $\vec{x} \in \vec{O}$. Noticing that:

$$x_3^{\ 2} = \left(x_3 - k_3 + k_3\right)^2 = \left(x_3 - k_3\right)^2 + k_3^{\ 2} + 2\left(x_3 - k_3\right) \cdot k_3 \quad (2.3)$$

with k_3 being the third component of the centroid of $\vec{O}$, and given that the contribution of the last right-hand term of Eq. (2.3) to the integral is zero, Eq. (2.2) becomes:

$$d^2(\vec{O},\vec{C}) = k_3^{\ 2} + \int_{\vec{x}\in\vec{O}} \left(x_3 - k_3\right)^2 f(\vec{x})d\vec{x} \qquad (2.4)$$

The first term in the right-hand side is simply the squared distance between the centroid of $\vec{O}$ and the infinite plane, while the second term is simply the variance of $\vec{O}$ in the direction normal to the plane (which can be deduced from the pdf of $\vec{O}$ and its covariance matrix). This is nothing but the variance decomposition theorem.

2.4.2 Expected square distance (ESD) between a probability density function and an infinite line or a point

Following a similar procedure when $\vec{C}$ is a line, we can choose a coordinate system so that $\vec{C}$ lies on the first axis. Then we get:

$$\begin{aligned} d^2(\vec{O},\vec{C}) &= k_2^{\ 2} + k_3^{\ 2} + \int_{\vec{x}\in\vec{O}} \left[\left(x_2 - k_2\right)^2 + \left(x_3 - k_3\right)^2\right] f(\vec{x})d\vec{x} \\ &= \sum_{i=2}^{3} k_i^{\ 2} + \sum_{i=2}^{3} \int_{\vec{x}\in\vec{O}} \left(x_i - k_i\right)^2 f(\vec{x})d\vec{x} \qquad (2.5) \end{aligned}$$

When $\vec{C}$ is a point, we can choose a coordinate system so that $\vec{C}$ lies at the origin. Then we get:

$$d^2(\vec{O},\vec{C}) = \sum_{i=1}^{3} k_i^{\ 2} + \sum_{i=1}^{3} \int_{\vec{x}\in\vec{O}} \left(x_i - k_i\right)^2 f(\vec{x})d\vec{x} \qquad (2.6)$$

The interpretation of Eq. (2.5) and (2.6) is the same as for Eq. (2.4) except that we now compute the distance between the centroid and a line and use the relevant dimension for the variance decomposition. This last set of equations will prove very useful when approximating the distance between a pdf and a finite plane.

2.4.3 Expected square distance (ESD) between a probability density function (pdf) and a finite plane

The anisotropic clustering of location uncertainty distributions (ACLUD) method we propose still assumes that active fault segments can be modeled as rectangular finite planes. If it proves rather easy to compute the Euclidean distance between a point and a finite plane, the problem is a bit more difficult when observations are given through their pdf's. Indeed, we shall see that, using the variance decomposition theorem, we can only provide theoretical approximations to the expected squared distance between a pdf and a finite plane.

Figure 2.1 illustrates the problem. The grey rectangle area represents a finite plane $\vec{C}$. Events may be located anywhere in the full 3D space that surrounds it. We now consider any object $\vec{O}$ in the 3D space and its projection Q along the direction normal to $\vec{C}$ onto the infinite plane containing $\vec{C}$. The object Q will be located within one or more of the nine sectors defined in Figure 2.1, each sector being indexed in roman numbers from I to III as shown in the figure. The object Q can overlap several sectors depending on the shape of the support of its pdf.

If Q is completely included within sector III, then the ESD between $\vec{O}$ and $\vec{C}$ can be computed using Eq. (2.4), as the infinite plane assumption is valid. If Q is completely included within a sector labeled (II), the ESD is computed using Eq. (2.5) (after an appropriate change of coordinates) as the infinite line assumption is valid. If Q is completely included within a sector labeled (I), the ESD should be computed between the pdf and the closest corner of the finite plane, using Eq. (2.6). Indeed, a similar approach has been used in Ouillon et al. (2008) to compute the Euclidean distance between a given hypocenter and a given finite plane.

In our case, the general problem is much more complex as we implicitly have to consider the distance between the finite plane and every point where the pdf of $\vec{O}$ is defined. This implies that the projection $\underline{Q}$ is characterized by a pdf that may overlap several distinct sectors, so that none of the above simple formulae (2.4), (2.5) and (2.6) can be used anymore. In that case, only a direct Monte Carlo approach provides an accurate estimate of the ESD. As it would prove computationally too heavy when handling large catalogs and sets of faults, we propose a simplification: we first consider only the centroid of $\vec{O}$ and its own projection. If the latter is contained within sector III, we use formula (2.4) as an approximation to the ESD. If it is contained within a sector labeled (II), we use formula (2.5). If it is contained within a sector labeled (I), we use formula (2.6). This approximation is obviously wrong when the size of the finite plane is much smaller than the spatial extent of the domain where the pdf of $\underline{Q}$ is defined. However, in practice we found that for most of cases, location uncertainties are much smaller than the size of potential fitting fault plane we can resolve.

2.4.4 Anisotropic clustering of location uncertainty distributions algorithm

Assume that an earthquake catalog provides the location of each event with a pdf. We can characterize the location with its centroid (hereafter, the hypocenter) and its covariance matrix. The new clustering algorithm we propose is the following:

1. Split randomly the earthquake catalog into 2 distinct subsets: the training set (which is the one to be fitted) and the validation set (which is the one used to qualify or discriminate different clustering models).
2. Initialize a number of N_0 faults with random positions, orientations and dimensions.
3. For each earthquake in the training subset, associate the earthquake to the closest plane according to the ESD. We thus get a partition of events into a set of N_0 clusters.

4. For each cluster *i*, compute the covariance matrix of the locations of its associated hypocenters, and find its eigenvalues and eigenvectors. By doing so, the dimensions and orientations of each cluster can be computed. The smallest eigenvalue $\lambda_{i,3}$ provides the thickness of the corresponding cluster.
5. For each cluster, compute the average individual variance ε_i of the hypocenters' location pdf in the direction normal to the cluster.
6. For each cluster, compare its thickness $\lambda_{i,3}$ with the average location uncertainty ε_i of its associated events. If $\varepsilon_i \geq \lambda_{i,3}$ for all clusters, the computation stops, as location errors alone can explain the finite thickness of each cluster. We then proceed to step 8. If there is at least one cluster for which $\varepsilon_i < \lambda_{i,3}$, then we proceed to step 7 as we need more planes to explain the data.
7. We split randomly the thickest cluster into m other planes, and go back to step 3 (increasing N_0 accordingly by m-1).
8. We compute the residual of the fit of the validation data set conditioned on the fault network model of the training data set (from step 6).
9. We repeat steps 1-8 many times (typically several thousands) and rank all models according to their validation fit residuals obtained in step 8.

For this study, in step 7 we use m=2. The proposed algorithm accounts for individual event location uncertainties, both in the computation of the ESD between an event and the planes and in the criterion used to continue or stop the fitting process. The stopping criterion thus doesn't assume a spatially uniform location uncertainty, but is adapted to the case of space-dependent location quality. This property is particularly welcome in the case of earthquakes for which location uncertainties heavily depend on the spatial structure of the stations networks.

One should also be aware that the full three-dimensional confidence interval is different from the confidence interval in 1D. In order to compute the variance of the pdf in the direction normal to the plane, we have to project the 68% three-dimensional confidence ellipsoid onto that normal direction. Yet, after the projection, the confidence level increases to higher levels so that the correct quantiles have to be estimated (Press et al. (2007) , page 811, figure 15.6.3).

By subdividing the data set, we implement a cross-validation technique to the predictive skill of the clustering approach. Our procedure separates randomly the full dataset into two independent subsets, generates the fault model that fits the training dataset and evaluates it by estimating how well it predicts the independent validation set. The process is repeated several times, each trial corresponding to different training and validation sets, and we select the one with the best validation result. How to generate the training and validation data sets is a question in itself. On the one hand, if there are not enough earthquakes in the training set, it will lead to a spurious fit with a very bad validation score; on the other hand, if there are not enough earthquakes in the validation set, residuals may fluctuate and depend strongly on the particular choice of the validation set. Using 95% of the data as the training set and 5% as the validation set are standard values used in pattern recognition algorithms (Bishop 2006). Yet, from synthetic tests where the original fault networks are known, we checked that it generally provides robust results.

The main assumption of this algorithm is that the hypocenter corresponds to the expectation hypocenter location (Lomax et al. 2000). In the framework of probabilistic earthquake location the hypocenter location is usually associated with the maximum likelihood point (Tarantola and Valette 1982). The assumption that the hypocenter is not very different from the maximum likelihood point would be valid if and only if the pdf of the location of the event is compact, i.e. small in size, which has no *a priori* reason to be true. We shall discuss later the conditions for which this assumption might be approximately valid in the case of natural earthquakes catalogs.

2.4.5 Validation strategies

The new clustering method automatically explores a very large solution space. In order to find the "best" solution, we follow a purely statistical strategy, i.e. cross validation. However, other validation strategies might be more appropriate. In the following, we will introduce three other criteria: one residuals-based statistical strategy called Bayesian Information Criterion (BIC, see Schwarz (1978)), and two metrics based on observed focal mechanisms.

2.4.5.1 Bayesian Information Criterion

BIC is a commonly used statistical criterion for model selection that takes both the likelihood function and model complexity into consideration. During clustering, it is possible to increase the likelihood by adding more faults, at the cost of increasing the complexity of the model. By adding a penalty term for the number of faults, the BIC merges the likelihood and complexity of the solution together. Assuming that the distribution of earthquakes across the fitting planes is a normal distribution, the BIC can be expressed as:

$$BIC = n \cdot ln(\hat{\sigma}) + k \cdot ln(n) \tag{2.7}$$

where n is the number of events used for the fit, $\hat{\sigma}$ is the unbiased variance estimation of the earthquake distribution across the fitting planes, and k is the number of faults in the tested model. Thus, by minimizing the BIC, we may find the best network from the solution space that provides both a large likelihood and a simple model structure. The difference with the cross validation scheme is that the latter is performed using the validation dataset, whereas the BIC uses the training dataset. It is also important to notice that, during the clustering process, we randomly partition the whole data set into training and validation sets. It means that, for each clustering run, the training set changes so that the computed BIC is not strictly derived from the same training set. However, considering that we deal with large datasets among which 95% of each single one is used as training sets, the BIC remains a robust estimator.

2.4.5.2 Focal mechanism μ -metrics

The focal mechanism of an earthquake describes the potential orientations of the rupture plane and slip vector. If events are clustered together on a given fault plane, we may expect them to be characterized by similar focal mechanisms, the latter being also consistent with the orientation of the fitting plane. This provides a mechanical approach to validation. At the end of each fit, we thus adopt the following procedure:

1. For each cluster, select the available focal mechanisms of events.
2. For each focal mechanism, compute the normal vector to each of the two nodal planes.
3. For each nodal normal vector, compute its dot product with the vector normal to the cluster (defined as pointing upwards). If one of the dot products is negative, replace the nodal normal vector by its opposite and change the sign of the dot product.
4. From both nodal normal vectors, choose the one that maximizes the dot product.
5. Once steps (2)-(4) have been fulfilled for each event of the cluster, stack all the selected nodal normal vectors, and compute the angle μ between the resultant and the normal vector to the cluster.

Step (5) is performed after weighting each selected nodal normal vector according to the magnitude M of the corresponding event. The weight is taken as 10^{aM}. If a=0, then all events have the same weight and the measured angular discrepancy is mainly controlled by the smallest events. If a=1/2, then each event is weighted proportionally to its empirically assumed slip amount, while it is weighted by its energy or moment if we set a=3/2 (and in that case the angular discrepancy is controlled by the largest event in the cluster). As the plane segments we infer from our network reconstruction are, among other parameters, characterized by a size (i.e. an area) and a direction, it thus makes sense to compare them with the average direction of rupture events weighted by their individual rupture area. This is why we choose to set a=1. Moreover, if the local Gutenberg-Richter b-value is close to 1, each magnitude range contributes equally to the estimated angular discrepancy, yet, it is not a necessary assumption of the methodology.

We first define a weighted average normal vector to the selected nodal plane of events on fault plane F_i as:

$$\vec{V}_{E_i} = \frac{1}{\left\| \sum_{k=1}^{m(i)} \vec{v}_{E_{i,k}} \cdot 10^{a \cdot M_{i,k}} \right\|} \sum_{k=1}^{m(i)} \vec{v}_{E_{i,k}} \cdot 10^{a \cdot M_{i,k}} \qquad (2.8)$$

where:

- $m(i)$ = the number of events in fault plane F_i;
- $\vec{v}_{E_{i,-}}$ = the normal vector to the selected nodal plane of a given event on fault plane F_i;

We then define a global angular discrepancy of the full set of planes as the μ_{fault} measure. It is formally expressed as:

$$\mu_{fault} = \frac{\sum_{i=1}^{n} cos^{-1}\left|\vec{V}_{F_i} \cdot \vec{V}_{E_i}\right|}{n} \tag{2.9}$$

where:

- n = the number of fault planes;
- $\vec{V}_{F_i}$ = the normal vector to fault plane F_i;

The weighting strategy of Eq. (2.9) implies that we simply compute an average angular misfit over all faults (hence the associated subscript on the left-hand side). Similarly, we can also perform the average over all events. We obtain:

$$\mu_{event} = \frac{\sum_{i=1}^{n} m(i) cos^{-1}\left|\vec{V}_{F_i} \cdot \vec{V}_{E_i}\right|}{\sum_{i=1}^{n} m(i)} \tag{2.10}$$

Minimizing both estimators will select networks where the orientation of inverted fault planes is the closest to the average orientation of the focal mechanisms. In summary, the μ-metric measures the magnitude weighted average direction of the normal vectors of the “observed” focal mechanisms to the normal vector of the fault plane derived within the clustering approach.

2.4.5.3 Focal mechanism σ-metrics

Events grouped together by our fitting procedure may also feature roughly similar focal mechanisms, whose orientation may be different from the one of the fitting plane (see sketch in Figure 2.2 and explanations below). Following the same procedure as above from step (1) to (4), we change step (5) as:

6. Once steps (2)-(4) have been fulfilled for each event, stack all the selected nodal normal vectors, and

compute the average angle between each individual selected nodal vector and the resultant stacked vector.

The associated measures are defined σ_{fault} and σ_{event} , depending on the way they are averaged. They are similar to standard deviation in statistics, yet we compute them using the L-1 norm (and not the L-2 norm). The reason is that, in the case when the distribution of angles is not Gaussian but fatter tailed, using the L-1 norm provides results less sensitive to large outliers. Using the same notations as above, the mathematical expressions are:

$$\sigma_{fault} = \frac{\sum_{i=1}^{n} \frac{1}{m(i)} \sum_{j=1}^{m(i)} cos^{-1} \left| \vec{v}_{E_{i,j}} \cdot \vec{V}_{E_i} \right|}{n} \qquad (2.11)$$

$$\sigma_{event} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m(i)} cos^{-1} \left| \vec{v}_{E_{i,j}} \cdot \vec{V}_{E_i} \right|}{\sum_{i=1}^{n} m(i)} \qquad (2.12)$$

σ measures the angular difference from each single normal vector to the fault plane from the clustering approach and then averages, which results in a quite different metric.

Figure 2.2 shows examples of applying μ and σ measures. On each plot, the black line indicates the trend of the fault zone, while the gray lines indicate the potential orientations of shorter individual ruptures within the fault zone, all events being clustered in the same macroscopic fault zone (see Section 6 for a further discussion of the influence of the fault zone complexity on the results of clustering). When the rupture planes are quasi-colinear with the fault trend, then both μ and σ values are small (Fig. 2a). Figure 2.2b shows a series of planes for which orientations oscillate around the trend of the fault. In that case, the μ value is still small while the σ value is larger. Figure 2.2c shows the case of an *en-échelon* distribution of rupture segments, which will provide a finite and possibly large μ value and a very small σ value. The last example (Fig. 2d) shows a series of alternating conjugate rupture planes, which will be associated with large values of both μ and σ. These two measures derived from focal mechanisms can quantify the degree of agreement of the reconstructed fault network with local focal mechanisms. They provide tools in model selection with consideration of tectonic knowledge compared to pure statistical approaches such as cross-validation or BIC.

2.5 Tests of the ACLUD method on synthetic catalogs featuring location uncertainties

The previous section has introduced a new clustering scheme to automatically reconstruct fault structures from seismicity catalogs including location uncertainty information. We apply the approach to synthetic catalogs to understand its sensitivity to different structural complexities.

2.5.1 Generation of datasets

Locating earthquakes results in a posterior probability density function of an event location (Moser et al. 1992; Tarantola and Valette 1982; Wittlinger et al. 1993). The pdf may possess any arbitrary shape and may be visualized using scatter density plots, which are obtained by drawing samples from the posterior pdf with their number being proportional to the probability (Lomax et al. 2000; Husen et al. 2003). From these samples, the 68% confidence ellipsoid can be computed by a singular value decomposition of the corresponding covariance matrix, and consists in a rough approximation of the spatial uncertainty of the location estimate. The expectation hypocenter is at the center of the confidence ellipsoid, and the maximum likelihood hypocenter will always be located within the densest part of the pdf, so that both locations do not necessarily coincide.

In this section, we generate synthetic earthquake catalogs using the NonLinLoc software package (Lomax et al. (2000), Version 5.2, http://alomax.free.fr/nlloc/). Compared to traditional, linearized approaches, NonLinLoc is superior in that it computes the posterior pdf using nonlinear, global searching techniques. The general method we use to generate a synthetic earthquake catalog is the following. We first impose the geometry of the original fault network, which consists in a collection of rectangular planes with variable locations, sizes and orientations. We then assume that all earthquakes occur exactly on those planes and generate P-waves. We then randomly distribute a given number of earthquakes on those planes. For each event, we randomly choose a set of 11 stations which constitute a set of observations. For a given velocity model, theoretical travel times between the true hypocenters and a set of given stations are computed. Random perturbations are added to the arrival times mimicking the uncertainty in picking waveform onset, which allows us to proceed to the inverse problem of computing the location of the events as well as their uncertainties by using NonLinLoc.

To generate the set of associated synthetic focal mechanisms, we first assume that the rake of the slip vector on each plane is zero. For each event, the strike and dip are assumed to be identical to the ones of the input plane to which it belongs. We then add an independent Gaussian random perturbation respectively to the strike, dip and rake of the event. Those perturbed angles are then used to compute the strike and dip of the auxiliary plane, thus providing a complete focal mechanism.

Note that we did not take account of the possible errors on the velocity model, which would provide systematic errors on both locations and focal mechanisms.

The catalog of relocated hypocenter locations including their scatter density clouds is then fitted with a set of finite planes, using the ACLUD algorithm as defined in the previous section. The best solution which depends on the validation technique is then compared to the original input fault network.

As a first test, we generated a very simple synthetic dataset consisting in three vertical faults featuring 4,000 events in all (thus similar to the one studied in Ouillon et al. (2008)) and characterized by their full pdf. The new clustering technique we propose successfully reconstructed the fault network whatever the validation criterion we used (see electronic supplement). We shall now test it on a more realistic and complex case.

2.5.2 Synthetic catalog with complex geometry inspired from Ouillon et al.

This synthetic dataset outlines a more complex and realistic case. Figure 2.3a shows the structure of the reconstructed fault network in the area of the 1992 M_W 7.3 Landers earthquake by Ouillon et al. (2008). It features 13 planes with a dip larger than 45° (the three other planes, dipping less than 45°, have been removed as they certainly are spurious planes – see Ouillon et al. (2008)). The original catalog used in Ouillon et al. (2008) includes 3,103 events, which we now assume to occur randomly and uniformly on those planes. We define a virtual station network, similar to the simpler one used in the example shown in the electronic supplement, in order to compute theoretical wave travel times to 11 randomly chosen stations, and add Gaussian errors with a standard deviation of 0.1 s to simulate picking errors. Figure 2.3b shows the spatial distribution of the relocated 3,103 events. To generate the set of synthetic focal mechanisms, we add an independent Gaussian random perturbation respectively to the strike, dip and rake of each event with a standard deviation of 10°. Those perturbed angles are then used to compute the strike and dip of the auxiliary plane, thus providing a complete focal mechanism. For a 80° dipping fault, we performed a Monte Carlo simulation in order to compute the angular difference between the normal vectors of the correct and perturbed mechanisms. We found that the mean value of the angular difference is 11.5°, which is comparable with the quality Class A and B focal mechanisms computed by the HASH approach (Hardebeck and Shearer 2002). As focal mechanisms are characterized by a 3D orientation, their statistics is very different from linear variables. Our approach is thus a simplified procedure to simulate uncertainties. Kagan (2005) introduces a more rigorous way to randomly rotate double-couple focal mechanisms in 3D, which one has to use if simulating broader distributions than ours.

Note that, in Section 3.5, we defined two statistical measures derived from focal mechanisms, which can be used to evaluate each reconstructed fault network. We can also assess the individual contribution of each cluster with respect to those global measures. We then similarly define for each cluster two individual measures of focal mechanism consistency, μ_F and σ_F :

$$\mu_F = cos^{-1}\left|\vec{V}_F \cdot \vec{V}_E\right| \qquad (2.13)$$

$$\sigma_F = \frac{1}{m}\sum_{i=1}^{m} cos^{-1}\left|\vec{v}_{E_i} \cdot \vec{V}_E\right| \qquad (2.14)$$

where:

- m = number of events within fault F ;
- $\vec{V}_F$ = normal vector to the given fault plane F ;
- $\vec{V}_E$ = weighted average normal vector to the selected nodal plane of events on fault F ;
- $\vec{v}_{E_i}$ = normal vector to the selected nodal plane of a given event on fault F ;

A large μ_F value indicates that the average focal mechanism rupture plane deviates significantly from the fitted fault plane. A large σ_F value indicates a significant dispersion of the orientations of focal mechanisms within the cluster.

We performed 6,000 runs with different initial conditions of the random number generator which controls the fault splitting step, and obtained as many solutions. We now discuss the results obtained using the six validation techniques discussed in Section 3.5.

Cross validation: Figure 2.4a shows the selected reconstructed network, featuring 14 planes. One can notice that two faults in the northern end are merged into a single plane. This is due to the fact that locations quality in this region is deteriorated due to a poor station coverage at the northern end. Such a poor coverage also occurs for the southern end, where the two crossing faults are reconstructed as a set of three faults. This kind of local overfitting is often observed in such situations, and is due to the splitting step of the clustering process.

BIC: Figure 2.4b shows the selected reconstructed network, featuring 15 planes, which is different from the one selected by cross-validation. Whereas the structure is now correctly inverted in the northern part, one can observe a small fault in the middle region pointed by the arrow, whose orientation is clearly rotated clockwise compared to the original synthetic network (Figure 2.3a). The reason is that the BIC gives more weight to the fault planes featuring more events. The density of events on this fault is the smallest among all 15 faults (for which this parameter ranges from 0.6/km^2 to 5.0/km^2). The reconstruction of such low event density faults can be unstable as their weight in the global criterion is very small. We also noticed that its individual σ_F value is the largest (for which this parameter ranges from 12° to 29°), indicating that the focal mechanisms of events clustered on this fault are very scattered.

μ **metrics:** both μ_{event} and μ_{fault} metrics select the same solution, shown in Figure 2.4c, featuring 14 planes. One can observe that two faults in the northern middle region have been merged into a single one (indicated by a small arrow). The distribution of μ_F and σ_F values of all 14 planes range from 1° to 43° and 12° to 23°, respectively. The individual μ_F value and σ_F value of this merged fault are both the largest over all 14 faults. This indicates that the focal mechanisms of the events clustered on this fault are neither consistent with each other nor with the orientation of the fitting plane. This thus makes the fault suspicious. More runs would be necessary to sufficiently sample the solution space and get a fully correct solution.

σ **metrics:** Figure 2.4d shows the reconstructed network chosen by both σ_{event} and σ_{fault} metrics, featuring 13 planes. Three faults in the central region are merged into a single large fault (see the arrow). This comes from the fact that the orientations of those three faults are very similar. The individual μ_F and σ_F values of this merged fault are close to the average of the values obtained on the other planes. We thus have no way to diagnose this cluster as abnormal. This may stem from the fact that the faults that generate those events are located close to each other and feature orientation differences less that the uncertainties on the focal mechanisms orientations.

Figure 2.5 shows the stereo plots of the original input faults and of the four solutions favored by the six different criteria. Plots in the left column indicate the orientations of fault traces. Dots in the right column indicate the directions of the normal vectors to the fault planes. Qualitatively, there is a nice agreement between all the reconstructed networks and the true network (first row of Figure 2.5).

This little example shows that inverting a complex but realistic structure, given realistic location uncertainty estimates, is not an easy task. However, the inverted networks, if not identical to the original one, are very similar to the original synthetic ones using the selection criteria. All validation criteria feature reasonable solutions: none of them is particularly better or worse than any of the others and the selections based on pure statistical techniques give similar fault networks as those based on tectonic constraints.

2.5.3 Comparison of the ACLUD method with the OADC method.

The OADC method uses a single, uniform and isotropic location uncertainty for the whole catalog as the clustering stopping criterion. For the synthetic Landers catalog, we computed an average location uncertainty of 1.10 km. Using this value as the stopping threshold, we performed 6,000 runs using the code of Ouillon et al. (2008). The OADC method does not feature the cross validation procedure, so that all events are used as the training data set. However, we can still rank all 6,000 solutions based on their final clustering global residuals. Figure 2.6 shows the four solutions chosen by the six following criteria: best global clustering residual, BIC, and the four focal mechanism criteria previously defined. All those four solutions selected from different criteria clearly miss the small-scale structure of the network**.** Obviously, clustering has been forced to end too early due to using an inappropriate average location uncertainty estimate, especially in the central region. As location uncertainties in the central region are smaller than close to the northern and southern edges (due to a better station coverage), the stopping criterion, that resembles the location uncertainty, should be smaller in the central region than in the edge regions. Clustering thus stopped too early in the central region and made the structure coarser. Comparing with our new method, we thus clearly see the advantage of using the true location uncertainty of each event. Comparing the four solutions, we notice that the three of them chosen by focal mechanism criteria are superior to the ones chosen by both the global clustering residuals and the BIC criteria (see Figure 2.6). These three solutions cover most of the input fault planes, yet do not include planes sampled by a small numbers of events. However, despite its simplicity, the main advantage of OADC is its fast convergence.

2.5.4 Synthetic data with background events.

The previous section showed that our technique is able to reasonably reconstruct the structure of the synthetic fault network. We now test a new assumption where the catalog of events consists in the same set as before, but now we add background events. In nature, such events also occur on faults but the latter are, for our approach, undersampled by seismicity; thus a clustering technique cannot reconstruct the structure. Specifically, we add another 20% background events to the synthetic data set uniformly distributed in the 3D space (see Figure 2.7). The latitude, longitude and depth ranges are identical to the ones of the fault-related events, providing a total number of 3,724 events. For the sake of simplicity, their focal mechanism is chosen randomly among the set of the original 3,103 events.

Our new clustering technique follows the same approach as the OADC method to detect and remove background events. The detection is based on a local density criterion, as well as on the impossibility to associate an event with a given cluster without increasing too much its thickness. However, background events are not removed from the dataset if they are located close to a fault, as they are then undistinguishable from other events.

Results obtained using the different selection procedures are shown in Figure 2.8, after 6,000 runs. Both purely statistical criteria (cross validation and BIC) select models with clearly spurious faults. For example, for cross validation, we observe a large nearly horizontal plane in the northern area while, in the southern region, original planes are divided into many small planes. Similarly, for the solution selected by the BIC, a large nearly horizontal plane is generated at latitudes 46.0° - 46.2°. Those low-dipping planes are indicated by numbers on Figure 2.8 and Figure 2.9. The best results emerge when using models selected by criteria based on focal mechanisms. Looking at the properties of each cluster (see Figure 2.9), we notice that the reconstructed horizontal faults marked 1 and 2 have very large μ_F values. This suggests that these shallow-dipping planes disagree with their associated focal mechanisms. Results chosen by cross validation and BIC clearly show the effect of these nearby background events, which distort the inverted network and require to introduce spurious shallow-dipping faults to decrease the variance of the fit. In contrast, due to the fact that background events come mostly with arbitrary mechanisms, the validation criteria based on focal mechanisms detect more efficiently the associated inconsistencies, and favor more realistic solutions.

2.5.5 Summary of synthetic tests

The synthetic tests show that our new ACLUD method successfully reconstructs fault networks, both in the case of simple or more realistic and complex structures. The tests show that, due to location uncertainties, faults that are close in space and orientation may merge into a single structure. Comparing with the previous OADC method proposed by Ouillon et al. (2008), the new method improves the results by considering location uncertainties of each individual event, thus allowing us to invert the structure more finely within areas benefiting from a better station coverage. The new method also improved the validation step, as we automated the computation of six criteria, two of them being purely statistical indices of the fit (cross validation and BIC), the four others being based on the comparison between the inverted network and the observed focal mechanisms. While all those criteria provide reasonable selected models in the absence of background events, criteria based on focal mechanisms outperform the others when such background events are present. We even obtain better solutions when including background events, which may be due to a different exploration of the solution space. For real datasets, this implies performing an extensive simulation effort to reconstruct a fault network, similar to larger scale Monte-Carlo simulations. The multiple selection criteria and their characteristics also suggest that the technique does not allow us to pinpoint single best solutions but rather emphasize that possible solution groups exist, which is likely a result of undersampling of the structures with earthquakes.

2.6 Application to the Landers aftershock series

We now apply our new clustering technique to a real dataset in the area of the 1992 M_w 7.3 Landers earthquake, already studied by Ouillon et al. (2008); this allows for a comparison of results. The catalog we used has not been published by the Southern California Earthquake data center (SCEDC) and we obtained the permission to use this data set by E. Hauksson (California Institute of Technology, personal communication). The catalog has been located using the NonLinLoc-method described in the electronic supplement. It contains 20 years of data from 1984 to 2004, with depth ranging from 1.37 km to 26.99 km. This catalog neither features the complete description of the original pdf of event locations, nor the corresponding covariance matrices that we need to input into our clustering scheme. Uncertainty is simply characterized by the lengths and orientations of the axes of the 68% confidence ellipsoid. Note that the corresponding derivation of the covariance matrix can be rigorously achieved only when the location pdf is Gaussian, a condition which generally holds only in areas well covered by a dense network of stations (Husen and Hardebeck 2010; Lomax et al. 2009). We assume that this is the case in the Landers area, due to the presence of numerous stations belonging to the permanent Southern California network, as well as due to the set of temporary stations installed during the Joshua-

Tree-Landers earthquake sequence. This is also the reason why we selected a subset of events that are most likely to be located with Gaussian uncertainties, i.e. those whose locations are particularly well constrained according to the criteria we defined in a companion work (Wang et al. 2013a). We finally retained only events located using more than 11 stations, with local magnitude M≥2, and located within an area well-covered by the station network (primary azimuthal gap smaller than 180°, ratio of the epicentral distance to the closest station over focal depth smaller than 1.5, see Bondár et al. (2004)), yielding a final subset of 3360 events (see Figure 2.10), comparable in size with our most complex synthetic example.

The focal mechanism catalog we used is computed by the HASH-method, using the locations derived by waveform cross-correlation and a 3D-velocity model (Hauksson et al. 2012; Yang et al. 2012). We only used the quality Class A and B focal mechanisms that show to about 60% focal mechanism errors of up to 20° (Hardebeck and Shearer 2002). Note, that we used locations from the unpublished catalog derived with NonLinLoc to for clustering; we associated the focal mechanisms using the event IDs to apply the validation metrics. The focal mechanism error provided is an average of the uncertainties in strike, dip and rake, mainly governed by the uncertainty of the rake angle. Given that the solutions are provided by using first motions polarities and S-wave amplitude ratios, we assume that the actual uncertainties of the strike and dip, which are important for our measures μ and σ, may be smaller.

As the clustering technique can be considered itself partly as a stochastic process, we performed 30,000 different runs in order to reasonably sample the complex landscape of the solution space. Figure 2.11 shows the fault networks corresponding to the best solutions selected from the six validation procedures. Plots present the horizontal projection of the fitting plane segments, as well as the epicenters of their associated events. For the sake of clarity, the clusters obtained for each fit are split into 2 subsets depending on their dip: clusters with dip larger than 50° (left plot) and clusters with dip smaller than 50° (right plot). As the Landers area is dominated by strike slip faulting on nearly vertical faults, we think, in the spirit of Ouillon et al. (2008), that the large-dip clusters may represent genuine underlying faults, while the low-dip clusters mainly represent spurious structures artificially introduced in order to decrease the local residual of the fit in areas of diffuse seismicity.

Each of the validation techniques yields a different solution. Clearly, there is a large number of events that are clustered on low-dip faults (dip < 50°) in the model selected by cross-validation. Looking at the properties of each cluster, we notice that there is a clear decrease of μ_F value with increasing dip, suggesting that low-dip planes disagree with their associated focal mechanisms. Thus, the solution selected by cross validation seems not to be realistic. The other validation processes yield solutions that offer a nice agreement in the northern part of the network (which can then be considered as reasonably well inverted), yet significant differences occur at other locations. If we leave aside the BIC solution for reasons explained in the section dealing with synthetic examples, we are left with four solutions that all agree well with focal mechanisms, and among which no definitive and objective choice can be made.

The fact that these validation techniques yield different selected solutions may come from the interplay of two main factors: the multiscale structure of individual faults and the spatial extent of earthquakes location uncertainties. Many studies show that faults feature a complex inner structure consisting of a complex subnetwork of sub-faults and secondary brittle structures (Tchalenko 1970; Tchalenko and Ambraseys 1970). If the time span of the catalog is much shorter than the typical time scale necessary to activate rupture on every substructure, then most of the sub-faults will feature very few events, precluding their detailed reconstruction. Furthermore, if location uncertainties are larger than the typical spacing of sub-faults, the solution to the fit of the full network is not unique either and different validation techniques will favor different solutions.

Following the same approach as Ouillon et al. (2008), we also computed the predicted surface traces of the reconstructed faults for each selected model. The idea is to prolong fault planes to the surface and compare them with the observed traces compiled by the CFM (see Figure 2.12). None of the six predicted trace maps fully agrees with the observed surface fault traces. It may stem from the fact that the catalog we used is only 20 years long, whereas surface fault traces derive from millions of years of tectonic deformation. The active part of this network is thus necessarily a subset of the full network, so that the correspondence between both sets of fault traces is necessarily imperfect. Surprisingly, Ouillon et al. (2008) obtained a solution with a more realistic predicted map of fault traces in the same area.

2.7 Discussion and Conclusions

2.7.1 Summary of the results

In this paper, we introduced a new technique (the ACLUD method) to reconstruct active fault networks which improves on the method of Ouillon et al. (2008) as it uses both earthquake locations and their estimated individual uncertainties. After a massive, yet non-exhaustive search through the very large solution space, the full set of potential solutions is submitted to six different validation procedures in order to select the corresponding best solutions. Two of the validation steps (cross-validation and BIC) process the fit residuals, while the four others look for solutions that provide the best agreement with independently observed focal mechanisms.

Tests on synthetic catalogs allowed us to qualify the performance of the fitting method and of the various validation procedures. The method is able to provide exact reconstructions in the case of very simple structures, yet is not able to find the input network when structures display more complexity and realistic location uncertainties. However, the solutions provided by each validation step are close to the expected one, especially for the BIC and focal mechanism-based techniques. Adding a uniform spatial background seismicity rate, both validation techniques based on fit residuals fail, while the ones based on focal mechanisms consistency show a much better agreement with the expected solution. The use of a uniform background density is compatible with a total lack of prior assumption about its spatial structure. Moreover, background events are generally spatially isolated as their magnitude is too small to trigger a sufficient number of close aftershocks (which would help in defining a local structure). Those low magnitude events are thus, naturally, prone to larger location uncertainties, which randomize their structure even more. Using more complex distributions, like fractal or multifractal ones, which should also be anisotropic, would require to define more arbitrary parameters.

We compared the results obtained by our new ACLUD technique with the ones obtained on the same dataset using the OADC code developed by Ouillon et al. (2008). Despite a slight difference in the nature of one of the validation procedures, we showed that the new method improves significantly on the OADC method, because accounting for individual location uncertainties of events allowed a more detailed fit of faults in areas where such uncertainties were small. It also showed that the results provided by the OADC method also improved when using validation steps based on focal mechanisms consistency. This last observation thus suggests the systematic use of such validation tools, whatever the underlying clustering technique. This also suggests that including focal mechanisms into the clustering scheme itself will provide a more consistent and efficient exploration of the solution space.

The technique has also been applied to a real data set, namely the Landers area. This study confirms that cross-validation provides a poor quality solution, as the network features a significant number of planes with a very low dip, at odds with the prior structural knowledge we have about the nature of faulting in that area. The obtained fault networks also show a poor agreement with focal mechanisms. Comparing the predicted map of fault traces for each of the six selected solutions to the actually observed map did not allow us to draw any conclusion. The reason why Ouillon et al. (2008) obtained a solution with a more realistic predicted map of fault traces in the same area remains unclear, as they did not use the same catalog. The latter may have been of lower quality than ours, which in turn allowed them to fit correctly the gross features of the network. In our case, a better assessment of locations and uncertainties may better reveal the genuine small-scale complexity of the network, which may in turn impact on the quality of the fit, for various reasons that we explain below.

2.7.2 Under-sampled multiscale faults

Many field observations suggest that faults feature a complex inner structure (Klinger et al. 2005; Tchalenko and Ambraseys 1970), consisting of a complex network of sub-faults and secondary brittle structures (like Riedel shears or flower structures, for instance). Some of the substructures may themselves feature a complex inner zone, which thus replicates itself in a more or less self-similar manner. This process necessarily holds down to a lower cutoff scale, which might be of the order of a few rock grain sizes, so that the full fault should ideally be modeled as a closely packed array of a very large number of potentially seismically active subfeatures. This view has been one of the arguments raised by Ouillon and Sornette (2011) to justify the use of a Gaussian mixture approach to cluster earthquakes. If we now assume that we can compile a catalog of all events occurring on such a fault, whatever their size and over a very long period of time, with vanishing location uncertainties, then our method would invert correctly the full underlying structure. If the time span of the catalog is much shorter than the typical time scale necessary to activate rupture on every substructure, then the sub-faults will be undersampled by the seismicity process, as most of them would feature very few events, if any. In that case, any method will fail to retrieve the correct structure of the fault zone, and our method would only provide a coarse-grained solution, which may not be necessarily unique. If we now add location uncertainties that are larger than the typical spacing of sub-faults, and sometimes comparable to the spacing of the macro faults, the coarse-graining problem will be transferred to even larger scales, so that the solution to the fit of the full network will not be unique either: different validation techniques will provide different preferred solutions.

In order to illustrate this reasoning, we extended the complexity of the synthetic Landers network of Section 4 down to smaller scales, using an algorithm inspired from the theory of Iterated Function Systems (Barnsley 1988; Hutchinson 1981), a popular technique used to build synthetic fractal sets. In a nutshell, this technique consists in replicating a given fault into another set of randomly rotated and scaled down copies of itself. The set of copies is then used to replace the original fault. The copies are themselves replaced by a similar set of rotated and scaled down copies as well, and so on, down to a given fine-scale resolution. For the sake of simplicity, this segmentation is imposed along the strike of the fault, each sub-plane extending to the same depth as the original fault. An example is shown in Figure 2.13, and features 220 sub-faults (instead of the 13 original planes). The small-scale structure appears to be very complex, yet the large scale structure is similar to the one presented in Figure 2.3. We then distribute the same set of 3,103 events over this new set of sub-faults. The network has been generated so that there is, on average, between 10 to 20 events on each segment, but some sub-faults may feature only one or two of them. (Details are given in the electronic supplement).

Using the same method as in Section 4, we generate a new catalog of events providing both their expected locations and their uncertainties. Focal mechanisms are first chosen as fully compatible with the orientation of the sub-fault to which the event is attributed, before we add a 10° uncertainty on strike, dip and rake. This catalog is then processed by our nonlinear fitting method, using 6,000 runs. This smaller number of runs is a consequence of the much larger duration of individual inversions due to the larger complexity of the dataset, which necessitates a longer time to explore the space of models.

Figure 2.14 shows the solutions selected by the six validation methods. None of them is able to reconstruct the full set of 220 planes, as expected. All proposed networks feature only 17 to 19 faults, as undersampled sub-faults are indeed merged into simpler structures in order to cluster a sufficient number of events (at least 4, as we imposed). None of the solutions are identical, reflecting the non-uniqueness of the solution provided by the different criteria.

2.7.3 Overfitting, underfitting and validation techniques

The two validation tools based on residuals, i.e. cross validation and BIC, were used in order to avoid problems of overfitting. However, we showed in the previous section that we primarily face a problem of underfitting. This observation necessarily questions the use of such validation strategies for clustering techniques. We also showed that both cross validation and BIC were unable to select the correct solution when a set of background events is superimposed over the more correlated set of earthquakes. This thus leads us to conclude that the use of such criteria is certainly much less adapted to the selection of the correct solution than the use of focal mechanisms, which bring their share of information about the dynamics of the network. Up to now, we only use part of the information contained within focal mechanisms, as we only checked the consistency of the orientation of one of the nodal planes and of the fitting planes. We thus deliberately forgot the rake. In the future, this observation should be included as well in order to better constrain solutions, thus providing a coherent set of slip vectors within the same fault.

2.7.4 Future developments

Our unsupervised clustering technique uses only the spatial information contained within seismicity catalogs. We showed that the model validation criteria derived from focal mechanisms are in better agreement with the true model when dealing with synthetics. A natural idea is then to include more prior seismic information into the clustering procedure itself, like waveform correlation coefficients, focal mechanisms similarities, and so on. However, the design of a cost function able to take account of all those different data necessitates defining a proper weighting strategy. We rather suggest using this extra knowledge to make decisions at decisive steps of the clustering process.

Despite the fact that earthquakes catalogs depict events as point processes, those events indeed define a collection of stress tensors (and their time histories during the rupture process), distributed over a set of finite planar, subplanar or fractal structures. Earthquakes define stress and strain singularities, which obviously interact through stress transmission: earthquakes are triggered by the accumulation of stress at plates’ boundaries as well as by stress fluctuations induced by previous events. Earthquakes are also increments of deformation that reveal the development and growth of faults. In return, earthquakes are constrained to occur on such faults. The geometry of the set of events is thus governed both by the applied boundary conditions and the mechanical interactions between events. The overall orientation of faults is mainly governed by the principal directions of the applied boundary stress tensor, while the inner structure and complexity of faults is mainly dominated by interactions between events.

These interactions may propagate over very large distances and time scales, through cascades of domino-effects. Indeed, faults are complex geological structures that are often considered as self-affine surfaces or self-similar aggregates of smaller scale planar features. This means that such objects are significantly correlated over a substantial range of spatial scales. The basic idea we have in mind is that such a correlation must also translate into the dynamical signature of faulting, i.e. the dynamics of the associated earthquakes. Here, we do not use the term 'dynamic' as associated to the temporal distribution of individual events (that is also given in earthquakes catalogs), but to the rupture process of individual events. The idea is that if two events occur within a short spatial distance and belong to the same fault, then there is a 'large' probability that their rupture processes will be similar (which is the basic meaning of correlation). This similarity should, on average, decrease with the distance between events. As all the information we have about the dynamics of faulting is contained within the recorded waveforms, it is thus reasonable to assume that events belonging to the same fault segment will radiate, on average, similar waveforms. Indeed, this similarity is observed and exploited for source model inversion and strong ground motion modeling in using small events as empirical Green's functions (e.g. Woessner et al. (2002)).

The most critical and arbitrary step of the clustering algorithm is the one where the locally worst cluster is split into two sub-clusters in order to improve the fit. The chosen cluster is the one with the largest thickness (so that it relies on arguments based on local fit residuals), and the split process is purely random. We suggest that, for a given number of clusters, we may first assess the μ_F and σ_F values for all individual faults. The one(s) with the largest values may then be the chosen ones to be split when increasing the number of planes, so that the splitting is now based on more mechanical grounds. We may also separate those clusters from the rest of the catalog, fit them separately, and put them back into the whole dataset. This would allow fitting separately less complex structures within smaller solution spaces, converging more quickly to a reliable solution. The randomness of the splitting may also be questioned, as we know that the standard k-means algorithm is very sensitive to initial conditions (i.e. the locations of the initial seeds), and that some of them are more optimal than others. In our case, the location, size and orientation of the new planes generated by splitting certainly have a large impact on the reliability of the final solution. Recently, both the k-means++ (Arthur and Vassilvitskii 2007) and the k-means|| (Bahmani et al. 2012) have been proposed in order to provide better initial conditions to k-means. In k-means++, the first seed is chosen randomly among the data points. All the other seeds are then chosen sequentially from the remaining data points with a probability proportional to their distance squared to the closest previous seed. The k-means|| is an improvement of k-means++ to deal with large datasets. This technique thus allows one to generate a more or less uniform set of seeds.

The most important obstacle to such clustering techniques is certainly the size of the catalogs to be processed. Up to now, we only considered sets of a few thousands of events, but the full California catalog for instance features up to half a million data points. Processing such large datasets is clearly out of reach of our current algorithm. We may improve it by parallelizing some steps (such as the computations of distances), and also by choosing more efficiently the initial conditions (as outlined above with the k-means++ approach). This limitation to process very large catalogs also holds for other clustering techniques, such as the Gaussian mixture expectation-maximization (EM) approach of Ouillon and Sornette (2011). In the latter paper, a catalog is approximated as a superposition of Gaussian kernels, whose optimal number is determined through a cross-validation strategy. A set of 4,000 events occurring in the Mount Lewis area necessitated about 100 Gaussian kernels for fitting. This large number of objects to fit the data is explained by the fact that the fitting procedure is very sensitive to density fluctuations along a given fault – whereas this is not in the case when fitting with planes. Such a fault, fitted by one single plane following our approach, may require several kernels in the EM approach, which increases the necessary computational resources. We would thus rather use our k-means-based approach to first fit the main faults, then switch to an EM approach to infer more precisely the structure of the fault zones, in the spirit of Ouillon and Sornette (2011), who were able to provide a typical segmentation scale – an information of prime importance to model high-frequency ground shaking.

The proposed clustering algorithm accounts for the full location pdf of each event and the full covariance matrix of the location errors, implemented as the stopping criterion. Four solution selection criteria based on focal mechanisms are used to find the optimal solutions. It implies that (1) location catalogs should contain detailed location error information, (2) focal mechanisms are known in great number, and with sufficient accuracy. The prerequisits currently limit the scope of this method to only a few suitable regions. Indeed, our method calls for increased efforts to generate catalogs that provide a more detailed description of the uncertainties as quantitative pattern recognition methods can now handle them – this is nothing else than is required by users of earthquake catalogs, ranging from earthquake forecasting to seismic hazard assessment. We used, for example, NonLinLoc (Lomax et al. 2000) to perform relocations as this method is able to provide the full description of possible locations, and it is an accepted method. This can be implemented for any network and may become a standard method in the future. Similarly, there are more and more methods to better describe the uncertainties of focal mechanisms. Yang et al. (2012) is only one example using the HASH-method from Hardebeck and Shearer (2002). Another example for better constraining focal mechanism uncertainties is using a Bayesian approach by Arnold and Townend (2007) and Lund and Townend (2007). We are convinced that this is not a limit of the method, but rather defines quality requirements for future studies.

Both OADC and ACLUD assume that faults can be modeled as perfectly planar objects with a vanishingly small thickness, so that each event can be attributed with certainty (i.e. unit probability) to the closest plane. It implies that location uncertainties alone explain the fit discrepancies. Instead of using planes, Ouillon and Sornette (2011) use an expectation-maximization scheme featuring Gaussian kernels to fit the data. Generalizing to fractal (or self-affine) fault models would necessitate to extend the method to stable laws (such as Cauchy or Lévy laws), which display power-law tails. Yet, no simple implementation of such kernels exists, even in 1D.

The proposed clustering approach retains the potential to improve the spatial forecasting skills of current forecast models, especially those that attempt short-term near real-time forecasts and are prone to be used for operational earthquake forecasting. Forecast models such as the Short-Term Earthquake Probability (STEP) model (Gerstenberger et al. 2005; Woessner et al. 2010) or the class of epidemic-type earthquake forecast (ETES) models (Helmstetter et al. 2006; Ogata and Zhuang 2006) have been shown to be mostly limited in their spatial predictive skill (Woessner et al. 2011). Thus, we expect that including the proposed method will improve the forecast skills at least during strong aftershock sequences and may help to improve current efforts to provide meaningful operation earthquake forecasting (Jordan et al. 2011).

2.8 Acknowledgments.

We would like to thank Egill Hauksson for providing us the Landers catalog and Stefan Wiemer for his valuable comments. This work was funded by the Swiss Science Foundation under Grant 200021-117829 and the European Commission through funds provided within the FP7 project GEISER, grant agreement no. 241321-2.

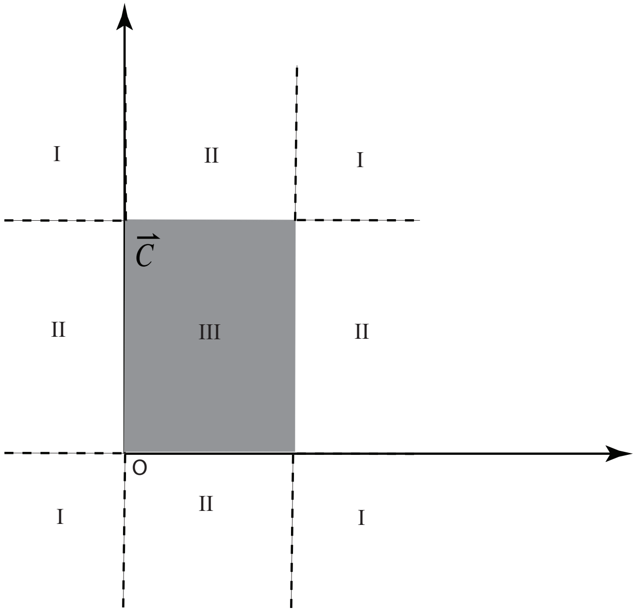


Figure 2.1: Partition of the 3D space in order to compute the expected square distance of a pdf and a finite plane (shown in grey). The roman indices I, II and III correspond to various approximations of the ESD (see main text, Section 3.3).

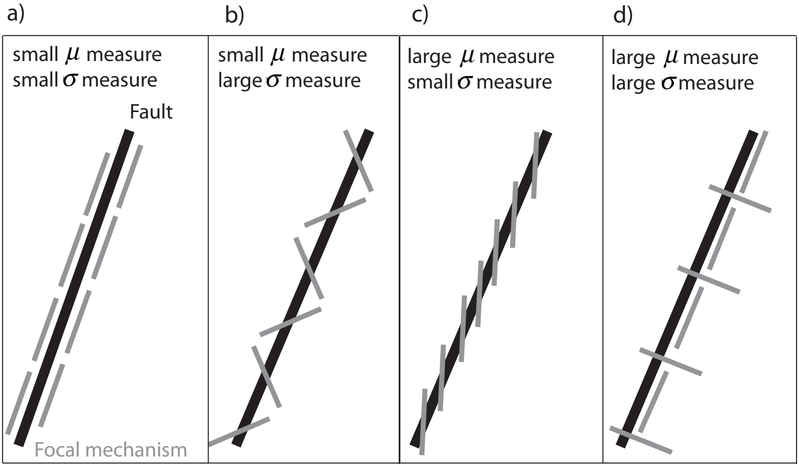


Figure 2.2: Examples of micro- and macro-structure relationships in fault zones to justify the use of different criteria based on focal mechanisms (see main text, Sections 3.5.2 and 3.5.3). Thick black line denotes general orientation of the fault zone (macro structure); thin gray lines indicate orientation of shorter individual fault planes within the fault zone (micro structure).

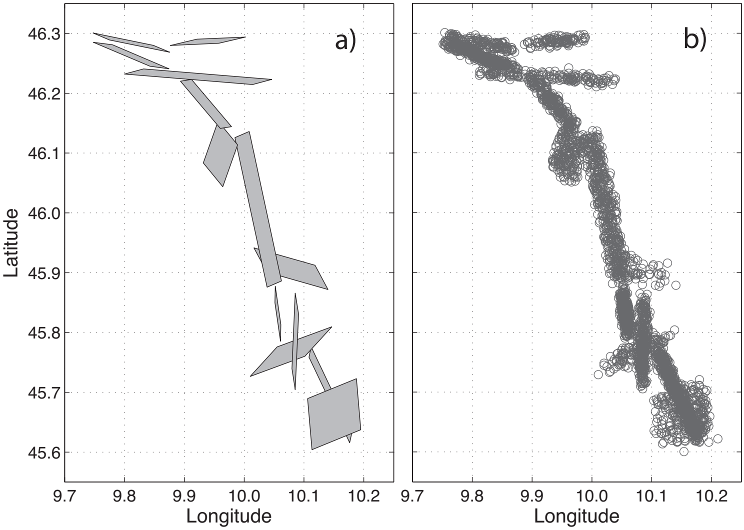


Figure 2.3: Synthetic data derived from the analysis of (Ouillon et al. 2008) on the Landers fault network. a) Fault network consisting of 13 faults. b) Epicenter map of the synthetic relocated 3,103 events.

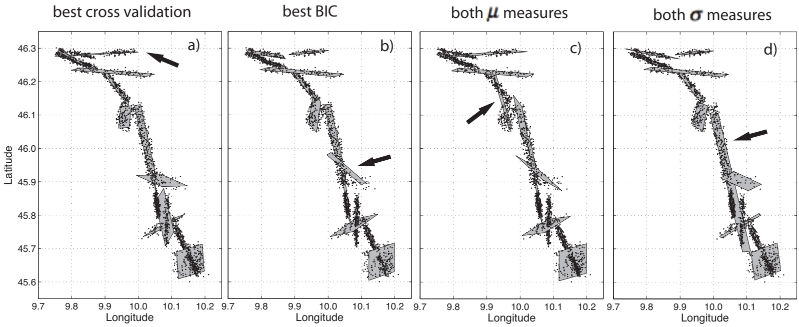


Figure 2.4: Result of our clustering method applied to the synthetic data consisting of 13 original fault planes and 3,103 events presented in Figure 3. Planes pointed by arrows are spurious faults discussed in Section 4.2.

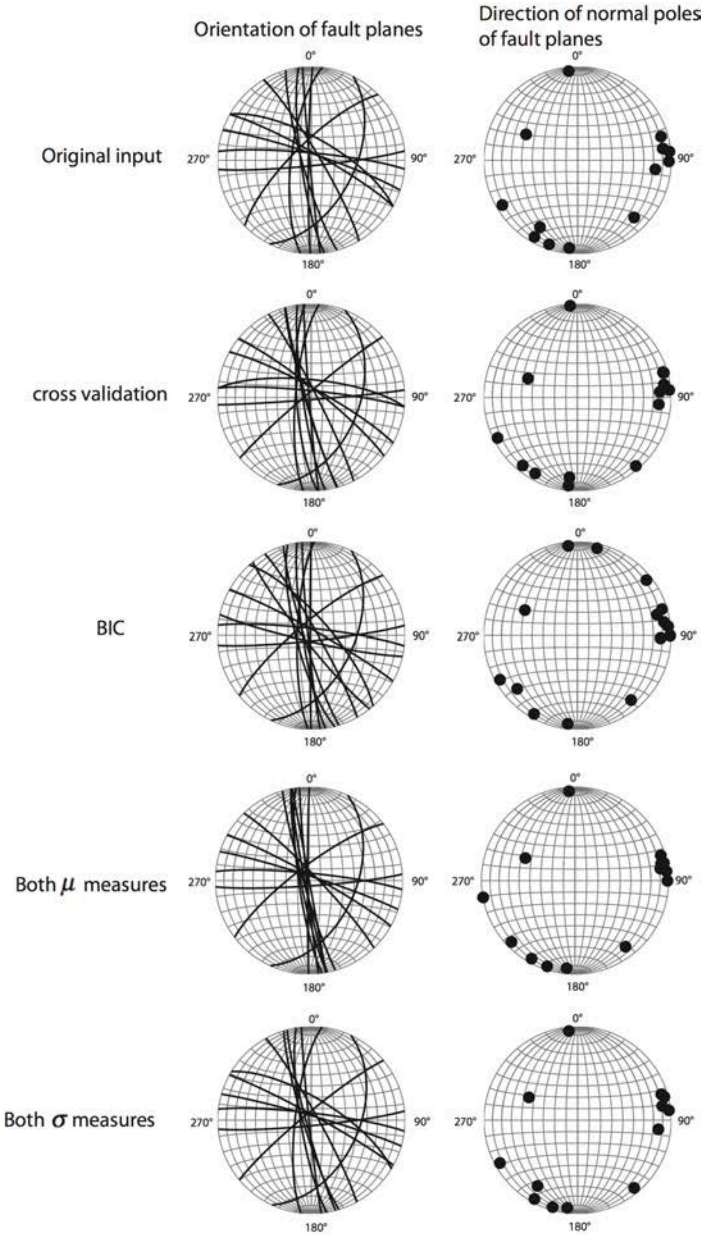

Orientation of fault planes
Direction of normal poles of fault planes
Original input
cross validation
BIC
Both μ measures
Both σ measures
0°
90°
180°
270°

Figure 2.5: Stereo plots of the original input network and solutions chosen by the six validation criteria. Curves in the left column indicate the orientations of fault traces. Dots in the right column show directions of the normal poles of fault planes.

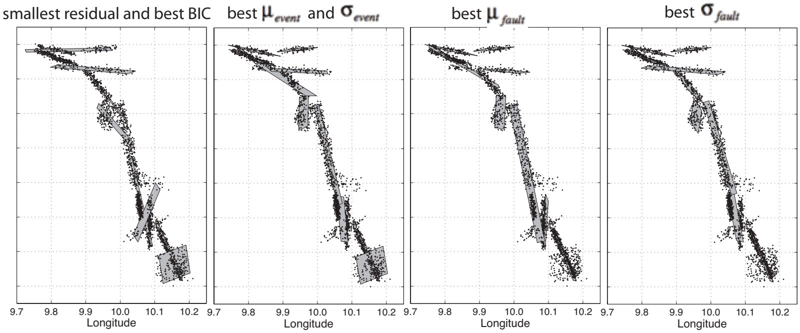


Figure 2.6: Result of the OADC clustering method of (Ouillon et al. 2008) on a synthetic dataset consisting of 13 original faults and 3,103 events.

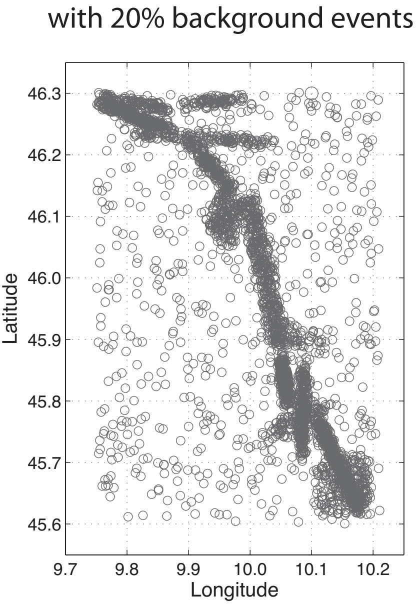


Figure 2.7: Epicentral map of the synthetic dataset with 20% background events, giving a total of 3,724 events.

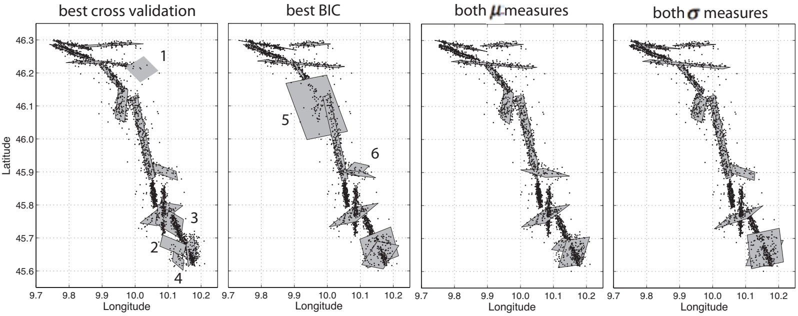


Figure 2.8: Result of our clustering method applied to the synthetic data set consisting 13 original faults with background seismicity. Solutions chosen by cross validation and BIC feature horizontal planes pointed by numbers are discussed in the text.

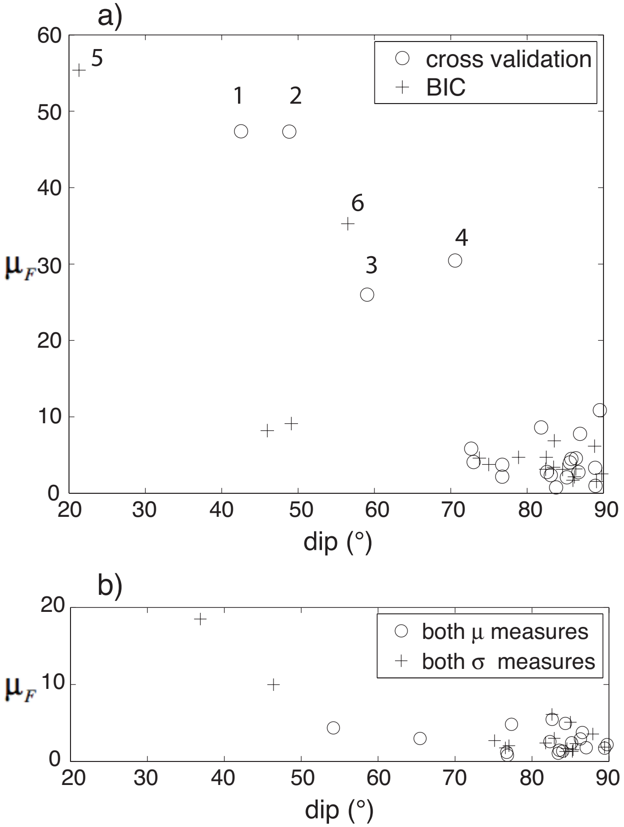


Figure 2.9: μ_F value as a function of dip of each reconstructed fault for solutions chosen by different validation criteria. The synthetic data set consists of 13 original faults with background seismicity. Solutions chosen by cross validation and BIC feature horizontal planes with large μ_F values pointed by numbers and are discussed in the text.

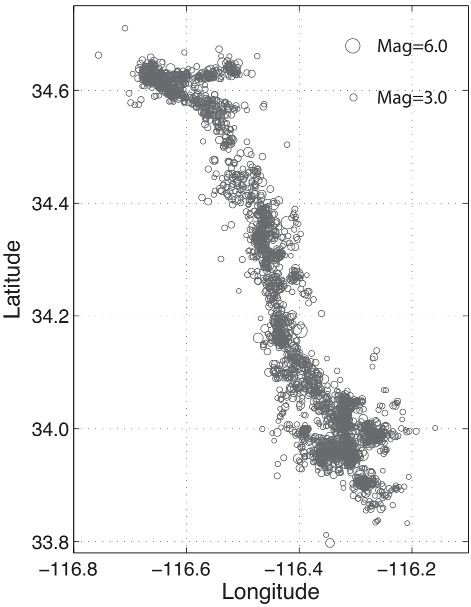


Figure 2.10: Epicentral seismicity map of the Landers area, 1984-2004. 3360 events were chosen with magnitude>2, with more than 11 observations, located within an area well-covered by the station network (primary azimuthal gap smaller than 180°, and ratio of the epicentral distance to the closest station over focal depth smaller than 1.5).

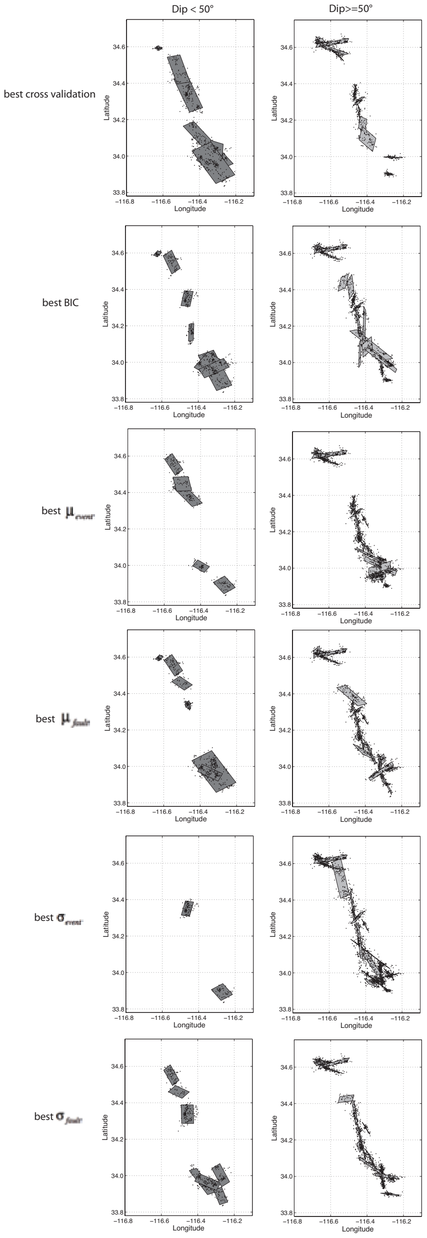

Dip < 50°
Dip>=50°
best cross validation
best BIC
best μ_{event}
best μ_{fault}
best σ_{event}
best σ_{fault}
Latitude
Longitude
34.6
34.4
34.2
34.0
33.8
−116.8
−116.6
−116.4
−116.2

Figure 2.11: Results of our clustering method applied to the Landers area using the six validation criteria. Results are presented separately for small dipping faults (dip < 50°, left) and large dipping faults (dip >= 50°, right). We assume the solutions using cross-validation and BIC as unrealistic due to the many low-dipping faults in comparison to the tectonically motivated validation measures.

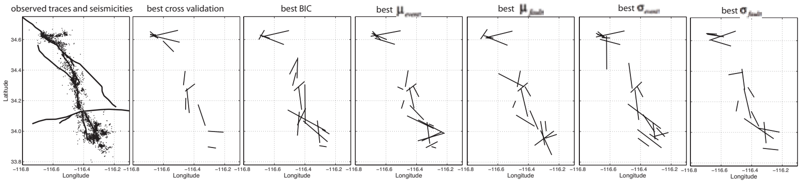


Figure 2.12: Observed surface traces and seismicity of the Landers area (left plot), and predicted sets of fault traces for each selected reconstructed network.

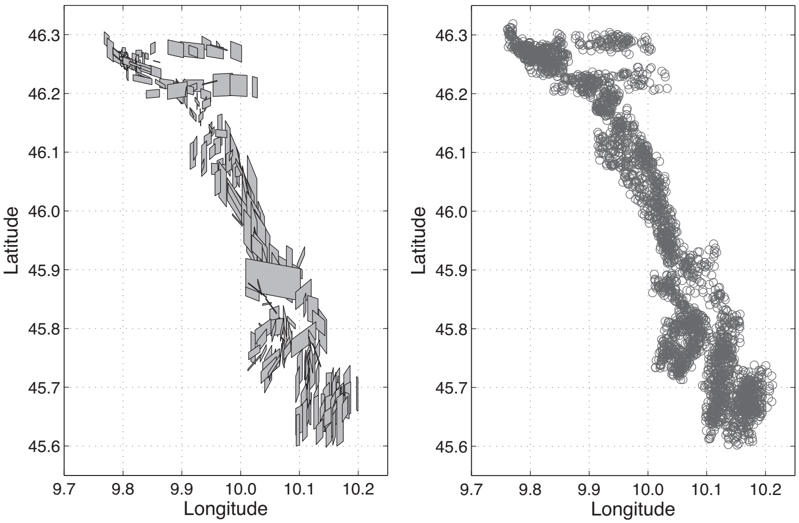


Figure 2.13: Synthetic multiscale fault network (left) and seismicity (right), consisting respectively of 220 sub-faults and 3,153 events.

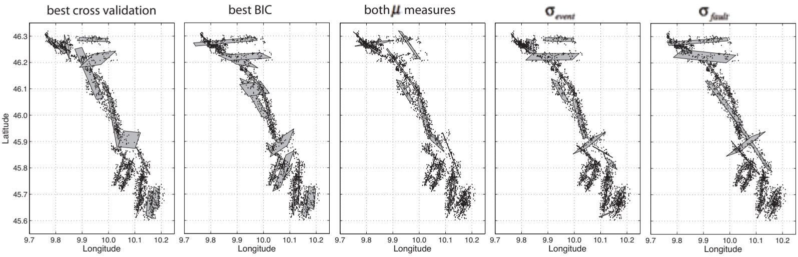


Figure 2.14: Result of our clustering method applied to the 220 synthetic multiscale fault network consisting of 3103 events. Only 17 to 19 planes were generated.

Chapter 3

Assessing earthquake location quality using seismic network criteria and its importance in fault network reconstructions

Y. Wang, S. Husen, J. Woessner, G. Ouillon, D. Sornette

Submitted to the Journal of Seismology

3.1 Abstract

We revisit station network criteria to assess earthquake location quality for local networks and study their importance in fault network reconstructions. Our study, based on a nonlinear earthquake location scheme, confirms that network criteria, such as number of observations, primary station azimuth gap and distance to the closest station are highly valuable to assess location quality. If the seismic velocity structure is accurately known, epicenter locations are well-constrained if the primary station gap is less than 180° and well-constrained focal depth estimates require a nearby station (distance to the closest station / event focal depth < 1.5). The use of classical error ellipsoids to describe location uncertainties can be misleading for earthquakes observed at less than 11 observations and with a primary station gap of more than 180°. Using two synthetic data sets for a simple, 45° dipping fault and a more complex fault structure derived from real data of the 1992 M_w=7.2 Landers aftershock sequence, we illustrate that, by using the highest-quality data selected by station network criteria, we reach better fault reconstructions of those parts of the fault structure that are sampled by the data. Using lower-quality data can lead to unstable and unreliable fault network reconstructions and may introduce artifacts, in particular in regions of a complex fault structure. Our results suggest the need for a careful assessment of the quality and reliability of reconstructed fault networks for real data applications, involving clustering of data sets of different qualities and realistic tests with synthetic fault network structures.

3.2 Introduction

Earthquakes define the location of sudden distributed slip on fault or fault networks. The interaction of tectonic processes on all scales and the myriad of earthquake ruptures cause fault networks to grow into complex structures. These complex structures are in retrospective illuminated by earthquakes of all sizes. Plotting the hypocenters displays a good first order image and together with additional seismic, geologic and geomorphologic information, a better understanding of the structure and evolution of a fault network can be achieved. However, details of fault networks are often blurred by observational and technical resolution constraints and methodological limitations. For example, random and systematic errors in earthquake locations will put limits on the detail of interpretation for a given distribution of seismicity (Husen and Hardebeck 2010; Waldhauser and Ellsworth 2000). Similarly, interpreting seismic reflection and refraction data or analyzing structural and tectonic data is limited by their uncertainties. Improving techniques to better constrain the geometry of fault networks is therefore a challenge of primary interest for earth scientists that investigate the structure, the kinematics and the dynamics of fault systems.

Over the last years, multiple initiatives have started to build large-scale fault network models compiling all possible information. The most prominent compilations are available at the USGS National Fault database, the Southern California Earthquake Center Community Fault Model (SCEC-CFM, Plesch et al. 2007) and more recently the European Database of Seismogenic Faults (EDSF, Basili et al. 2013, Giardini et al. 2013). These datasets form the base for applications such as seismic hazard assessment (Wesson et al. 2003), large scale earthquake rupture simulations, modeling fault interactions via stress transfer and so on. They are also used for more detailed studies such as comparing aftershock hypocenter locations with main shock properties (Hauksson 2010; Powers and Jordan 2010; Woessner et al. 2006) or investigating possible recurrence models on single faults (Page et al. 2011). However, these large scale models generally address only the gross structural features. Moreover, assembling these models requires major efforts that need multiple years, resulting in a product that depicts smaller geometrical complexity than would be needed for other applications, such as earthquake forecasting models (Schorlemmer and Wiemer 2005), earthquake rupture simulators (e.g. Tullis 2012; Richards-Dinger and Dietrich 2012) or understanding the complexity of the crustal stress field (Hardebeck and Michael 2006).

One approach to reconstruct the active part of a fault network solely from the spatial location of earthquake hypocenters was proposed by Ouillon et al. (2008). In a nutshell, the Optimal Anisotropic Dynamic Clustering (OADC) fits the spatial structure of a set of events with a set of finite-size plane segments. The number of segments used is increased until the residuals of the fit become comparable to the average hypocenters location uncertainty. One can then estimate the position, size and orientation of each plane segment. The main shortcoming of the OADC (Ouillon et al. 2008) is its rough account of location uncertainties, assumed to be constant for the whole catalog. While other methods have been proposed using a kernel approach (e.g. Ouillon and Sornette 2011), Wang et al. (2013b) improved the OADC method by introducing a new active fault reconstruction algorithm: Anisotropic Clustering of Location Uncertainty Distributions (ACLUD). ACLUD uses the full location uncertainty description as given by the posterior probability density function (PDF) of the earthquake location problem (Moser et al. 1992; Tarantola and Valette 1982), while OADC only used a mean uncertainty value of all events. The spatial resolution of both reconstruction methods is thus directly linked to earthquake location uncertainties and we presume that better constrained data allows one to pursue higher resolution fault reconstruction result.

Earthquake location uncertainties are a consequence of systematic and random errors in the data used to locate an earthquake. The main source for systematic errors is the imperfect knowledge of the Earth's seismic velocity structure (Husen and Hardebeck 2010). Since the true velocity structure is unknown these systematic errors are difficult to assess. Random errors in earthquake locations result from uncertainties in the observations (seismic arrival times) and from the imperfect geometry of the seismic network that recorded the earthquake (Husen and Hardebeck 2010). The latter leads to the well-known problem that uncertainties in focal depth are often larger than in epicenter. Hence, location uncertainties are rarely isotropic and can vary for each earthquake, which contradicts the assumptions used in OADC. Since ACLUD, in contrary to OADC, is able to use the full uncertainty information as provided by the PDF, we can investigate the influence of the location uncertainties on the fault reconstruction process and address the following questions: can we expect better clustering results if we only use more accurate and precise earthquake locations? Which fault structures can be resolved when using only high quality locations in comparison to all data?

Since the PDF of the earthquake location problem can be complex and multimodal in shape, a visual analysis of each PDF would be required to assess earthquake location quality. Obviously, this becomes unfeasible for large data sets. Instead, we will use so-called network criteria to select earthquake locations of different qualities. The motivation of establishing a set of network criteria follows the idea that the number and geometry of stations that record an earthquake determine how well the hypocenter estimate is constrained (e.g. Lee and Stewart 1981). Important network criteria to consider are: 1. the number of observations (*nobs*), 2. the primary and secondary azimuthal gap without observation (*GAP*) (Bondár et al. 2004), and 3. the ratio between the distance to the closest station and focal depth (*DIST*). Assessing earthquake location errors has a long history (Bondár et al. 2004; Husen and Hardebeck 2010). Much of the recent work has been developed within the efforts to monitor the Comprehensive Nuclear Test Ban Treaty CTBT (e.g. Bondár and McLaughlin 2009a; Bondár et al. 2004; Yang et al. 2004). These studies are based on high-quality ground-truth events with known location accuracies (Bondár and McLaughlin 2009b; Bondár et al. 2004) recorded at local, regional and teleseismic distances. An important drawback of the CTBT studies is that they use global, one-dimensional reference velocity models, such as IASPEI91 or ak135, for relocation. While these velocity models may be applicable at regional and teleseismic distances, their use for local networks is highly questionable. Moreover, these studies focused mainly on the accuracy of epicenter estimates, which is of primary concern for the monitoring efforts of the CTBT. Nevertheless, these studies developed so-called network criteria to assess the accuracy of local earthquake locations.

In the first part of our study we therefore establish a set of network criteria to assess earthquake location quality for local networks. We do so by relocating 10,000 earthquakes randomly distributed on a simple, single fault with randomly chosen station geometries and varying number of stations. For each relocated hypocenter, we analyze mislocations (distance between true and relocated hypocenter location) as a function of three different network criteria, i.e. *nobs*, primary *GAP* and *DIST*, to assess their influence on earthquake location quality. We further compute the percentage of true hypocenter locations that are located inside the 68 % confidence ellipsoid as given by the PDF of the earthquake location problem to assess the reliability of the computed location uncertainties. Our results confirm that network criteria are valuable parameters to assess earthquake location quality. For example, epicenter locations are well-constrained if the primary *GAP* < 180°, while focal depth is well-constrained if *DIST* < 1.5. Our results also show that the 68 % confidence ellipsoid is a viable approximation of the location uncertainties if *nobs* ≥ 11.

In the second part of our study, we then investigate how data sets of different earthquake location qualities, as selected by the set of network criteria established in the first part, affects fault network reconstruction. We do this by generating synthetic data for a simple, 45° dipping fault and a more complex fault structure derived from real data of the 1992 M_w=7.2 Landers aftershock sequence. Subsets of different earthquake location qualities are then processed with ACLUD and reconstructed fault structures are compared to the true fault structures. Our results show that using the highest-quality data only leads to a better fault reconstruction of those parts of the fault structure that are sampled by the data. Using lower-quality data may introduce artifacts, in particular in regions of a complex fault structure.

3.3 Assessing network criteria

3.3.1 Synthetic data

We generate synthetic data to assess location uncertainties in earthquake catalogues. In contrast to real earthquake catalogues, the true locations are known for synthetic data. This allows us to assess location precision, i.e., location uncertainties as computed by the location program, as well as location accuracy, i.e., the difference between the true and relocated hypocenter. For real data, location accuracy can only be evaluated using sources with a known location, e.g., explosions (Bondár et al. 2004; Husen et al. 2003), and few such sources exist. Moreover, the use of synthetic data enables us to generate data sets with a large range of different

GAP and *DIST* values, which is necessary to analysis how station distribution affects earthquake location.

We generated synthetic data using the following procedure:

(1) We spatially distributed 10,000 events uniformly on a vertical fault plane with a length of 100 km and a width (depth extent) of 20 km (Figure 3.1). The strike of the fault is 0° and dip 90°. Those 10,000 events represent the true locations.

(2) For each event we randomly selected 6, 8, 11 and 22 stations from a regular grid of 88 stations (Figure 3.1). This yielded four different data sets (R6, R8, R11, R22), each consisting of 10,000 events with varying station geometries. By randomly selecting stations, we built data sets with strongly varying geometries, encompassing a large range of different *nobs*, *GAP* and *DIST* values.

(3) For each observation, we computed synthetic travel times using a 1-D velocity model representing a 30 km thick crust (P-wave velocity of 6.0 km/s) over a mantle (P-wave velocity of 8.0 km/s) (Table 3.1). We computed synthetic travel-times using a finite-difference scheme of the Eikonal equations (Podvin and Lecomte 1991). To simulate picking uncertainties, we added Gaussian noise with zero mean and a standard deviation of 0.1 s to the synthetic travel times. The chosen pick uncertainty corresponds to high-quality data, which can be achieved using automated, quality-weighted picking algorithms (Diehl et al. 2009).

(4) We relocated each event using the same velocity model used to compute synthetic travel times (Table 3.1).

In total, we created four data sets of relocated earthquake locations, each consisting of 10,000 earthquakes. The varying network geometries allow us to comprehensively study the effect of network geometry on earthquake location precision and accuracy. We note that our procedure will favor station geometries with small *GAP* values for earthquakes observed at a large number of stations (*nobs* > 11). This is due to the fact that true hypocenter locations are located in the center of the regular grid of 88 stations (Figure 3.1). We think, however, that such a setup reflects a realistic monitoring situation in which stations are distributed around the target tectonic structure.

We relocated earthquakes using the NonLinLoc software package (Lomax et al. (2000), Version 5.2, http://alomax.free.fr/nlloc/). Compared to traditional linearized approaches, NonLinLoc is superior in that it computes the posterior probability density function (PDF) using nonlinear, global searching techniques. The PDF represents the complete probabilistic solution to the earthquake location problems, including comprehensive information on uncertainty and resolution (Moser et al. 1992; Tarantola and Valette 1982).

In the work of (Tarantola and Valette 1982), *a priori* information on data and model uncertainties are assumed to be Gaussian. For NonLinLoc, data uncertainties are described by a Gaussian distribution with a standard deviation corresponding to the estimated arrival time uncertainty. We arbitrarily model data uncertainties with mean of zero and a standard deviation of 0.1 s, since the same velocity model is used for relocation model uncertainties.

For each location estimate, uncertainties are described by the posterior PDF with no *a priori* assumption of their shape. They can be visualized using scatter density plots, which are obtained by drawing samples from the posterior PDF with their number being proportional to the probability (Husen et al. 2003; Lomax et al. 2000). From these samples, the 68 % confidence ellipsoid can be computed by singular value decomposition of the corresponding covariance matrix. The 68 % confidence ellipsoid describes the location uncertainties as computed by linearized location algorithms such as HypoEllipse (Lahr 1989). It forms a compact approximation of the spatial uncertainty of the location estimate. The expectation hypocenter is at the center of the confidence ellipsoid, and the maximum likelihood hypocenter will always fall within the densest part of the PDF. Figure 3.2 shows the scatter density plot and corresponding 68% error ellipsoid of one event of the data set located with six randomly drawn stations. In this example, the *GAP* is 128° and *DIST* is 3.8.

3.3.2 Results

We present our results in terms of mislocation (the difference between the true and relocated hypocenter) in both epicenter and focal depth direction, and location uncertainties as computed by the 68 % confidence ellipsoid. The former is related to the accuracy of an earthquake location estimate, while the latter characterizes the precision of an earthquake location estimate. Since we are interested in assessing accuracy and precision in terms of network criteria, we further group the results using a combination of three network criteria, including *nobs*, primary *GAP*, and *DIST*. We did not use the secondary *GAP* as suggested by (Bondár et al. 2004) since it is often not computed for earthquake bulletins. Since the primary *GAP* and *DIST* depend on the hypocenter location, they were computed using true hypocenter locations. This ensures that each group contains the same events for all data sets (R6, R8, R11, and R22). The primary *GAP* is directly related to network geometry and provides a quantitative measure of how well an event is surrounded by stations (Bondár et al. 2004). For example, events with a primary *GAP* < 180° are usually considered to be well-locatable (Kissling 1988). The *DIST* criterion is considered to be important to constrain estimates of focal depth. For example, one observation from a station at an epicentral distance less than the focal depth is often required to constrain estimates of focal depth (Chatelain et al. 1980; Gomberg et al. 1990; Husen et al. 2003). We present epicenter mislocations in Table 3.2 and focal depth mislocations in Table 3.3. We analyze the performance of each group and data set by computing cumulative density functions (CDF) for mislocation in epicenter and focal depth (Figure 3.3). From these CDFs, we choose mislocation values at the 90th percentile to discuss the effect of *nobs*, *GAP* and *DIST* on earthquake location accuracy for the different synthetic data sets. In this paper, we use Δepi_{90} and $\Delta depth_{90}$ as notations for epicenter and focal depth mislocation at the 90th percentile, respectively. We chose the 90th percentile since the relatively low number of earthquakes for some groups does not allow a homogeneous sampling of the CDF at higher percentile levels (see Figure 3.3). To analyze earthquake location precision, we compute the number of events for which the true hypocenter is not within the bounds of the 68 % confidence ellipsoid (Table 3.4). Theoretically, this number should correspond to 32 % of the total number of events. Any significant deviation from this value could suggest that it may not be appropriate to use the 68 % confidence ellipsoid to assess earthquake location precision.

3.3.2.1 Epicenter mislocation

Table 3.2 shows Δepi_{90} for each data set obtained by randomly selecting station networks (R6, R8, R11 and R22) grouped by different network criteria. The corresponding CDF plots are shown in Figure 3.3. For all of these data sets, having more observations leads to smaller epicenter mislocations. The smallest epicenter mislocations are obtained for those events with smallest *GAP* ($0° \leq GAP < 90°$). There exists, however, a significant increase in epicenter mislocation for events with $GAP \geq 180°$. While the increase in Δepi_{90} is only moderate for earthquakes grouped by $0° \leq GAP < 90°$ and $DIST < 1.5$ to earthquake grouped by $90° \leq GAP < 180°$ and $DIST < 1.5$ (e.g. 0.8 km to 1.0 km for data set R8 in Table 3.2), Δepi_{90} increases by nearly a factor of two for earthquake grouped by $180° \leq GAP < 270°$ and $DIST < 1.5$ (1.8 km for data set R8 in Table 3.2). This suggests that a $GAP < 180°$ is important to constrain epicenter estimates. No significant differences in Δepi_{90} are found between events with a nearby station ($DIST < 1.5$) and with no nearby station (*DIST* ≥ 1.5). Compare, for example, Δepi_{90} for events of data set R8 grouped by 90° ≤ *GAP* < 180° and *DIST* < 1.5 (1.0 km) and grouped by 90° ≤ *GAP* < 180° and *DIST* ≥ 1.5 (0.9 km) in Table 3.2. This implies that the distance to the closest station is not important to constrain the epicenter estimate if the correct velocity model is used for relocation.

Our observations that the primary *GAP* is mainly controlling mislocation in epicenter is consistent with the findings that, if an earthquake occurs outside of a network, it is difficult to constrain its epicenter (Lee and Stewart 1981). This is linked to the properties of the Jacobian matrix of the linearized earthquake location problem, which contains the partial derivatives of the travel times with respect to the epicenter coordinates and focal depth. From linear algebra, it is known that, if a matrix has a column that is nearly a multiple of another column, it is a rank-defective matrix with a very small singular value. In this case, the Jacobian matrix is difficult to invert and, hence, the earthquake location problem is ill-conditioned or poorly constrained. If an earthquake occurs outside of a network, it is likely that the columns of the Jacobian matrix containing the partial derivatives with respect to the epicenter coordinates are nearly proportional to each other (Lee and Stewart 1981). Consequently, the earthquake location problem becomes ill-conditioned and the epicenter is poorly constrained.

3.3.2.2 Focal depth mislocation

Table 3.3 shows $\Delta depth_{90}$ for each data set obtained by randomly selecting station networks (R6, R8, R11 and R22) grouped by different network criteria. The corresponding CDF plots are shown in Figure 3.3. Similar to the results obtained for epicenter mislocation, we observe smaller mislocations with increasing number of observations. On the contrary, only a moderate increase in focal depth mislocation is observed for earthquakes with *GAP* ≥ 180°. For example, $\Delta depth_{90}$ is equal (1.3

km) for earthquakes of data set R11 grouped by 90° ≤ *GAP* < 180° and *DIST* < 1.5, and by 180° ≤ *GAP* < 270° and *DIST* < 1.5 (Table 3.3). This suggests that *GAP* is not a critical parameter to constrain estimates of focal depth if the correct velocity model is used for relocation. Focal depth mislocations, however, strongly depend on the distance to the closest station if the correct velocity model is used for relocation. $\Delta depth_{90}$ is about four times smaller if a station nearby (i.e. *DIST* < 1.5) observed the event than if not (Table 3.3). Interestingly, $\Delta depth_{90}$ is comparable for events of data set R22 grouped by 90° ≤ *GAP* < 180° and *DIST* ≥ 1.5 (3.6 km) to events of data set R6 grouped by 90° ≤ *GAP* < 180° and *DIST* < 1.5 (2.3 km) (Table 3.3). This suggests that the lack of a nearby station can be partly compensated by a large number of stations if the correct velocity model is used.

Our observations that the distance to the closest station, as described by the *DIST* parameter, is important to constrain focal depth is in agreement with previous studies, which have shown that stations within a focal depth's distance are important to constrain focal depth (Chatelain et al. 1980; Gomberg et al. 1990; Husen et al. 2003). Similarly to the situation where an earthquake occurs outside the network, the Jacobian matrix of the linearized earthquake location problem becomes ill-conditioned if no nearby station observes the earthquake (i.e. *DIST* > 1.5). In this case, the partial derivatives of the travel times with respect to focal depth become similar, and, consequently, the corresponding column of the Jacobian matrix becomes a multiple of the first column, which contains only ones (Gomberg et al. 1990; Lee and Stewart 1981). Following this logic, the observation that a lack of a nearby station can be partly compensated by a large number of stations can be explained by an improved range of epicentral distances, which yields larger variations of the partial derivatives with respect to focal depth. For a larger number of stations, chances are higher in our synthetic data that a more distant station will be selected.

3.3.2.3 Location precision

To assess the reliability of location uncertainties, we computed the number of events for each data set, for which the true location is not covered by the 68 % confidence ellipsoid (Table 3.4). This number should be close to 32 % if the 68 % confidence ellipsoid presents a reliable approximation to the true location uncertainties. From Table 3.4, we notice that the number of events for which the true locations are not covered by 68 % confidence ellipsoid decreases gradually with increasing number of observations. Only for data set R22, the number is close to the expected value of 32 %. This suggests that the 68 % confidence ellipsoid is not an appropriate approximation of the true location uncertainties for a significant number of events in data sets R6 and R8.

The observed deviations from the expected 32 % may appear small; however, with a simple, rigorous hypothesis test using a binomial distribution (Stegman 1989), we can show the significance of the deviations. The null hypothesis is that the 68 % confidence ellipsoid is a correct error approximation. Given the n = 10,000 event relocations, we expect with a probability of 0.5 that, for a given event, the true hypocenter is located inside the 68 % confidence ellipsoid; in other words, we expect with a 50-50 chance that for 6800 events the true hypocenter location falls within the confidence ellipsoid. Observing a much smaller or higher number of events leads to reject the null hypothesis. For data set R22, we observe 6700 events with the true hypocenter location located inside the 68 % confidence ellipsoid. The probability to observe this number based on the binomial test is 0.016. The probabilities decrease dramatically to 9.3 x 10^{-11} (R11, 35 %), 4.6 x 10^{-37} (R8, 38 %) and 8.85 x 10^{-64} (R6, 40 %), leading to reject the null hypothesis for these data sets. This implies that, for a significant number of events in data sets R6, R8 and R11, the 68 % confidence ellipsoid is not an appropriate approximation of the location uncertainties. For data set R22, we can expect a few events for which the confidence ellipsoid may not be appropriate.

For earthquakes observed at few stations, the PDF can have a non-ellipsoidal shape and, hence, the confidence ellipsoid is not a good approximation of the true location uncertainties (Husen and Hardebeck 2010; Lomax et al. 2000). The difference between the maximum likelihood and the expectation hypocenter location can be used as a measure to identify PDFs that are non-ellipsoidal, which usually requires visual inspection of a large number of earthquake locations (Husen and Smith 2004). We therefore check whether earthquakes observed at few stations show a larger number of PDFs that are non-ellipsoidal. For each data set (R6, R8, R11 and R22), we computed CDFs of the difference between maximum likelihood location and expectation location in both epicenter and focal depth (Figure 3.4). For each data set, we analysed the PDF for those events for which the difference between maximum likelihood and expectation hypocenter location in both directions was larger than the 99th percentiles of the corresponding CDF. We found that, for data sets R6 and R8, 10 out of 10 and 8 of 13 events, respectively, showed PDFs that were non-ellipsoidal. In contrast, the corresponding numbers for data sets R11 and R22 were 1 out of 16 and 0 out of 12. This confirms that there are many events in data sets R6 and R8 for which the PDF is non-ellipsoidal in shape.

3.4 Application to fault network reconstruction

We will now investigate how data quality impacts fault network reconstruction. We do this by using two synthetic fault structures: i) a simple 45° dipping fault plane, and ii) a complex fault network for the Landers area in southern California consisting of 13 planes with dips > 45°. For each fault structure, we compute synthetic earthquake catalogs including location uncertainties. Different subsets are selected by applying different selection criteria: i) no selection criterion (i.e. full data set), ii) *GAP* < 180°, termed soft selection criterion in the following, and iii) *GAP* < 180°, *nobs* ≥ 11, *DIST* ≤ 1.5, termed stringent selection criterion in the following. Each subset is clustered using the fault reconstruction method ACLUD (Wang et al. 2013b) and reconstructed fault networks are compared against true (input) fault structures.

3.4.1 Anisotropic clustering of location uncertainty distributions (ACLUD)

Ouillon et al. (2008) proposed the optimal anisotropic data clustering (OADC) method to reconstruct the active part of a fault network from the spatial location of earthquake hypocenters. It is inspired from the seminal k-means method (MacQueen 1967), which partitions a given dataset into a set of (*a priori* isotropic) clusters by minimizing the global variance of the partition. Ouillon et al. (2008) generalized this method to the anisotropic case with a new algorithm, which, in a nutshell, fits the spatial structure of the set of events with a set of finite-size plane segments. The number of segments used increases until the residuals of the fit become comparable to the average hypocenters location uncertainty. One can then estimate the position, size and orientation of each plane segment. More details on the OADC method can be found in Ouillon et al. (2008).

The main shortcoming of the OADC method is the implicit assumption that location uncertainties are uniform and isotropic for the whole catalog. This implies that location uncertainties are equal in all directions and identical for all earthquakes in a catalog. Obviously, this assumption is rather unrealistic since location uncertainties depend strongly, due to picking and velocity model errors, on station network geometry (see section 2). To improve on the OADC method, Wang et al. (2013b) introduced the anisotropic clustering of location uncertainty distribution (ACLUD) method to reconstruct active fault networks. The ACLUD method extends the OADC method by taking into account the detailed and individual location uncertainties of each event. This is achieved by introducing the expected squared distance between a probability density function and a finite plane to associate earthquake locations as described by the probability density function and the closest plane (Wang et al. 2013b). Furthermore, the ACLUD method introduces a dynamic stopping criterion that links the average location uncertainty in the direction normal to a given earthquake cluster and the thickness of that cluster. This allows one to adapt locally the resolution of the fit to the location uncertainties.

Another advantage of the ACLUD method is that it allows for a massive search through the entire solution space of possible reconstructed fault networks (Wang et al. 2013b). This is achieved by computing a large number of solutions. Since search for solution is a stochastic process, thus different runs converge to different solutions. The full set of potential solutions is submitted to six different validation procedures in order to select optimal solutions (Wang et al. 2013b). Two of the validation steps (cross-validation and Bayesian Information Criteria (BIC)) are purely statistical approaches that process the data fit of all solutions. The four other validation procedures use independent information from observed focal mechanisms to identify solutions that provide the best agreement. Wang et al. (2013b) show that, compared to cross-validation, BIC shows stronger stability, especially when the earthquake catalog contains spatial background seismicity, i.e. events that cannot be clustered and are automatically removed from the plane fitting procedure. Accordingly, we will use the BIC to select the optimal solution since we will focus on synthetic data that do not contain focal mechanisms.

3.4.2 Synthetic data

In order to analyze the effect of data selection criteria on fault reconstruction, we focus on tests with synthetic data that allow us to compare the reconstructed solutions with the original input. Two synthetic fault structures are used. The first simulates a simple fault structure consisting of a 45° dipping fault. The second is more complex and uses the synthetic fault structure in the area of the 1992 M_w 7.3 Landers earthquake (Ouillon et al. 2008). It features 13 planes with dips larger than 45° (Figure 3.5). For each synthetic fault structure, we generated synthetic earthquake catalogs including location uncertainties as described in the following.

Following the same approach we applied in section 2, a 45° dipping fault was placed in the middle of the station grid shown in Figure 3.1. We then uniformly distributed 2000 events on the fault. Synthetic travel times were computed for each of the 2000 events to all stations using a simple two-layer model (Table 3.1). Gaussian distributed errors with a mean of 0.1 s were added to simulate realistic picking errors. Out of this set of synthetic travel times, we randomly chose 6, 8, 11 or 22 stations for each event for relocation. Earthquakes were relocated using the same simple two-layer velocity model (Table 3.1) and a non-linear probabilistic earthquake location technique (NonLinLoc) as described in section 2. As a result, we obtained a set of 2000 relocated earthquakes including a full description of location uncertainties as given by their scatter density clouds. These data served as input to the ACLUD method.

The aim of the second synthetic fault structure was not only to use a complex fault network but also to use a realistic station network to compute synthetic travel times. The geometry of the 13 fault planes was based on clustering results obtained using a set of 3103 earthquakes observed within two weeks following the 1992 M_w 7.3 Lander mainshock at stations of the Southern California Seismic Network (Ouillon et al. 2008). For the synthetic fault structure, however, we assume that these 3103 earthquakes occur randomly and uniformly on the 13 fault planes (Wang et al. 2013b). As a consequence, we lose the original information about which earthquake was observed at which station. Instead, we had to use the following approach to select a set of stations for each earthquake:

1. For each of the 71 real stations that were operational in our study region, we computed the number of events it observed from the 3103 earthquakes used in the study of Ouillon et al. (2008). From these numbers, we computed an observation hash table that represents the fraction of events observed at each station (Figure 3.5).
2. For each synthetic earthquake location, we randomly chose 6, 8, or 11 or 22 stations. The probability of a station to be chosen was proportional to the numbers given in the observation hash table. Hence, stations that observed a higher fraction of the original 3103 earthquakes had a higher probability to be chosen.
3. For each selected station, we computed synthetic travel times using a simple two-layer model (Table 3.1). Gaussian distributed errors with a mean of 0.1 s were added to simulate realistic picking errors.

The goal of our approach was to create synthetic travel time data that showed distributions in *nobs*, *DIST*, and *GAP* as close as possible as for the real data. Following our standard procedure, we relocated all 3103 synthetic earthquakes using the same simple two-layer velocity model (Table 3.1) and a non-linear probabilistic earthquake location technique (NonLinLoc) as described in section 2. As a result, we obtained a set of relocated earthquakes including a full description of location uncertainties as given by their scatter density clouds.

As we discussed in Section 2, network criteria *nobs*, *GAP*, and *DIST* can be used to select well-constrained hypocenter locations. In order to test the effects of applying network criteria in fault network reconstruction, we created three different earthquake catalogs for each synthetic fault network. The first catalog is simply using all data without applying any selection criteria. The second catalog is generated by applying what we call a soft selection criterion, i.e. *GAP* < 180°. For the third catalog, we apply what we call a stringent selection criterion, i.e. *GAP* < 180°, *DIST* ≤ 1.5 and nobs ≥ 11. By applying ACLUD to all three earthquake catalogs, we can analyze whether applying different selection criteria will lead to different fault network reconstructions.

3.4.3 Results

3.4.3.1 Simple fault structure

Our simple fault structure consists of one fault plane dipping at an angle of 45°. From the original 2000 earthquakes distributed uniformly on the fault plane, we are left with 1732 earthquakes and 259 earthquakes when applying soft and stringent selection criteria, respectively (Table 3.5). Each earthquake catalog is processed with ACLUD, computing 1000 solutions for each catalog with different initial conditions of the random number generator that controls the fault splitting step (Wang et al. 2013b). One solution thereby corresponds to one fault reconstruction result. We first analyzed the number of reconstructed faults for all 1000 solutions for each catalog (Figure 3.6). As can be inferred from Figure 3.6, more than 90 % of the solutions converge to more than five faults if either no selection criteria (full catalog) or the soft selection criterion are applied. If the stringent selection criterion is applied, more than 90 % of the solutions converge to one, two or three faults. Given the original number of one fault, apparently most of the solutions are overfitting, no matter how we selected the data. However, using the stringent data selection criterion clearly reduces the over fitting problem.

Among all 1000 solutions, we selected the best solution according to BIC. They are shown in Figure 3.7 for each earthquake catalog. There are seven and six faults generated if no selection criterion and the soft selection criterion are applied, respectively. Both solutions feature one large fault plane and several smaller fault planes (Figure 3.7b,c). Strike and dip of the large fault plane are in good agreement with the original input, i.e. strike = 180°, dip = 45°. The smaller fault planes are likely generated to cover events that were offset from the main fault plane by the relocation process. If the stringent selection criterion is applied, the best solution according to BIC features two faults (Figure 3.7d). Note that, due to the constraint of $DIST \leq 1.5$, most of the shallow events are removed in this subset, thus only the deeper part of the fault is reconstructed. However, compared to the two previous results, the reconstructed fault structure is simpler and agrees better with the true structure of one dipping fault, although a significantly smaller number of earthquakes has been used.

As discussed in section 2, different selection criteria have different effects on location quality. Earthquakes that pass our soft selection criterion (i.e. $GAP < 180°$) have well constrained epicenter locations but uncertainty in focal depth can be large. In order to obtain well constrained focal depths, a sufficient number of observations ($nobs \geq 11$) and a station nearby are needed (i.e. $DIST \leq 1.5$). These criteria correspond to our stringent selection criterion. It becomes obvious that, in order to resolve a dipping fault structure, hypocenter locations need to be well constrained in epicenter and focal depth. Hence, it can be expected that reconstruction of a low dipping fault becomes more stable only if earthquakes are used that pass our stringent selection criterion, which is what we observe. The fact that earthquakes that pass our stringent selection criterion are well constrained in epicenter and focal depth is supported by the observation that, on average, they show smaller shifts after relocation, compared to the original location, and they exhibit smaller volumes of the 68 % confidence ellipsoids (Table 3.6*)*. It is interesting to note that, even with well-constrained hypocenter locations, we are not able to fully recover the simple structure of a 45° dipping fault. Our best result, according to BIC, using best data (stringent selection criterion) shows two fault planes. As we use an error-free velocity model and Gaussian picking errors, the precision of fault reconstruction result is mainly controlled by the geometry of the station network. This suggests that the current station network setup does not allow us to perfectly resolve the one 45° dipping fault.

3.4.3.2 Complex fault structure

Our synthetic complex fault structure consists of 13 planes with a dip larger than 45° (Figure 3.5a). Earthquakes are distributed uniformly on each plane. With no selection criterion applied, the earthquake catalog consists of 3103 events. Applying soft and stringent selection criteria leaves 2257 and 400 events, respectively (Figure 3.8). Each earthquake catalog was processed with ACLUD, computing 6000 solutions for each catalog with different initial conditions of the random number generator (Wang et al. 2013b). One solution thereby corresponds to one fault reconstruction result. Similar to the results for a simple structure, we first analyzed the number of reconstructed faults for all solutions (Figure 3.9). Using all data (i.e. no selection criterion applied), the majority of solutions (about 90 %) show five or less fault planes, which is significantly lower than the true number of 13 faults. Applying the soft selection criterion, we obtain a bi-modal distribution with about 45 % of the solutions showing 5 or less faults and 40 % of the solutions showing more than 15 faults (Figure 3.9). About 80 % of the solutions show 10, 11 or 12 faults if the stringent selection criterion is applied. Our interpretation is that clustering results become more stable and closer to the true number of faults if only well-constrained hypocenter locations are used (i.e. when using the stringent selection criterion), although only a fraction of the original number of events are used. Interestingly, using all earthquakes (no selection criterion applied) also yields stable clustering results but the recovered structure seems oversimplified, given that about 90 % of the solutions show five or less fault planes.

For each earthquake catalog, Figure 3.10 shows the best solution as defined by the BIC. For all three solutions, large dipping faults are more or less correctly recovered when using the constrained earthquake data. Discrepancies between the different solutions, however, exist in the northern and southern parts, and for shallow dipping faults. For example, a spurious NNW-SSE striking, shallow dipping fault is introduced in the northern part (marked with an arrow in Figure 3.10a) when all earthquakes are used. This fault is not introduced if the soft or stringent selection criterion is applied. In the southern part, our synthetic fault structure shows two shallow dipping fault planes, which are not correctly recovered if the full earthquake catalog is used or the soft selection criterion is applied; if the stringent selection criterion is applied, two shallow dipping fault planes are introduced but strike and size of these planes are slightly different compared to the original fault planes (Figure 3.10). These observations suggest that earthquake locations, which are poorly constrained in focal depth, contaminate well constrained hypocenter locations in these regions, introducing spurious, mainly large dipping fault planes. This is consistent with our observations in the previous section, which demonstrated that well constrained hypocenter locations in epicenter and focal depth are needed to correctly recover dipping fault structures. Interestingly, the small fault in the central part (marked with an arrow in Figure 3.10a) is reconstructed when clustering the full catalog but not when using the better quality catalogs. We also notice that some solutions recovered this small fault if the soft criterion is used. Due to the low event density on this fault and the complex structure (overlapping), the reconstruction is highly unstable (for more discussion, see Wang et al. 2013b).

Applying our stringent selection criterion, only a fraction of all earthquake locations can be used for fault network reconstruction (Figure 3.8). In particular, the earthquake distribution becomes sparse at latitudes 34.4° N - 34.5° N due to a gap in the station distribution in this region (Figure 3.5). Consequently, no fault planes are reconstructed if the stringent selection criterion is applied. In the remaining regions, the original fault network is quite well recovered despite a significantly lower number of events. It is interesting to note that the original fault network at latitudes 34.4° N - 34.5° N is quite well recovered if the full earthquake catalog is used or the soft selection criterion is applied (Figure 3.10). This is likely due to the relatively simple structure of the fault network in this region, which consists of large, non-overlapping fault planes. South of 34.1° N, where the original fault network is more complex with overlapping fault planes at different dips, the fault network reconstruction is significantly worse if the earthquake catalog is used or the soft selection criterion is applied (Figure 3.10). Overall, our results demonstrate that the original fault network is more reliably reconstructed if well constrained hypocenter locations are used, i.e., data is selected using our stringent selection criterion. This effect becomes particularly important for complex fault structures with overlapping fault planes at different dips. It should be noted, though, that in certain regions where station distribution becomes sparse, no information on the original fault network can be recovered.

3.5 Discussion

3.5.1 Network criteria

In the first part of our manuscript, we assessed how network criteria (*nobs, GAP, DIST*) can be used to select well-constrained hypocenter locations. Our results show that well-constrained epicenter locations can be obtained if the following network criteria are met: *nobs* $\geq$ *11* and *GAP* < 180°; a nearby station is not needed to constrain epicenter locations. Well-constrained hypocenter locations can be obtained if the following network criteria are met: *nobs* $\geq$ 11*,* *GAP* < 180°, and *DIST* $\leq$ 1.5. The motivation for establishing a set of network criteria to select well-constrained hypocenter locations follows the idea that the number and geometry of stations that record an earthquake determine how well the hypocenter estimate is constrained (Lee and Stewart 1981). This idea has a long history in seismology and much of the recent work in the literature has been developed within the efforts to monitor the Comprehensive Nuclear Test Ban Treaty CTBT (Bondár and McLaughlin 2009a; Bondár et al. 2004; Yang et al. 2004). As already mentioned, these studies are based on high-quality ground-truth events with known location accuracies (Bondár and McLaughlin 2009b; Bondár et al. 2004) recorded at local, regional and teleseismic distances. We have already stressed that an important drawback of the CTBT studies is that they use global, one-dimensional reference velocity models, such as IASPEI91 or ak135, for relocation. While these velocity models may be applicable at regional and teleseismic distances, their use for local networks is highly questionable. Moreover, these studies focused mainly on the accuracy of epicenter estimates, which is of primary concern for the monitoring efforts of the CTBT. Nevertheless, these studies developed so-called network criteria to assess the accuracy of local earthquake locations. For example, using Monte Carlo simulations of network geometries for two explosions, (Bondár et al. 2004) found that epicenter locations of local networks are accurate to within 5 km with a 95 % confidence level if the following network criteria are met: (1) *nobs* > 10, all within 250 km, (2) primary *GAP* < 110^{0}, (3) secondary *GAP* < 160^{0} and (4) at least one station within 30 km. In a more recent study, these criteria were updated using data from 47 GT0 explosions, of which 35 were located at the Nevada Test Site (Bondár and McLaughlin 2009b). The new results revealed that epicenter locations of local networks are accurate to within 5 km with a 95 % confidence level if the following network criteria are met: (1) *nobs* > 10, all within 150 km, (2) *Δu* < 0.35, (3) secondary *GAP* < 160°, and (4) at least one station within 10 km, where *Δu* describes the absolute deviation between the best-fitting uniformly distributed network of stations and the actual network (Bondár and

McLaughlin 2009b). The latter two constraints also ensure that focal depth is resolved with an accuracy of 7 km at the 95 % confidence level. Compared to these findings, our criteria seem less stringent. This could be explained by the fact that these studies used global velocity models and blasts located at the surface. Sources located close to the surface are difficult to relocate due to expected near-surface velocity heterogeneity (Husen et al. 2003). Global velocity models have no resolving power at shallow depth, and thus introduce significant velocity model errors. Due to the inherent coupling of seismic velocities and hypocenter locations, these errors will affect hypocenter locations as well (Pavlis 1986; Thurber 1992). In our study, we avoid this by using the same velocity model to compute synthetic travel times and to relocate the earthquakes. The main motivation for doing so is that, for fault network reconstruction, earthquake location precision is more important than accuracy. Due to their systematic nature, velocity model errors will mainly affect absolute locations but not necessarily relative locations between a set of earthquakes (Pavlis 1986), which are important to constrain fault networks.

Our results suggest that the 68 % confidence ellipsoid is not an appropriate approximation of the location uncertainties if earthquakes are observed at less than 11 stations (i.e. data sets R6 and R8). For these data sets, a significant number of events did not contain the true hypocenter locations within the bounds of the 68 % confidence ellipsoids (Table 3.4). If an earthquake is observed at only a few stations, the PDF can have a non-ellipsoidal shape and, hence, the confidence ellipsoid is not a good approximation of the true location uncertainties (Husen and Hardebeck 2010; Lomax et al. 2000). Visual inspections of some of these events revealed that they often suffer from poor control on focal depth (i.e. *DIST* > 1.5), which in all cases yield elongated PDFs. In order to check whether earthquakes with poor control on focal depth are more likely to show non-ellipsoidal PDFs, we randomly chose 40 events from data set R11. Out of these, 28 events showed non-ellipsoidal PDFs, of which 16 events had poor control on focal depth (*DIST* > 1.5). This suggests that events with poor control on focal depth are more likely to show non-ellipsoidal PDFs. Following the same approach, we found no evidence that the earthquakes with a primary *GAP* > 180° are more likely to have non-ellipsoidal PDFs. It should be noted that the conclusion on whether a PDF has an ellipsoidal shape or not is somewhat subjective. As discussed in section 2.2.3, the difference between the maximum likelihood and expectation hypocenter location can be used as a measure to identify PDFs that are non-ellipsoidal. Nevertheless, a universal threshold does not exist since the shape of the PDF depends non-linearly on station geometry and parameterization of the velocity model (layered model or gradient model).

Our results have important consequences if earthquakes are only observed at a few stations. For these earthquakes, location uncertainties, as computed by linearized location methods such as HYPOELLIPSE (Lahr 1989), become unreliable since they assume location uncertainties to be ellipsoidal in shape, which follows out of the underlying *Chi-square* or *F*-statistics (Boyd and Snoke 1984). It should be noted as well that our results do not account for errors in the velocity model. It has been shown however that, in the presence of these errors, formal uncertainties as computed by the location programs can be misleading (Pavlis 1986). This is due to the systematic nature of these errors, which does not obey the underlying assumption of Gaussian distributed measurement and model errors.

3.5.2 Application to fault network reconstruction

In the second part of this paper, we showed the effect of applying different data selection criteria on fault network reconstruction. We found fault reconstruction can be highly unstable and unreliable if no selection criterion or the soft selection criterion are applied, even in the case of a single 45° dipping fault plane. We showed that even the optimal results selected from the solution space contain several spurious faults. These faults are likely introduced to fit hypocenter locations shifted by a larger distance in the relocation process. It should be expected that these hypocenter locations are associated with larger uncertainties, which in principle should allow ACLUD to assign them to the correct plane fault. Instead of using all samples of the scatter density cloud for a given earthquake location PDF, ACLUD uses the covariance matrix and the associated 68 % confidence ellipsoid to compute the expected square distance between the PDF and a given fault plane and to compare the size of the average confidence ellipsoid in the direction normal to the cluster with the thickness of each cluster (Wang et al. 2013b). By computing the number of events, for which the 68 % confidence ellipsoid does not contain the true hypocenter location, we found in section 2 that the number of events with a non-ellipsoidal shaped PDF increases for earthquakes observed at less than 11 stations. For these events, the 68 % confidence ellipsoid is likely not a good approximation of the true uncertainties (see also the discussion in section 4.1). In our tests with synthetic data shown in section 3, 50 % of the events are observed at less than 11 stations if no or the soft selection criterion are applied; only if the stringent selection criterion is applied, we find that events are observed at 11 or more stations. Consequently, when no or the soft selection criterion are applied, we can expect that a larger number of events show a non-ellipsoidal shaped PDF. For these events, the 68 % confidence ellipsoid, as used in ACLUD, is not a good approximation of the true uncertainties, which may lead to the observed introduction of additional or spurious fault planes. Only if the stringent selection criterion is applied, does the vast majority of events show an ellipsoidal shaped PDF and reconstructed fault networks become closer to the true structure.

For our synthetic data, only 10 % of the events are retained if the stringent selection criterion is applied. For the single fault structure, we still have a sufficient amount of data for the fault reconstruction. In contrast, the network reconstruction does not resolve certain regions for the Lander fault network (e.g. region between 34.40 N and 34.50 N in Figure 3.10). In these regions, the station network is not as dense, leading to less well-constrained hypocenter locations in these regions. Nevertheless, poorer quality data does recover the true fault network in this region relatively well, which is likely due to the simple fault structure in this region (i.e. large, non-overlapping faults). This poses the question of how to perform fault network reconstruction for real data, where the true fault network is unknown and data selection may significantly reduce the number of usable events. One approach could consist of first using highest-quality data (e.g. applying the stringent selection criterion) in fault network reconstruction. This would outline so-called well-resolved regions, where data quality is high, and, hence, fault network reconstruction is reliable. Data of lower quality can then be used for the remaining regions but care should be exercised in interpreting the obtained results, as artifacts are likely. Moreover, fault networks in these regions should clearly be labeled as poorly or only fairly-well resolved. For real data applications, our results also demonstrate the usefulness of tests with synthetic fault networks, which allows one to identify regions where data quality and fault network reconstruction can be problematic. In order to utilize the power of these tests, they need to be as realistic as possible, e.g. the real station network should be used to compute synthetic travel times. Overall, the identification of well- and poorly-resolved regions in fault network reconstruction is similar to the situation in seismic tomography, where the heterogenous distribution of sources (earthquakes) and receivers (stations) demands a careful assessment of the quality or resolution of the obtained model (Husen et al. 2000; Thurber et al. 2007).

3.6 Conclusions

We have used synthetic data and a nonlinear probabilistic earthquake location technique (NonLinLoc) to comprehensively assess the influence of the number and distribution of stations on the accuracy and precision of local earthquake location estimates. Our study confirms that network criteria, such as *nobs,* primary *GAP, DIST,* are highly valuable to assess location quality. Epicenter locations are well-constrained if the primary *GAP* < 180° and well-constrained focal depth estimates require a nearby station (*DIST* < 1.5). It should be noted that these results only apply if the seismic velocity structure is accurately known. Moreover, our results imply that the use of classical error ellipsoids to describe location uncertainties can be misleading for earthquakes observed at a low number of stations (nobs < 11) and with a primary *GAP* of more than 180°.

Investigating two synthetic datasets computed for a simple fault structure (i.e. a single, 45° dipping fault) and for a complex fault structure derived from a real data application to the 1992 M_w 7.3 Landers aftershock sequence (Ouillon et al. 2008; Wang et al. 2013b), we have illustrated the influence of earthquake data quality on fault network reconstruction. By applying three different data selection criteria, we found that fault network reconstruction can be unstable and unreliable when using poorly constrained location data. In turn, this implies that using high-quality data selected by the proposed network criteria leads to high-quality fault network reconstruction results. Our results thus suggest the need for a careful assessment of the quality and reliability of reconstructed fault networks, involving clustering of data sets of different qualities and realistic tests with synthetic fault network structures.

Based on the assessment of hypocenter location quality, we have illustrated its influence on one application: reconstructing fault networks. This forms an example for many other studies that use earthquake catalogs as essential input data. The lesson learned in this study should be investigated in multiple other applications, such as detailed mapping of earthquake parameters (Mignan et al. 2011; Schorlemmer et al. 2004; Woessner and Wiemer 2005), building earthquake forecast models (e.g. Hainzl et al, 2010. Zhuang et al. 2011, Zhuang et al. 2012 and reference therein), or studies that evaluate earthquake forecasting models (e.g. Woessner et al. 2011; Zechar et al. 2013 and references therein). The conclusion from our results is simple, but nicely illustrated: higher quality input data serves to develop better constrained images of faults structures within the crust and its related processes. Low quality or heterogeneous quality data sets often blur the true resolvable results.

3.7 Acknowledgements

We thank Anthony Lomax for providing us with his NonLinLoc software and his excellent comments. We are grateful for the comments by J. Zechar and improving our non-native English mix to a more precise and concise wording. This work was funded by the Swiss Science Foundation (grant SNF 200021-117829) and the European Commission through funds provided within the FP7 project Geiser, grant agreement no. 241321-2.

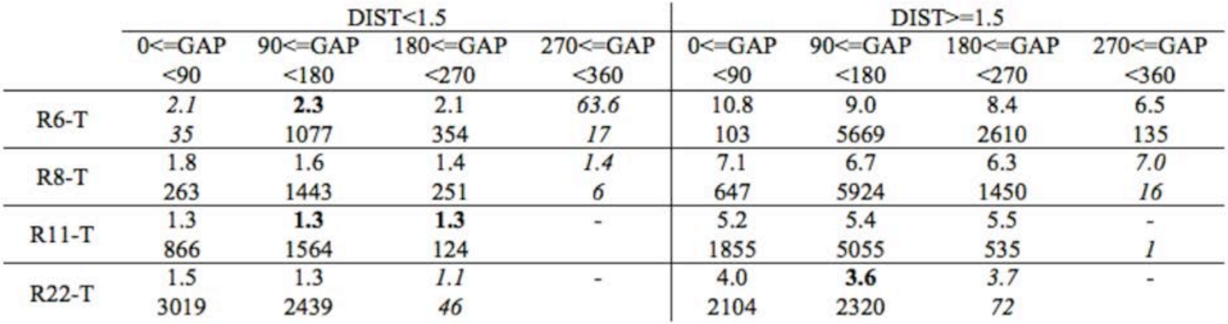
Table 3.1: P-wave velocity model used to compute synthetic travel times and to relocate synthetic data.

Depth (km)	P-wave velocity (km/s)
0.0	6.0
30.0	8.0

Table 3.2: Mislocation in epicenter at the 90th percentile for data sets located with increasing number of stations (R6-R12). Values are grouped by network criteria DIST and GAP as indicated. The lower values in each row indicate the number of events in each group. Numbers in italic mark groups for which the number of events is low (< 50); bold numbers indicate values discussed in text

	DIST<1.5				DIST>=1.5			
	0<=GAP <90	90<=GAP <180	180<=GAP <270	270<=GAP <360	0<=GAP <90	90<=GAP <180	180<=GAP <270	270<=GAP <360
R6-T	*1.0* *35*	1.4 1077	**2.3** 354	*6.0* *17*	1.0 103	1.2 5669	**2.1** 2610	6.7 135
R8-T	**0.8** 263	**1.0** 1443	**1.8** 251	*3.3* *6*	0.8 647	**0.9** 5924	1.6 1450	*4.0* *16*
R11-T	0.7 866	0.8 1564	1.3 124	-	0.6 1855	0.8 5055	1.2 535	- *1*
R22-T	0.5 3019	0.5 2439	*0.7* *46*	-	0.4 2104	0.5 2320	*0.6* *72*	-

Table 3.3: Mislocation in focal depth at the 90th percentile for data sets located with increasing number of stations (R6-R12). Values are grouped by network criteria DIST and GAP as indicated. The lower values in each row indicate the number of events in each group. Numbers in italic mark groups for which the number of events is low (< 50); bold numbers indicate values discussed in text.



	DIST<1.5				DIST>=1.5			
	0<=GAP <90	90<=GAP <180	180<=GAP <270	270<=GAP <360	0<=GAP <90	90<=GAP <180	180<=GAP <270	270<=GAP <360
R6-T	*2.1*	**2.3**	2.1	*63.6*	10.8	9.0	8.4	6.5
	35	1077	354	*17*	103	5669	2610	135
R8-T	1.8	1.6	1.4	*1.4*	7.1	6.7	6.3	*7.0*
	263	1443	251	*6*	647	5924	1450	*16*
R11-T	1.3	**1.3**	**1.3**	-	5.2	5.4	5.5	-
	866	1564	124		1855	5055	535	*1*
R22-T	1.5	1.3	*1.1*	-	4.0	**3.6**	*3.7*	-
	3019	2439	*46*		2104	2320	*72*	

Table 3.4: Percentage of events for which true locations are not covered by the 68 % error ellipsoid. Percentages were computed using all 10,000 events in each data set. Observed deviations from the theoretical value of 32 % are caused by a significant number of events for which the 68 % error ellipsoid is an unreliable description of the true error.

Data set	Percentage of events
R6	40%
R8	38%
R11	35%
R22	33%

Table 3.5: Number of events under three different data selection criteria.

	No selection	Soft selection	Stringent selection
One 45° dipping fault	2000	1732	259
Synthetic Landers	3103	2257	400

Table 3.6: Average mislocation in epicenter and in focal depth, and average volumes of the 68 % confidence ellipsoids for earthquake catalog under three different data selection criteria.

	No selection	Soft selection	Stringent selection
Mislocation in epicenter	0.54	0.44	0.31
Mislocation in focal depth	1.62	1.59	0.59
Volume of 68% con. ellipsoid	10.71	7.29	1.29

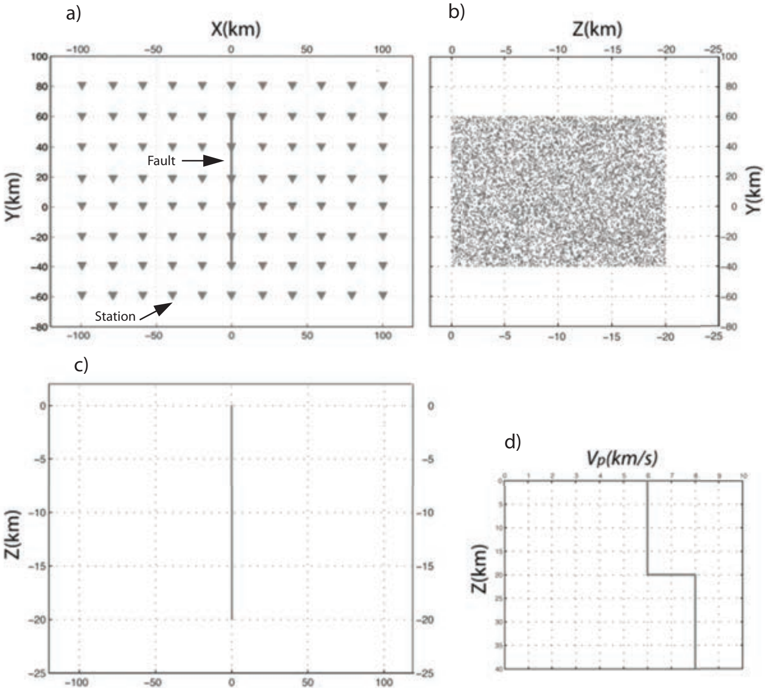


Figure 3.1: Network design, fault location and earthquake distribution to compute synthetic data: a) Map view: Triangles represent stations (88 in total, 20 km spacing). The black line indicates the fault surface trace along which earthquakes were distributed. b) Vertical cross-section (y-z-section). Dots represent location of 10,000 earthquakes uniformly distributed along the fault plane. c) Same as b) but along X-axis. d) Synthetic one-dimensional (1-D) velocity model used to compute synthetic travel times.

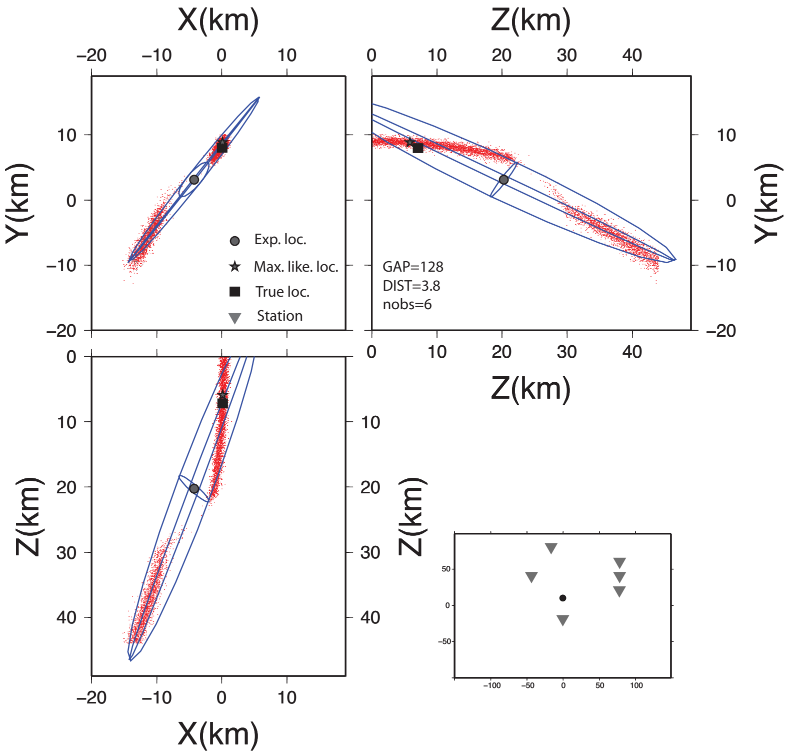


Figure 3.2: Density scatter plot drawn from the *posterior* PDF (red dots) as computed with NonLinLoc for an earthquake observed at six randomly chosen stations (data set R6). Map view (upper left) and vertical cross sections (upper right and lower left) are shown together with location network (lower right). Ellipses represent projection of the 68 % confidence error ellipsoid. Square represents the true location. Maximum likelihood location (star) and expectation location (circle) show difference to true location. Network criteria are as indicated.

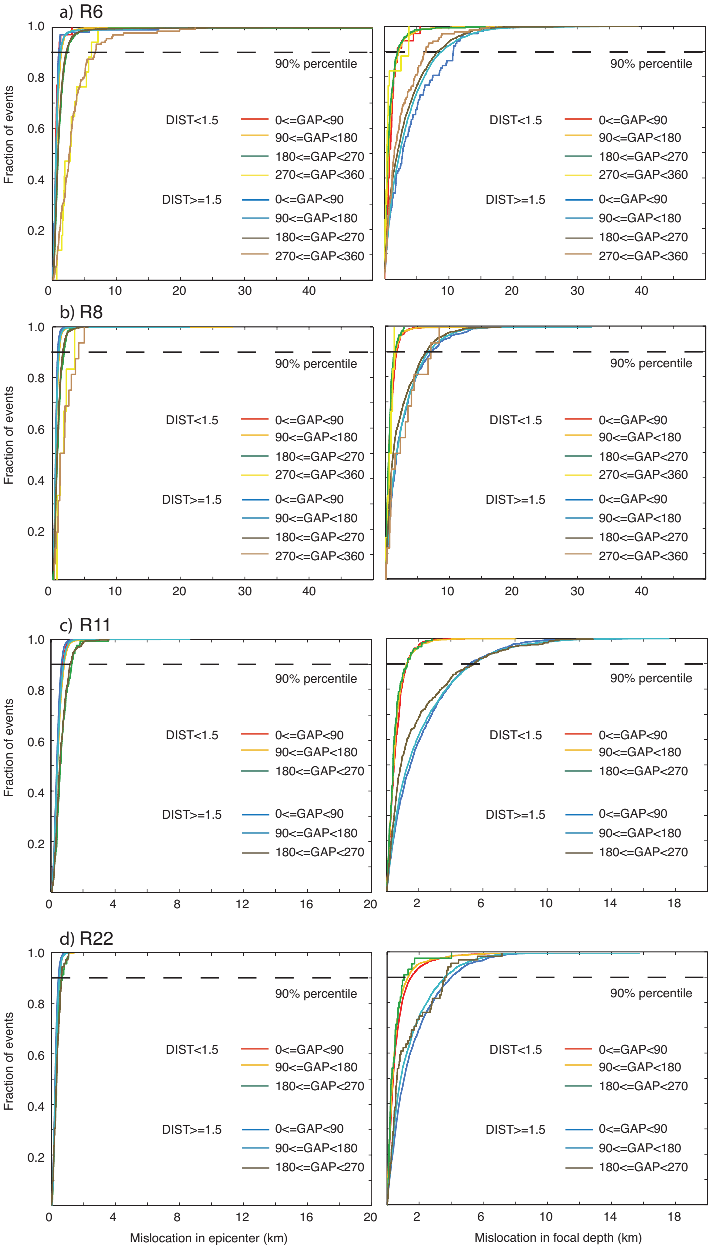

a) R6
b) R8
c) R11
d) R22
Fraction of events
1.0
0.8
0.6
0.4
0.2
0
10
20
30
40
4
8
12
16
20
2
6
10
14
18
90% percentile
DIST<1.5
0<=GAP<90
90<=GAP<180
180<=GAP<270
270<=GAP<360
DIST>=1.5
0<=GAP<90
90<=GAP<180
180<=GAP<270
270<=GAP<360
Mislocation in epicenter (km)
Mislocation in focal depth (km)

Figure 3.3: Cumulative Density Functions (CDF) of mislocation in epicenter (left) and focal depth (right). CDF were computed for 10,000 events of data sets a) R6, b) R8, c) R11, and d) R22 grouped by different network criteria (*GAP, DIST*) as indicated. The 90th percentiles are computed for each sub-group and are listed in Tables 2.2 and 2.3. At the 90th percentile, 90 % of the events show a mislocation that is equal or lower than the corresponding value.

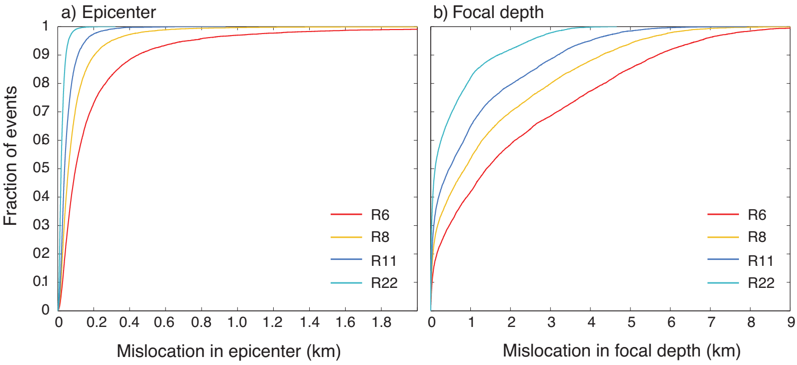


Figure 3.4: Cumulative Density Functions (CDF) of absolute difference between maximum likelihood and expectation hypocenter locations in a) epicenter and b) focal depth for data sets R6, R8, R11, R22 as indicated. Since hypocenter locations with fewer observations are less well-constrained, differences between maximum likelihood and expectation hypocenter locations increase with decreasing number of observations. In addition, focal depth is commonly less well-constrained than epicenter, yielding larger differences in focal depth between maximum likelihood and expectation hypocenter locations.

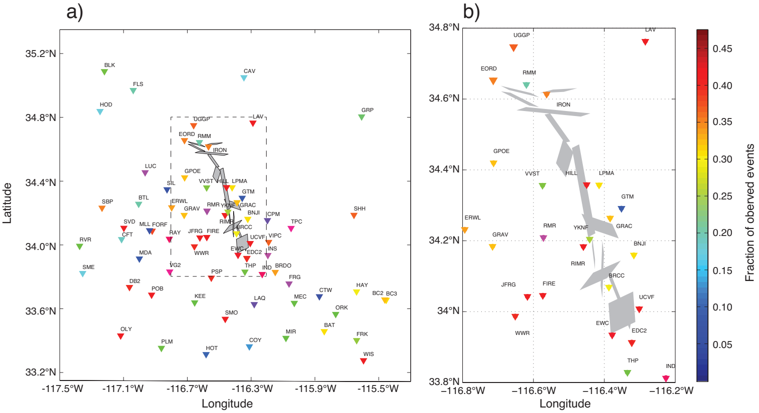


Figure 3.5: Synthetic Landers fault network with real station network. a) Active stations in the region with longitude as [-117.4°, -115.4°] and latitude as [33.25°, 35.25°] during 2 weeks span after the main shock on 28.06.1992. b) Synthetic Landers fault network from (Ouillon et al. 2008) consisting of 13 faults.

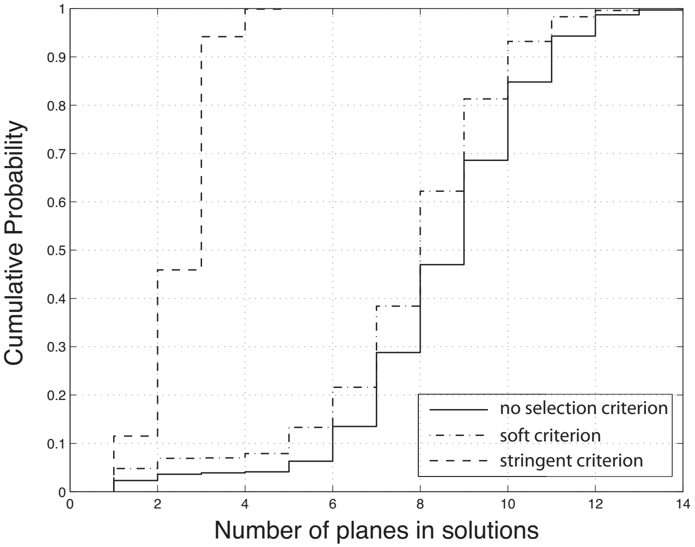


Figure 3.6: Cumulative Distribution Function (CDF) of the number of clustered planes for one 45° dipping fault structure. The clustering method was applied to three data sets under different data selection criterion. The whole dataset contains 2000 events. There are 1732 events under the soft criterion (GAP <= 180°). There are 259 events under the stringent criterion (GAP <= 180°, DIST <= 1.5 and nobs >= 11).

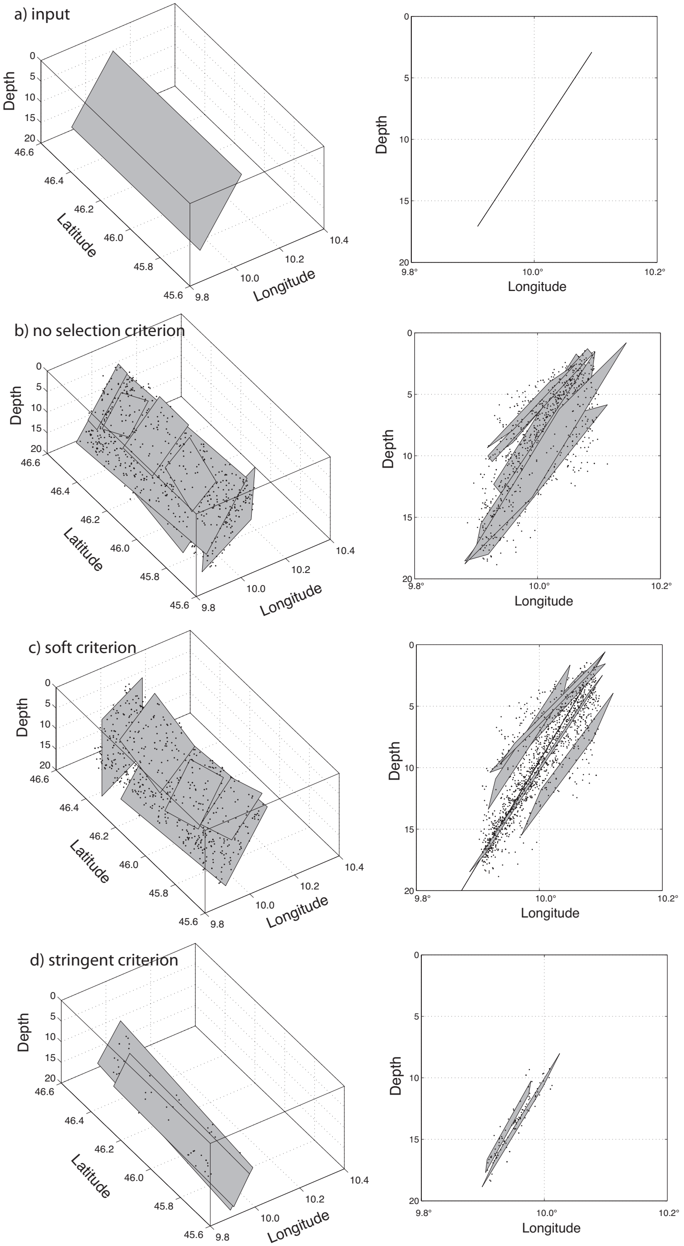

a) input
Depth
0
5
10
15
20
46.6
46.4
46.2
46.0
45.8
45.6
Latitude
9.8
10.0
10.2
10.4
Longitude
Depth
0
5
10
15
20
9.8°
10.0°
10.2°
Longitude
b) no selection criterion
Depth
Latitude
Longitude
Depth
Longitude
c) soft criterion
Depth
Latitude
Longitude
Depth
Longitude
d) stringent criterion
Depth
Latitude
Longitude
Depth
Longitude

Figure 3.7: Result chosen by the BIC of the clustering method applied to the synthetic data consisting of one 45° dipping fault 2000 events. With no data selection criterion, 7 planes were generated. Under the soft selection criterion, 6 planes were clustered. Under the stringent criterion, 2 planes were generated.

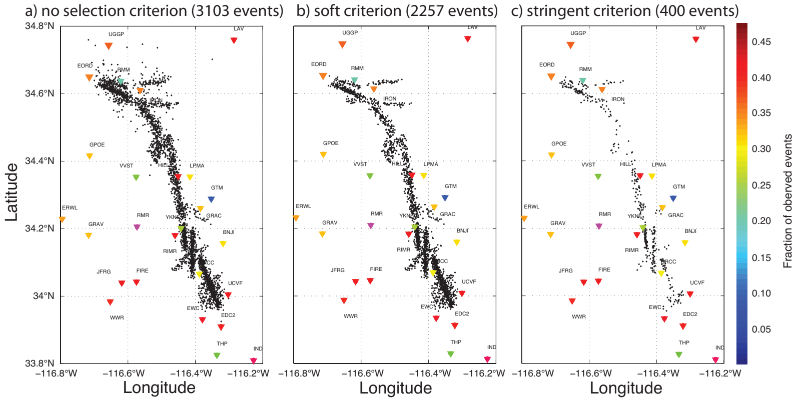

a) no selection criterion (3103 events) b) soft criterion (2257 events) c) stringent criterion (400 events)

Figure 3.8: Synthetic data derived from the Landers fault network of (Ouillon et al. 2008) and real station network. a) 3103 events in the whole data set. b) 2257 events under the soft data selection criterion. c) 400 events under the stringent data selection criterion.

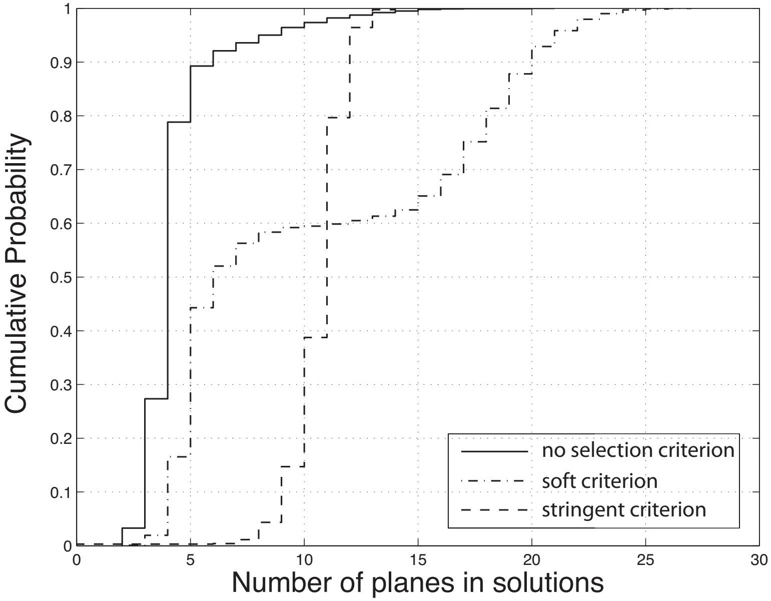
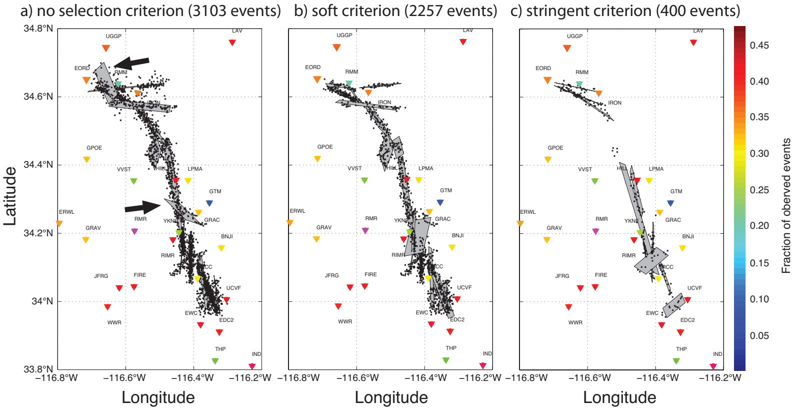


Figure 3.9: Cumulative Distribution Function (CDF) of the number of clustered planes for the synthetic Landers data set consisting of 13 faults and 3103 events. The clustering method was applied to three data sets under different data selection criterion.


a) no selection criterion (3103 events)

b) soft criterion (2257 events)

c) stringent criterion (400 events)

Figure 3.10: Result chosen by the BIC of the clustering method applied to the synthetic Landers data consisting of 13 faults. With no data selection criterion, 15 planes were generated. Under the soft selection criterion, 18 planes were clustered. Under the stringent criterion, 10 planes were generated.

Chapter 4

Fault network reconstruction and event-size distributions analysis of the Basel induced seismicity sequence

4.1 Introduction

For the Deep Heat Mining project, a private/public consortium stimulated the reservoir in Basel, Switzerland, to establish an Enhanced Geothermal System (EGS) with the aim to supply the region with an alternative source of energy. From December 2nd to 8th, 2006, approximately 11500 m^3 of water were injected into a 5km deep well at the site of Kleinhüningen (Häring et al. 2008). The experiment in Basel was closely monitored by a six-sensor borehole array which recorded more than 11,000 events of which over 3,500 could be located (Deichmann and Giardini 2009; Häring et al. 2008). The located events range from moment magnitudes M_w 0.1 to 3.2, with three events above M_w 3.

Enhanced Geothermal Systems generally use and injection and an extraction borehole to circulate the water through the reservoir. Thus, the first stimulation facilitates the definition of the site of the second borehole. The induced seismicity is thought to indicate the direction of the fluid-flow as earthquakes occur along the cracks or fault planes that cannot resist the applied stresses anymore because of normal stress reduction by increasing pore pressure. Anyway, if planes that are favorably oriented compared to the background stress field are certainly prone to activation, the role of pore pressure diffusion in the rupture along misoriented planes is still a matter of debate.

Due to the close monitoring of the seismicity with the borehole array, high-quality hypocenter locations (T. Kraft, pers. Communication) that include detailed uncertainty information and a well-defined focal mechanism dataset exist (Deichmann and Ernst 2009). With these datasets, Deichmann et al. (2013) attempt to define active fault planes using their location and clustering techniques, as well as their expertise on the tectonic setting of the Basel volume. Despite the logic procedure Deichmann et al. (2013) followed, the result is depending on the personal expertise and a process that involves detailed hands-on analysis without objective criteria. In contrary, Wang et al. (2013b) proposed the Anisotropic Clustering of Location Uncertainty Distribution (ACLUD) method to reconstruct fault networks from high-quality datasets, considering the individual hypocenter uncertainties as described by a probability density function or a scatter density cloud. The method has been tested on synthetic and real datasets on the scale of a M7-type earthquake, using also network quality criteria to investigate the influence of poorly located events on the resulting reconstructed fault network (Wang et al. 2013a).

The Basel induced seismicity sequence provides for this method a unique dataset to apply the fault network reconstruction method at a scale of a 1km^3 volume, thus on a scale that is by a factor of 10,000 smaller than the volume that was investigated before. With this new application, we address various questions: Does the automatic fault network reconstruction lead to a network solution that is compatible with pre-existing natural fractures? Does the method allow to reproduce the complexity of the detailed analysis by a hands-on approach as performed by Deichmann et al. (2013)? Is the approach able to indicate the principle directions and does it also resolve active planes that are not oriented preferentially to a homogeneous background stress field?

Provided that the method generates fault networks that resemble the current physical and mechanical understanding of reservoirs, we further investigate the b-value of the Gutenberg-Richter relation (Gutenberg and Richter 1944) for each fault plane. The b-value is one of the important parameters within the seismic hazard assessment in Geothermal systems. We thus want to check whether we find a similar pattern as Bachmann et al. (2012), who find a radial decrease of b-values from the well-head where fluids are injected, to the outer parts of the volume, so that the largest events all occur on the outer boundary of the seismicity cloud where the b-value is closer to the usual tectonic b-value.

In this chapter, we first describe the datasets and then shortly review the Anisotropic Clustering of Location Uncertainty Distribution (ACLUD) method. We then present the results of the clustering technique and interpret them through the frequency-magnitude size distribution parameter b, discussing the possible implications and further use of the applied method.

4.2 Data description

We use the high-precision earthquake catalog of T. Kraft (ETH Zurich, personal communication) to automatically reconstruct the fault network in the stimulated reservoir. The catalog has been obained by first identifying clusters of events with highly similar waveforms. In a second step master events of each cluster were relocated relative to each other using a double-difference relocation technique (Waldhauser and Ellsworth 2000). Finally, events within each clusters were relocated using a master-event relocation technique (Console and Digiovambattista 1987; Deichmann and Garciafernandez 1992). Arrival times for all observations were improved by using cross-correlation measurements (Rowe et al. 2002). The combination of cross-correlation measurements and relative relocation techniques yields high-precision earthquake locations with associated uncertainties of a few meters (Richards et al. 2006). In total the earthquake catalog contains 1,915 events in the period December 2nd 2006 to March 30th 2007, out of the about 10,500 events initially located (Deichmann and Giardini 2009; Häring et al. 2008).

Location uncertainties where computed using a Monte-Carlo approach using picking uncertainties of between 1 ms and 2 ms.

We represent location uncertainties by means of density scatter clouds, as needed for the ACLUD method. As we discussed in Chapter 2, ACLUD uses the covariance matrix and the associated 68% confidence ellipsoid derived from the scatter density cloud to compute the expected square distance between the location uncertainty probability density function (PDF) and a fault plane, and to compare the size of the average confidence ellipsoid in the direction normal to the cluster with the thickness of each cluster. Average location uncertainty, as computed by taken the mean of all three half-axes of the 68 % confidence ellipsoid, is about 4 m (Figure 4.2). This is comparable to the results from Deichmann et al. (2013).

The 68% confidence ellipsoid is a viable approximation of the location uncertainty only if the PDF is linear in shape (Wang et al. 2013a). We, therefore, selected only events that were recorded by all six borehole stations of the Basel network operated by Geothermal Explorers Ltd. (Häring et al. 2008). In order to have an optimal station configuration we required as well a P- and a S-wave observation at each station. Thus, we only selected earthquakes with a total of 12 observations. This selection yields a subset of 1100 events (see Figure 4.1). There are three events with Mw≥2.0 marked as circles, i.e. one Mw=2.0 on 2006-12-06, one Mw=2.1 and one Mw=2.2 on 2006-12-08. The largest magnitude events of the entire sequence do not remain within the catalog as these did not fulfill the stringent cross-correlation coefficients required by the relocation procedure, i.e. they did not belong to any of the earthquakes clusters.

We used the focal mechanism catalog by Deichmann et al. (2013) for the validation procedures of the reconstructed fault networks by ACLUD. Focal mechanisms were determined by first-motion polarities targeting events ($0.7<=M_L<=3.4$) that were recorded not only by the six borehole stations but also at the local surface station network to ensure high-quality solutions with a small azimuthal gap (see also Deichmann and Ernst 2009). The focal mechanism catalog contains only 185 events, not sufficient to apply the clustering approach to this dataset. Thus we decided to only use this as an independent validation dataset. It should be noted that only 49 events out of the 185 events were also in the earthquake catalog of T. Kraft. Hence, we had a set of 49 events with focal mechanisms to calculate the validation metrics.

4.3 Anisotropic clustering of location uncertainty distributions (ACLUD)

Ouillon et al. (2008) proposed the optimal anisotropic data clustering (OADC) method to reconstruct the active part of a fault network from the spatial location of earthquake hypocenters. It is inspired from the seminal k-means method (MacQueen 1967), which partitions a given dataset into a set of (*a priori* isotropic) clusters by minimizing the global variance of the partition. Ouillon et al. (2008) generalized this method to the anisotropic case with a new algorithm, which, in a nutshell, fits the spatial structure of the set of events with a set of finite-size plane segments. The number of segments used increases until the residuals of the fit become comparable to the average hypocenters location uncertainty. One can then estimate the position, size and orientation of each plane segment. More details on the OADC method can be found in Ouillon et al. (2008).

The main shortcoming of the OADC method is the implicit assumption that location uncertainties are uniform and isotropic for the whole catalog. This implies that location uncertainties are equal in all directions and identical for all earthquakes in a catalog. Obviously, this assumption is rather unrealistic since location uncertainties depend strongly, due to picking and velocity model errors, on station network geometry (Wang et al. 2013a). To improve on the OADC method, Wang et al. (2013b) introduced the anisotropic clustering of location uncertainty distribution (ACLUD) method to reconstruct active fault networks. The ACLUD method extends the OADC method by taking into account the detailed and individual location uncertainties of each event. This is achieved by introducing the expected squared distance between a probability density function and a finite plane to associate earthquake locations as described by the probability density function and the closest plane (Wang et al. 2013b). Furthermore, the ACLUD method introduces a dynamic stopping criterion that links the average location uncertainty in the direction normal to a given earthquake cluster and the thickness of that cluster. This allows one to adapt locally the resolution of the fit to the location uncertainties.

Another advantage of the ACLUD method is that it allows for a massive search through the entire solution space of possible reconstructed fault networks (Wang et al. 2013b). This is achieved by computing a large number of solutions. Since search for solution is a stochastic process, thus different runs converge to different solutions. The full set of potential solutions is submitted to six different validation procedures in order to select optimal solutions (Wang et al. 2013b). Two of the validation steps (cross-validation and Bayesian Information Criteria (BIC)) are purely statistical approaches that process the data fit of all solutions. The four other validation procedures use independent information from observed focal mechanisms to identify solutions that provide the best agreement.

4.4 Event size distribution

The cumulative number of earthquakes, N, in a given volume generally follows a power law distribution and can be expressed as

$$\log_{10} N(M) = a - bM$$

where a and b are constants that describe the productivity and the relative size distribution, respectively (Gutenberg and Richter 1944). Higher b-values indicate more small events relative to larger events, and vice-versa.

We estimate the b-value with the maximum-likelihood technique

$$b = \frac{\log_{10}(e)}{\left[\langle M \rangle - (M_c - {}^{\Delta M_{bin}}\!/_{2})\right]}$$

with $\langle M \rangle$ being the mean magnitude of the sample, ΔM_{bin} the bin width and M_c the magnitude of completeness. The completeness is estimated from the data sample using the MAXC-approach by Woessner and Wiemer (2005), adding a 0.2 increment to the algorithmic solution as the MAXC approach often tends to slightly underestimate the actual completeness level.

4.5 Clustering result

Automatic fault network reconstruction with the ACLUD-method is a stochastic fitting procedure that depends on the initial conditions. Depending on the initial condition and due to the highly non-linear process, the fault network evolves by adding complexity to the network when the dynamic stopping criteria are reached. Increasing complexity refers to adding more fault planes to the system and reassessing the overall fit. We performed 6,000 runs in order to sample the complex solution space, with each solution corresponding to one reconstructed fault network.

The reconstructed fault networks converge to a number of three-to-thirteen faults planes (Figure 4.3) to explain the hypocenter distribution. This result indicates that automatic fault plane reconstruction leads to stable results within a non-unique solution space and sample the possible complexity, with the most likely networks (90%) being constituted by six-to-nine individual fault planes. Using the six validation procedures, we selected the best solutions for each of those (Figure 4.4 and Figure 4.5). Fault planes and the events that formed the cluster to which a plane was fitted are colored correspondingly, with the fault planes made transparent to be able to view the seismicity. The colors correspond to estimated *b*-values of the events cluster of each fault plane. Cold colors indicate high b-values, hot colors indicate low b-values. Three events with $M_W \geq 2.0$ are marked with larger circles (not on scale) on all solutions except for the solution selected with the Bayesian Information Criterion (BIC) – obtained by selecting 5% of the data as the validation dataset. In this case, the best solution is found for a subset of data that does only contain two of the larger events.

At first, we focus on the geometry of the solutions. The six best solutions show quite diverse orientations resembling the complexity of the underlying fault network. The orientation of the fault planes is visualized in map-view (Figure 4.5) and by plotting the planes on stereonet projection as well as with the normal vectors to the fault planes, to illustrate the strike and dip of the faults, colored again according to the *b*-values. (Figure 4.6, Faults with unreliable b-values are black). It is not a surprise to obtain diverse networks: (i) considering the variety of natural planar structures in the crystalline basement of the area (see Figure 3, in Häring et al. 2008), and (ii) given the validation criteria that weight different properties of the data set. The solutions indicate, however, a consistent picture of steeply dipping faults (see Table 4.1-4.6) and the majority of fault planes oriented in a NNW-SSE direction. This corresponds to the orientation of the maximum compressive stress axis (S_{Hmax} =144° ± 14°) of the background tectonic stress field in this area.

Compared to the solution of Deichmann et al. (2013), the automatic solution samples the same diversity of orientations. The philosophically closest validation scheme to the approach by Deichmann et al. (2013) is the σ_{fault} solution which provides a measure of fault mechanisms homogeneity averaged over all faults. However, it corresponds to a complex solution featuring 12 planes and very few events per plane. We therefore rather prefer the solution with the σ_{event} validation scheme, as this weights equally the focal mechanisms throughout the entire cloud, resulting in only 6 faults to fit the seismicity.

4.6 Analysis of event-size distribution

Studies of micro-earthquakes on faults (Schorlemmer and Wiemer 2005) have shown that the b-value, when mapped with high quality data at high resolution, varies in the Earth's crust over distances of a few kilometers or less. These studies, combined with the analysis of regional and global focal mechanism data (Gulia and Wiemer 2010; Schorlemmer et al. 2005) as well as laboratory work (Amitrano 2003) indicate that the b-value is inversely proportional to the differential stress σ_D and thus may qualitatively be used as a stress meter at depth in the Earth's crust, where generally no direct measurements are possible. In particular, Bachmann et al. (2012) studied the distribution of the Basel seismicity cloud and inferred that there are high b-values (up to 2) close to the borehole, decaying toward the outside down to values more around 1, a value that is expected within general tectonic settings, implying that larger magnitude events are relatively more likely to occur on the edges of the stimulated reservoir rather than at the locations with the highest fluid-pressures.

Using the reconstructed fault network, we investigate the event-size distribution per plane to understand the characteristics of our approach in terms of a possible relationship to the underlying physics and its implications for seismic hazard due to an Enhanced Geothermal systems.

The entire seismicity cloud contains 1110 events after selecting only the high quality data. Each final solution contains a subset of this, with mainly around 1040 events, thus about 50 events are not clustered. As one example and to define a base average event-size distribution, we selected the final data used in the clustering procedure for the σ_{event} validation scheme, finding an average b-value of b=1.91 ± 0.11 (Figure 4.7). The b-values of the singles planes range overall between $1.05 \leq b \leq 6.2$ (Figure 4.8), with a concentration of b-values around 1.9. The figure separates the validation criteria on the y-axis and plots the planes according to their b-value. The standard deviation of the maximum likelihood estimate of the b-values are plotted in the y-direction, on scale with the b-values on the x-axis, only for a better visualization. Some of the b-values and standard deviations are large, as many of the fault planes have only a few number of events above the completeness magnitude threshold to fit the power-law. Therefore, we show the same plot as a function of the number of complete events used for the b-value fit (Figure 4.9). Few data results in large uncertainties, thus many of the b-values computed are not reliable. Based on previous studies on synthetic catalogs, we assume that about 50 events lead to reasonably stable results. We find that the highest and lowest results arise from very small clusters of events – thus these are not reliable. For clusters that have more than 50 events to fit, the values all range around b=1.9.

As we discussed before, we colored the fault planes according to the estimated b-values (Figure 4.6), excluding all those planes featuring less than 50 events above the completeness threshold as the quality criterion. We find that most of the reliable values indicate b-values around 1.9, thus do not show large variability throughout the cloud. All these planes are generally oriented NNW-SSE with some variations. Planes with very different orientations have in general to few events to estimate reliable b-values. Thus, we find that most of the events according to our results occur on planes that are oriented favorably within an assumed homogeneous background stress field and that most of the ruptures occur on such fault, yet with an event-size distribution that is strongly influenced by the fluid-pressures.

4.7 Discussion and Conclusion

In this chapter, we have applied the automatic fault network reconstruction method (ACLUD) to one of the available datasets of the Basel induced seismicity cloud. Similarly to the result from Wang et al. (2013b), each of the six validation techniques yields a different solution. The fact that these validation techniques yield different selected solutions may come from the interplay of two main factors: the multiscale structure of individual faults and the spatial extent of earthquakes location uncertainties. Many studies show that faults feature a complex inner structure consisting of a complex subnetwork of sub-faults and secondary brittle structures (Tchalenko 1970; Tchalenko and Ambraseys 1970). If the time span of the catalog is much shorter than the typical time scale necessary to activate rupture on every substructure, then most of the sub-faults will feature very few events, precluding their detailed reconstruction. Furthermore, if location uncertainties are larger than the typical spacing of sub-faults, the solution to the fit of the full network is not unique either and different validation techniques will favor different solutions (more discussion, see Wang et al. 2013b).

In this chapter, we showed that the method can handle datasets at very different scales, given that high-quality data in terms of location and uncertainty description is given. In an exercise not described in this chapter, we arbitrarily increased the uncertainty information by a factor of 10 and thus found generally only one plane per validation scheme. With the high-quality data, the method reproduces the actual complexity of the fault network on the scale of a 1km^3 volume in a relatively fast manner. The results are comparable to those obtained by Deichmann et al. (2013). However, since the latter are based on expert inference, these are not comparable quantitatively.

Given the six validation schemes, we favor the σ_{event} validation scheme as this reproduces the major orientation of the fault structure knowing the overall homogeneous background stress field orientation (Häring et al. 2008). However, since there is strong evidence of local stress heterogeneities, the other networks resemble equally likely solutions and we are not able to falsify any of the solutions.

Our results imply that fast automatic fault network reconstruction need to rely on high-quality, cross-correlation located data with an adequate uncertainty description. We have used only events with 6 P-wave and 6 S-wave picks for which the locations can be computed with high accuracy. Only with such data, it is possible to resolve small-scale structures. The complexity of the fault network is very likely sampled by all the validation schemes, and there is independent evidence from Deichmann et al. (2013) and as well by the mapping of natural fractures (Häring et al. 2008) that the complexity is real. We show, however, that major orientations are sampled and can be resolved automatically which might indicate industrial use for siting an extraction borehole.

Using the hypothesis that b-values can be used as stressmeters, we mapped the b-values on the fault planes. Many of the faults have too few events to reliably compute this value, however, setting a threshold of at least 50 events to be fit, we find that the values range around 1.9. The structures on which the b-values can be computed are generally oriented favorably to the background stress field. In contrast to Bachmann et al. (2012), we do not find strong variations with distance to the injection well head. This is however, not unexpected since our sampling technique on the faults cross-cut along the entire seismicity cloud, while Bachmann et al. (2012) use a technique to sample the local events and separate largely distant events. Fault planes that are not favorably oriented to the assumed homogeneous background stress-field seem to show different b-values, however, the number of events that could be used within the clustering approach is not large enough to provide a conclusive statement and to make inferences on the stress-field on a solid ground.

Within this analysis, we also faced some major drawbacks due to the available data: the catalog by T. Kraft is genuine as it contains high-quality hypocenter locations due to the strong requirements in the cross-correlation approach, the full pick information and hypocenter location uncertainty information. However, due to the cross-correlation constrains, many events are missing and not included in the data set, e.g. the three largest magnitude events that caused the termination of the project. Combining the Kraft-catalog with the Deichmann-catalog (Deichmann et al, 2013) would result in a catalog with different quality locations, so this is not an option. We only did this for the focal mechanisms by using the event IDs. For our type of analysis, we would desire a catalog that is located relative to the well-head, and then use a cross-correlation technique with possible additional secondary master events that include as much as possible events. One could even use multiple of such dataset that are produced with varying constraints on the waveform cross-correlation requirements to understand this as quality measure for the reconstructed networks.

One of the major questions for EGSs arise due to the related seismic hazard: Bachmann et al. (2011) proposed a method to estimate hazard in a one-dimensional approach. However, the clustering approach outlines a pathway to implement a more complex structure as it samples activated structures. Combining all solutions in a Bayesian Framework (Marzocchi et al. 2012) might provide a better spatial resolution of the associated hazard and help to mitigate the seismic hazard for future EGS projects.

Table 4.1: Matrix of strike and dip value for solution favored by BIC, b-value, standard deviation, number of events.

	Plane	Strike(°)	Dip(°)	b-value	std-b	Nr. events
BIC	1	351	70	2.70	0.31	95
	2	158	82	1.85	0.13	399
	3	95	56	5.21	1.32	13
	4	151	79	2.04	0.16	264
	5	211	45	1.83	0.57	14
	6	181	80	1.69	0.17	158
	7	31	79	1.55	0.16	106

Table 4.2: Matrix of strike and dip value for solution favored by cross validation, b-value, standard deviation, number of events.

	Plane	Strike(°)	Dip(°)	b-value	std-b	Nr. events
Cross validation	1	336	80	1.94	0.18	165
	2	206	84	1.75	0.21	62
	3	149	83	1.52	0.16	98
	4	153	83	1.85	0.12	372
	5	160	78	2.00	0.20	258
	6	353	66	1.67	0.25	100

Table 4.3: Matrix of strike and dip value for solution favored by μ_{event}, b-value, standard deviation, number of events.

	Plane	Strike(°)	Dip(°)	b-value	std-b	Nr. events
μ_{event}	1	145	89	1.42	0.24	77
	2	95	56	5.21	1.32	13
	3	159	77	1.81	0.09	643
	4	321	78	2.05	0.23	135
	5	154	83	1.55	0.17	82
	6	179	39	3.23	0.37	91

Table 4.4: Matrix of strike and dip value for solution favored by μ_{fault}, b-value, standard deviation, number of events.

	Plane	Strike(°)	Dip(°)	b-value	std-b	Nr. events
μ_{fault}	1	164	89	1.41	0.22	55
	2	137	75	2.27	0.33	25
	3	166	82	1.64	0.21	112
	4	151	85	1.81	0.09	569
	5	266	41	1.61	0.21	60
	6	321	88	2.04	0.65	12
	7	213	55	5.21	1.32	12
	8	172	59	2.22	0.22	196
	9	14	45	1.89	0.45	20

Table 4.5: Matrix of strike and dip value for solution favored by σ_{event}, b-value, standard deviation, number of events.

	Plane	Strike(°)	Dip(°)	b-value	std-b	Nr. events
σ_{event}	1	4	41	1.87	0.38	58
	2	15	64	1.70	0.21	147
	3	154	86	1.93	1.26	21
	4	154	80	1.91	0.15	267
	5	171	83	1.46	0.25	41
	6	335	81	2.03	0.19	171
	7	148	82	1.77	0.35	35
	8	151	80	2.38	0.26	168
	9	168	55	2.12	0.21	95
	10	326	86	1.48	0.36	48

Table 4.6: Matrix of strike and dip value for solution favored by σ_{fault}, b-value, standard deviation, number of events.

	Plane	Strike(°)	Dip(°)	b-value	Std-b	Nr. events
σ_{fault}	1	253	50	6.20	1.77	12
	2	25	53	2.13	0.48	34
	3	189	81	1.85	0.26	89
	4	153	77	2.06	0.16	235
	5	148	81	1.73	0.11	424
	6	309	90	2.90	0.81	14
	7	312	23	3.58	1.24	26
	8	348	77	2.07	0.18	90
	9	180	86	1.47	0.36	23
	10	300	45	1.06	0.16	36
	11	162	63	2.21	0.52	23
	12	146	59	1.28	0.20	33

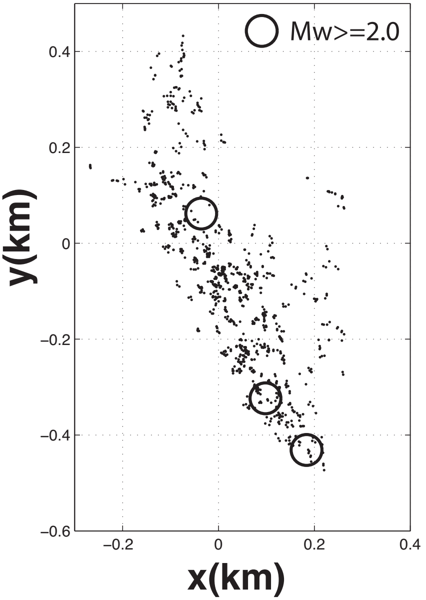


Figure 4.1: 1110 events observed by all 6 borehole stations with both P phases and S phases. Three large events (Mw>=2.0) are marked in circles, one Mw=2.0 on 2006-12-06, one Mw=2.1 and one Mw=2.2 on 2006-12-08.

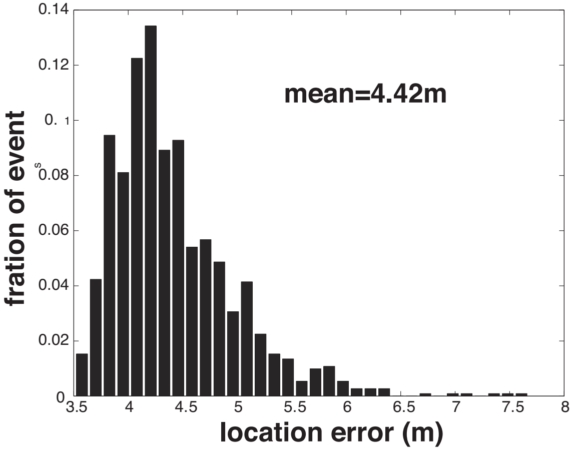


Figure 4.2: Histogram of location error for 1110 events used in this study. Assuming location uncertainty isotropic, one sphere and its radius are used to approximate the 68% confidence ellipsoid and overall location error for each event.

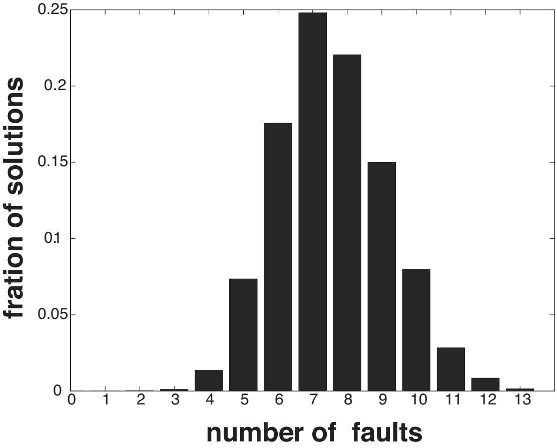
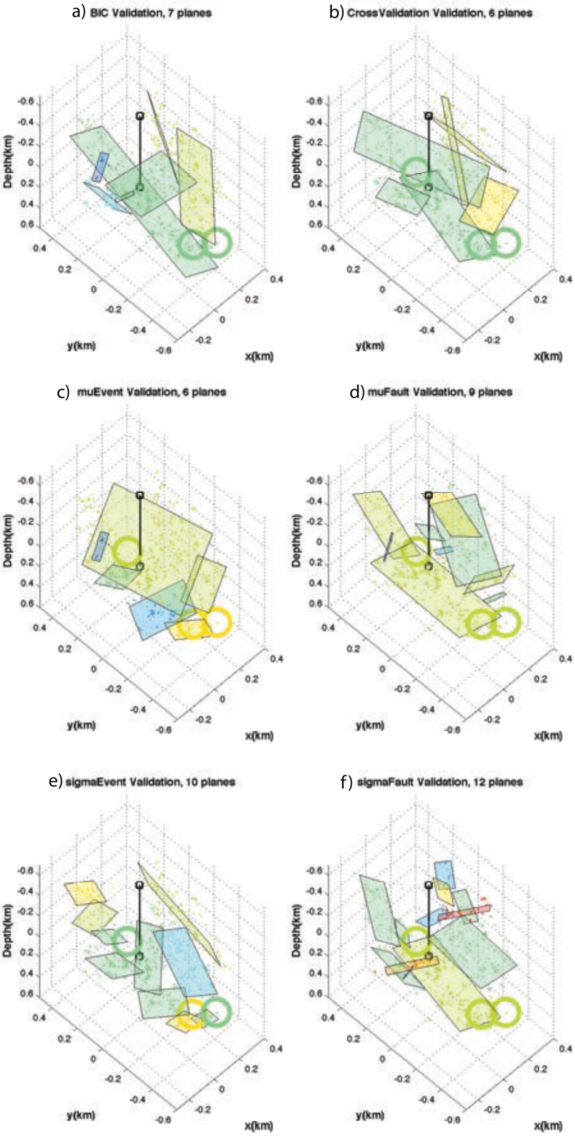


Figure 4.3: Histogram of the number of clustered faults.



a) BIC Validation, 7 planes

b) CrossValidation Validation, 6 planes

c) muEvent Validation, 6 planes

d) muFault Validation, 9 planes

e) sigmaEvent Validation, 10 planes

f) sigmaFault Validation, 12 planes

Figure 4.4: 3D-view of the reconstructed fault networks per validation procedure. Faults and their assigned events are colored correspondingly to b value in Figure 4.8. Events with Mw≥2 are indicated as large circles.

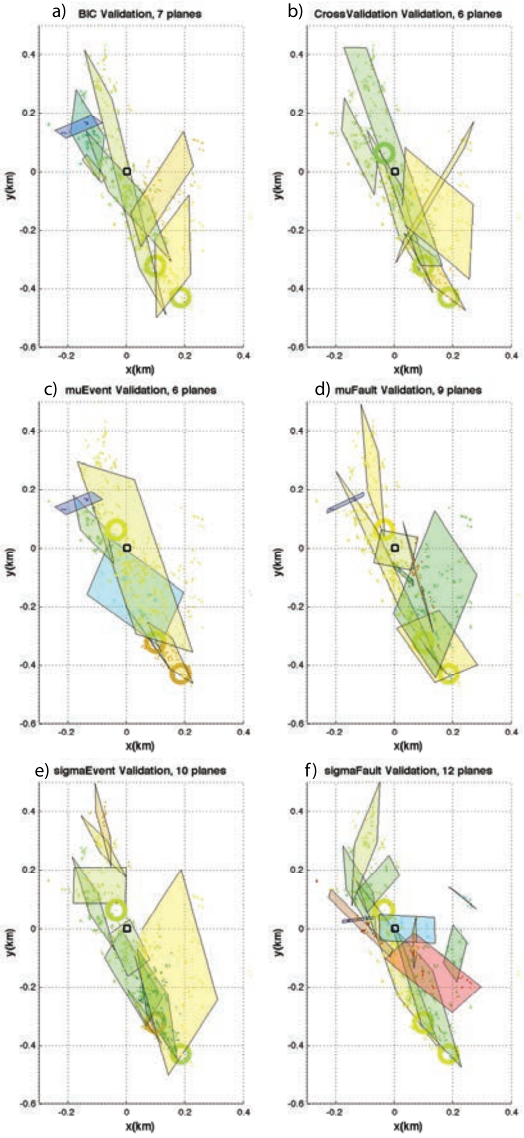

a) BIC Validation, 7 planes
b) CrossValidation Validation, 6 planes
c) muEvent Validation, 6 planes
d) muFault Validation, 9 planes
e) sigmaEvent Validation, 10 planes
f) sigmaFault Validation, 12 planes
y(km)
x(km)
0.4
0.2
0
-0.2
-0.4
-0.6
-0.2
0
0.2
0.4

Figure 4.5: Map-view of the reconstructed fault networks per validation procedure. Faults and their assigned events are colored correspondingly to b value in Figure 4.8. Events with M_W≥2 are indicated as large circles.

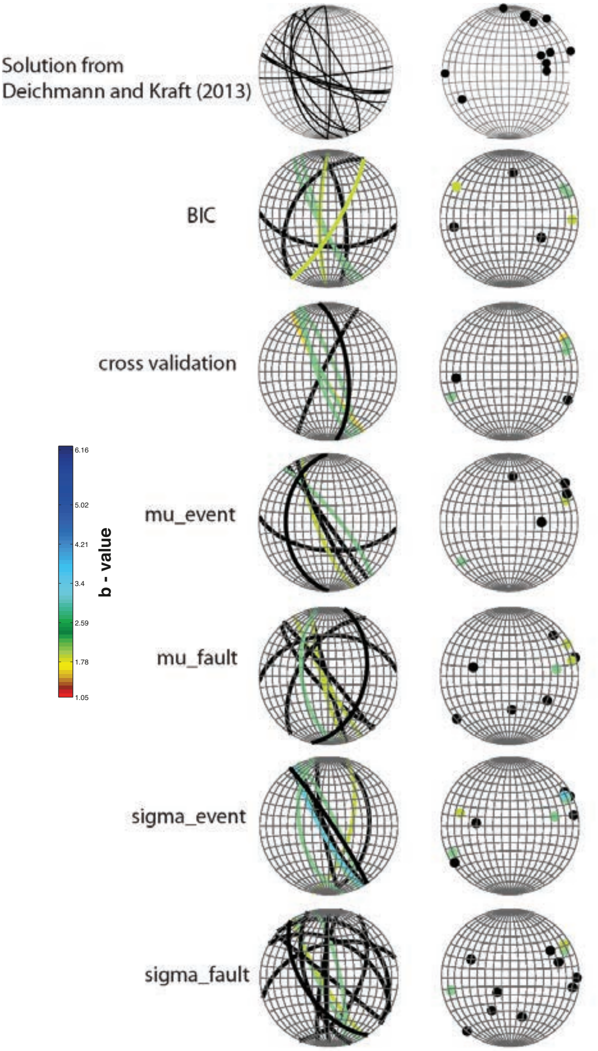

Solution from
Deichmann and Kraft (2013)
BIC
cross validation
mu_event
mu_fault
sigma_event
sigma_fault
6.16
5.02
4.21
3.4
2.59
1.78
1.05
b - value

Figure 4.6: Stereo plots of the fault network from Deichmann et al. (2013) and solutions favored by six validation criteria. Curves in the left column indicate the orientation of fault traces. Dots in the right column show directions of the normal poles of fault planes. Faults with >= 50 events are colored corresponding to b value in Figure 4.8.

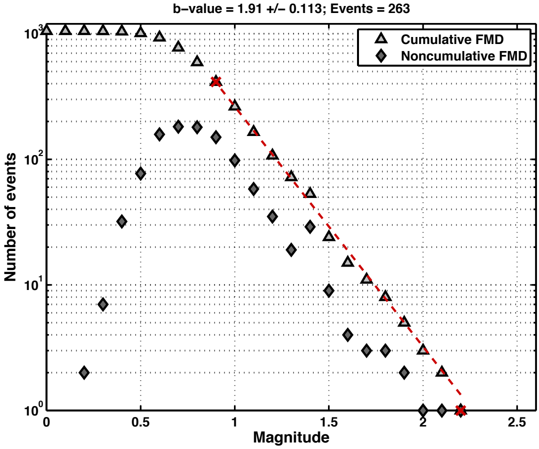


Figure 4.7: b-value for the solution of sigma_event for the entire seismicity.

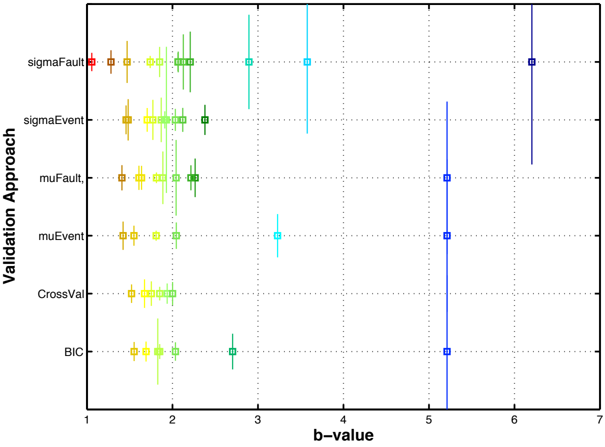


Figure 4.8: b-value of each fault plane for all solutions chosen by six validation criteria. Uncertainty of b-value is presented in vertical direction. Color corresponds to b-value.

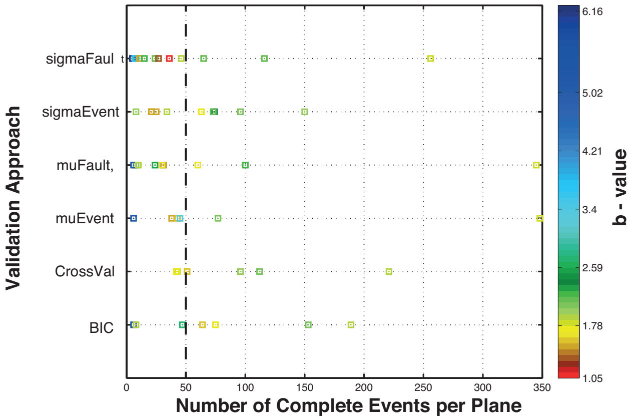


Figure 4.9: Number of events in each fault plane for all solutions chosen by six validation criteria. Color corresponds to Figure 4.8.

Chapter 5

Discussion and conclusions

In this thesis, I introduced a new anisotropic statistical clustering technique, the Anisotropic Clustering of Location Uncertainty Distributions (ACLUD) method, to reconstruct active fault networks from observed seismicity. This method significantly improves the Optimal Anisotropic Data Clustering method (OADC) previously developed by Ouillon et al. (2008) by utilizing the location uncertainty information within the clustering and hence the fault reconstruction procedure. Due to its dependence on the location uncertainty information, I investigated the influence of hypocenter location accuracy and precision on the reconstruction result in detail, based on the analysis of seismic network quality criteria. At the same time, I refined the seismic network criteria definition for seismic networks that locate events within the extent of the network. I applied the method to multiple synthetic datasets on the scale of an M7-type earthquake, i.e. choosing the 1992 Landers (CA) earthquake in Southern California, and on the much smaller scale of a 1km^3 volume due to induced seismicity during the Basel Enhanced Geothermal System (EGS) Project. In all cases, I find complex fault networks as the result of the approach, that indicate possible solutions depending on the validation scheme.

I achieved major improvements compared to OADC in both, the training phase, i.e. when performing the fit of a given dataset or subset, and the validation phase, i.e. when quantifying the ability of the solution to explain another set or subset of independent data. In the training phase, the new method accounts for individual location uncertainties of each event through their expectation and covariance matrix. This allows us to elegantly compute a distance between a given event and a given plane (or rather its expected squared value), so that each event can be associated to one plane; this also allows for full control over the local quality of the reconstruction as the method automatically generates more complex fault network structures wherever the local residuals of the fit becomes larger than the locally estimated location uncertainties. A more complex structure in this sense implies to introduce more fault planes to fit the observed seismicity. Keeping its inner k-means computation dynamics and stochastic scheme, the fitting method is strongly nonlinear, thus different runs generally converge towards different local minima of the residuals.

This observation resembles the well-known influence of human subjectivity when interpreting purely geometrical datasets in geosciences. I thus found it necessary to introduce six different validation schemes in order to select the best solution: two of them based on the residuals of the fit, a simple cross validation scheme and a Bayesian information criterion: the four others are based on the compatibility of the fault network with observed focal mechanisms, checking for consistency when comparing the reconstructed fault planes with the observed potential failure planes deduced from double-couple source solutions.

Testing the method itself on synthetic examples led us to develop a fresh and integrated approach that, I believe, should be implemented more generally when dealing with the statistics of seismicity in the spatial domain. Our approach indeed incorporates the process of earthquake data acquisition and its potential influence on the results of the data clustering scheme. I thus investigated in detail the station network criteria in order to assess the quality of earthquake location for local networks, and their importance in fault network reconstructions. Our results confirm that network criteria, such as the number of observations, the primary station azimuth gap and the distance to the closest station, are highly valuable to assess location quality. Epicenter locations are well-constrained if the primary station azimuth gap is smaller than 180°, whereas well-constrained focal depth estimates require a nearby station with a distance less than 1.5 times the focal depth. It should be noted that these results only apply if the seismic velocity structure is accurately known; I show that these results do not hold when the velocity model is systematically different. Though the latter might not be true for the currently used velocity models in earthquake location, and as I am only using P-phases for the relocation in the exercise, this result is not immediately valid for current location procedures, it is however notable that locations are inherently mislocated even with a very good station coverage. I additionally show that the use of classical error ellipsoids to describe location uncertainties can be misleading for earthquakes observed at a low number of stations and with a primary station azimuth gap of more than 180°. One consequence of this is that the location errors are often underestimated and within applications, the influence of location uncertainty again is underestimated for other applications.

Investigating synthetic datasets displaying different degrees of complexity clearly showed that the effect on fault reconstruction is fundamental: using high-quality subsets of data selected by the proposed network criteria leads to high-quality fault network reconstruction results. On the contrary, using lower-quality data can lead to unstable and unreliable fault network reconstructions (even in the simplest cases) and may introduce artifacts, particularly within regions featuring a complex fault structure. I believe that such a detailed study of the dependence of statistical outcomes as a function of data quality remains to be done in various other domains such as detailed mapping of earthquake parameters (Mignan et al. 2011; Schorlemmer et al. 2004; Woessner and Wiemer 2005), building earthquake forecast models (e.g. Hainzl et al. 2010. Zhuang et al. 2011, Zhuang et al. 2012 and reference therein), or studies that evaluate earthquake forecasting models (e.g. Woessner et al. 2011; Zechar et al. 2013 and references therein).

The data selection criteria and clustering technique have then been applied to the Landers 1992 event area, which had previously been studied by Ouillon et al. (2008) and benefited both from a high quality local seismic network and a large amount of geophysical observations and field investigations. It turns out that the six validation schemes provide six different solutions. I interpret this variability as resulting from the scale-invariant complexity of fault networks; the major implication is that the smallest scales within a given network are necessarily undersampled by a finite set of seismic events. If location uncertainties are smaller than the typical scale defined by the actual sampling rate of each substructure, then the resulting best solution becomes non-unique and depends on the criterion we use.

I then investigated the performance of ACLUD on a very different scale, applying the method to seismicity observed during the Basel Enhanced Geothermal System Project. The volume samples a comparable number of events, however, on a much smaller spatial domain. The Landers volume is on the scale of a 120km long fault system, using a 10km wide swath of seismicity that reaches to a depth of about 25km crustal thickness, with earthquake magnitudes between 2 to 7.3. The Basel dataset in comparison remained within a volume of 1km^3, with magnitudes between 0 and 3.2. The results for the latter again show a variability of the obtained solution with the validation criterion, suggesting that the multi-scale mechanical segmentation process of faults (and its associated sampling problematic) still holds within the upper crust down to very small scales. Using the hypothesis that b-values can be used as stressmeters, I mapped the b-values on the fault planes. Many of the faults have too few events to reliably compute this value, however, setting a threshold of at least 50 events to be fit, I find that the values range around 1.9. The structures on which the b-values can be computed are generally oriented favorably to the background stress field.

On a more technical aspect, the ACLUD method features a dynamic stopping criterion which consists in comparing the local variance of the fit with data location uncertainty within each cluster: clustering stops once the variance is comparable to the uncertainty. In other words, the dynamic stopping criterion is minimizing the *Type I errors* in statistics. However, in practice, the effort to reduce one type of error generally results in increasing the other type of error. If we define an index as 0 when there is no fault, and 1 when there is a fault, we then can describe *Type I error* and *Type II error* as shown in Table 5.1. For both cases where (Reality=0, Reconstruction=0) and (Reality=1, Reconstruction=1), the decisions made by ACLUD are correct. However, if Reality=0 but Reconstruction=1, i.e. in reality there is no fault but ACLUD reconstructs one, ACLUD makes a *Type I error*. On the contrary, if Reality=1 but Reconstruction=0, i.e. in reality there is a fault but ACLUD fails to reconstruct it, ACLUD makes a *Type II error*. The dynamic stopping criterion minimizes the *Type I errors* by terminating clustering once variance is comparable with uncertainty. However, by doing so, it may result in increasing the *Type II errors*, i.e. ACLUD failing to reconstruct an existing fault due to data quality reason. Given a fixed number of data, there is no simple way to reduce both types of error.

In cases of poorly located data, one may get a too coarse fault network solution. Inspired by what we discussed above, we may want to relax the dynamic stopping criterion and allow the algorithm to further explore the potential fault reconstruction solution space by comparing the variance of the fit with a predefined value, e.g. 60th quantile of the local data uncertainty. Indeed, by doing so, it would result in increasing the *Type I errors* due to over-fitting data and reducing the *Type II errors*, which allows one to reconstruct the potentially existing faults. Thus a careful assessment of the quality and reliability of the reconstructed fault networks, involving a comparison with a prior model (coming from structural geology investigations, for instance), is needed.

The latter comment suggests some further improvements of the method by incorporating external geological or geophysical knowledge within the training phase, whereas we currently take it into account in the validation phase only. For instance, during the inversion, the complexification step may be still be controlled by the local variance of the fit and location uncertainties, but also by the dispersion of orientations of the observed focal mechanisms, maybe also helped by criteria based on the similarity of waveforms recorded at nearby stations or the orientation of nearby faults compiled in catalogs such as the California Fault Model. This would open new doors within the field of pattern recognition techniques, as we would use genuine physical parameters to identify more directly the singularities of the strain field within the complex rheological medium that is the Earth's crust.

There are several potential applications of the outcomes of the proposed fault reconstruction algorithm. Once a satisfying fault network is selected, we may use its architecture to quantify the spatial and temporal statistical properties of earthquakes and clusters. A first track consists in investigating the correlations of seismic activity at the fault scale to better understand the mechanics of the network as a whole. The fault network can be taken as a natural candidate to partition the data in order to analyze the spatial and temporal variations of seismicity parameters. We may then study quantities such as

- the Gutenberg-Richter b-value and the productivity term (Schorlemmer and Wiemer 2005; Wiemer and Wyss 2002) and their dependence on the styles of faulting as performed for the data set of the Basel EGS-project;
- the statistical estimation of the tails of the size distribution based on extreme value theory (Pisarenko et al. 2008; Pisarenko and Sornette 2003) ;
- the size of the possible largest event (i.e. characteristic earthquake) that may occur on a given fault segment.

Another potential application is the modeling of fault-to-fault interactions that control the overall dynamics of a fault network, contributing to a clearer picture of the intimately interwoven dynamics of earthquakes and faults. For example, there is a growing awareness and an intense research activity based on the fact that a significant fraction of earthquakes are at least triggered by preceding events. A better understanding of the links between earthquakes and faults will allow us to improve on the purely statistical models of triggered seismicity such as the ETAS model (Ogata 1988) by including more realistic geometries and tensorial information associated with the reconstructed fault networks. This will improve present attempts to develop better space-time models of earthquake triggering, which still lack information on fault localization by assuming diffuse seismicity patterns unrelated to faults (Ogata and Zhuang 2006) and their irreversible nucleation and growth processes. More mechanically based approaches may also be used together with the available focal mechanisms that provide precious insight into the slip patterns on each plane. Those patterns may in return be used to compute more precisely fault-fault interactions through the estimation of stress transfer among faults. This will allow improving the validation schemes of the forecasting techniques based on Coulomb stress interactions by incorporating even the smallest events of a catalog as their failure planes will be known at higher confidence levels.

Reality=0
Reconstruction=0
	Reality=1
	Reconstruction=0
	(*Type II error*)

Reality=0
Reconstruction=1
(Type I error)
	Reality=1
	Reconstruction=1

Table 5.1: Type I error and Type II error.

Appendix A

Supplementary Material for Chapter 2

A.1 k-means including location uncertainties (uk-means)

The k-means method assumes that the uncertainty of the spatial location of data points is negligible. This assumption holds in disciplines such as image analysis, where the coordinates of the data points are given by red, blue and green color contents at each pixel of a picture. In the case of real physical systems, the story is different. For earthquakes, location uncertainty is an inherent property due to wave arrival time inaccuracy, velocity model errors, station network geometry, or outdated data sources (historical seismicity catalogs, for instance). When taking uncertainty into account, data can no longer be described by a point-process, but by a more or less complex probability distribution function (hereafter pdf).

Chau et al. (2006) claim that location uncertainties can significantly affect the results provided by clustering techniques such as k-means. They thus introduce the *uk-means* algorithm (where the 'u' letter stands for 'uncertain'), which incorporates uncertainty information and provides, when considering synthetic samples, more satisfying results than the standard algorithm.

For the general case of a set of objects $\{\vec{O}_1, \vec{O}_2, ..., \vec{O}_n\}$ within an m-dimensional space and a set of cluster $\{\vec{C}_1, \vec{C}_2, ..., \vec{C}_k\}$, *k-means* assigns each object to the "closest" cluster barycenter according to the Euclidean distance measure $d(\vec{O}_i,\vec{C}_j)$: $\mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$, where $i = 1,...,n$ and $j = 1,...k$. However, when O_i is no longer a point but a pdf, the distance must be estimated differently. Chau et al. (2006) propose to use the expected squared distance $ESD(\vec{O},\vec{C})$, defined as the integral of the weighted square norm $\left\|\vec{O}_i - \vec{C}_j\right\|^2$ over the whole probability space of O_i. Denoting the pdf of O_i as $f(\cdot)$, we have:

$$\begin{aligned} &\forall \vec{x} \in \overrightarrow{O_i}, f(\vec{x}) \geq 0 \\ &\int_{\vec{x} \in \overrightarrow{O_i}} f(\vec{x}) d\vec{x} = 1 \end{aligned} \tag{A1}$$

Then, we define (Lee et al. 2007):

$$ESD(\overrightarrow{O_i}, \overrightarrow{C_j}) = \int_{\vec{x} \in \overrightarrow{O_i}} \left\| \vec{x} - \overrightarrow{c_j} \right\|^2 f(\vec{x}) d\vec{x} \tag{A2}$$

Where $\overrightarrow{c_j}$ is the barycenter of the cluster $\overrightarrow{C_j}$. Monte Carlo techniques prove to be too heavy to compute $ESD(\overrightarrow{O_i}, \overrightarrow{C_j})$ empirically, especially when dealing with large datasets. A simpler technique consists in using the simple theorem of variance decomposition. Lee et al. (2007) thus rewrite Eq. (A2) as:

$$\begin{aligned} ESD(\overrightarrow{O_i}, \overrightarrow{C_j}) &= \int_{\vec{x} \in \overrightarrow{O_i}} \left\| \vec{x} - \overrightarrow{k_i} \right\|^2 f(\vec{x}) d\vec{x} + \left\| \overrightarrow{k_i} - \overrightarrow{c_j} \right\|^2 \\ &= ESD(\overrightarrow{O_i}, \overrightarrow{k_i}) + \left\| \overrightarrow{k_i} - \overrightarrow{c_j} \right\|^2 \end{aligned} \tag{A3}$$

where $\overrightarrow{k_i} \in \mathbb{R}^m$ is the centroid of the spatial distribution of the uncertain object $\overrightarrow{O_i}$ and is defined as $\overrightarrow{k_i} = \int_{\vec{x} \in \overrightarrow{O_i}} \vec{x} f(\vec{x}) d\vec{x}$. By definition, $ESD(\overrightarrow{O_i}, \overrightarrow{k_i})$ is simply the variance of that spatial distribution and can be easily computed once for all from its pdf. The second term in the right hand side in Eq. (A3) is simply the square of the distance between two points in a Euclidean space. Using this new distance definition and following the same procedure as standard k-means, data featuring uncertainty information can be processed easily.

Note that when observations come without uncertainties, all pdfs variances are set to 0, so that we recover the classical version of k-means.

A.2 Generating synthetic catalogs with a simple geometry with 3 vertical planes

The general method we propose to generate a synthetic earthquake catalog is the following: we first impose the geometry of the original fault network, which consists in a collection of rectangular planes with variable locations, sizes and orientations. We then assume that all earthquakes occur exactly on those planes and generate P waves. We then compute, assuming a given velocity model, the theoretical travel times between the true hypocenters and a set of stations which locations have been predefined. Random perturbations are added to the waves' arrival times, allowing proceeding to the inverse problem: computing the location of the events as well as their uncertainties. To generate the associated synthetic focal mechanisms, we first assume that the rake of the slip vector on each plane is zero. For each event, the strike and dip are assumed to be identical to the ones of the input plane to which it belongs. We then add a Gaussian random perturbation to the strike, dip and rake of the event with a standard deviation of 10°. Those perturbed angles are then used to compute the strike and dip of the auxiliary plane, thus providing a complete focal mechanism. The inverted location catalog is then fitted with a set of finite planes, using our algorithm for 1,000 clustering runs.

Earthquakes are located using the NonLinLoc software package (Lomax et al. (2000), Version 5.2, http://alomax.free.fr/nlloc/). Compared to traditional, linearized approaches, NonLinLoc is superior in that it computes the posterior probability density function (PDF) using nonlinear, global searching techniques. The PDF represents the complete probabilistic solution to the earthquake location problem, including comprehensive information on uncertainty and resolution (Moser et al. 1992; Tarantola and Valette 1982; Wittlinger et al. 1993).

The best solution (which may depend on the validation technique) is then compared to the original input fault network. Note that in real catalogs, the origin of the location uncertainties also lies in the uncertainties about the real velocity model – an ingredient that we neglect here: the sole uncertainties stem from the wave picking process and the geometry of the stations network (Bondár et al. 2004).

The first synthetic dataset consists in 4,000 events, uniformly distributed over a network featuring three vertical planes (see Figure A2). Faults A and C have a length (along strike) of 40 km, a width (along dip) of 20 km, and feature 1,000 events each. They share a common strike of 90°E and a common dip of 90°. Fault B has a length of 100 km, a width of 20 km, and features 2,000 events. Its strike is 0°E and its dip is 90°.

We distributed a set of 88 stations on a regular grid with a spatial extent of 240 km by 180 km and a cell size of 20km. For each event, we randomly selected 11 stations out of the complete set of 88 stations as observations, and computed the theoretical arrival times, to which a Gaussian error with a standard deviation of 0.1 s was added to simulate real pickings. A simple 1-D layered velocity model was used. Using NonLinLoc, we generate a synthetic earthquake catalog consisting of 4,000 events characterized by their full pdf.

Figure A2a shows the distribution of the 4,000 relocated earthquakes, which are slightly shifted away from the original fault planes. As we use an error-free velocity model and Gaussian picking errors, location uncertainties are mainly controlled by the geometry of the stations network. Events located with a better station network coverage are likely to be characterized by a better location quality. Using the relocated data set, with our clustering technique we generated 1,000 reconstruction solutions.

The solutions are chosen by cross validation, using both μ and σ_{event} measures, and consist of three planes with similar structure, which are shown in Figure A2b. Table A1 lists the parameters of the true and the reconstructed faults. All those solutions show a nice agreement with the true fault network. Figure A2c shows the solution chosen by BIC. Fault B is divided into two sub-faults at the intersection with fault C. From other tests we performed, we also noticed that, when faults cross each other, our model has difficulties in deciding which plane one event belongs to. Yet, the structure is still nicely inverted. Figure A2d displays the solution chosen by σ_{fault}. One small fault is generated at the northern edge of fault B. This comes from the fact that locations quality close to the northern and southern edges is not as good as in other parts due to a poorer station coverage. We also performed similar tests on catalogs generated with different numbers of observations and different Gaussian picking errors, and obtained similar results.

A.3 Generating multiscale synthetic fault catalogs

The method we use to generate multiscale fault networks is largely borrowed from the concept of Iterated Function Systems (hereafter IFS), proposed by Hutchinson (1981) and popularized by Barnsley (1988), which provide a basic algorithm to generate deterministic fractal objects. IFS consist in replacing a given large scale Euclidean object by a series of replications of itself at smaller and smaller scales. In our case, this consists in segmenting the fault at smaller and smaller resolutions.

We shall first consider a vertical fault, over which events are distributed. The case of a fault with arbitrary strike and dip is solved by simply performing the necessary rotations of such a vertical fault. The fault is chosen such that its length is 1. The general case is solved by a simple scaling up (or down) to the real length of the fault.

We first consider only the trace of the fault, which is its intersection with the free surface, and will perform the segmentation along its strike. The x axis is chosen to stand along the fault's strike, the y axis is normal to the fault, and the z axis is pointing downwards.

STEP 1

Consider a fault L1 with length 1 which extremities are [0,0] and [1,0].

STEP 2

Define a group of linear applications, for example 10 different functions F_i, with i=1,2,...10. F_i is a linear transformation function which reads:

$x' = A_i x - B_i y + C_i$
$y' = B_i x + A_i y + D_i$
with $A_i^2+B_i^2=Q_i^2$.

The coefficients A, B, C and D are chosen such that this application transform a given fault segment into a downscaled, slightly rotated and offset copy of itself (so that $Q < 1$). The various parameters of the set of functions may be chosen by hand or randomly.

STEP 3

i. Choose randomly one of the N functions F_i previously defined and apply it to L_1 so that one gets a new segment S_1 and its extremities

ii. Repeat step (i) a few times (p times, for instance, with p small). Doing so, a set of new small segments S_j, with j=1,…,p, is generated. Store the coordinates of their extremities. Remove the original, large scale segment L_i.

iii. The new dataset now consists in the set [S_j]. Apply steps (i-ii) to each of its members.

iv. Iterate step (iii) a few times so that, at each iteration, the full set of newly created segments [S_j] replaces the previous one. The total number of segments thus increases after each iteration while their sizes decrease.

STEP 4

Rescale the final lengths of the segments so that the extent of the set fits within [0;1] along its average direction.

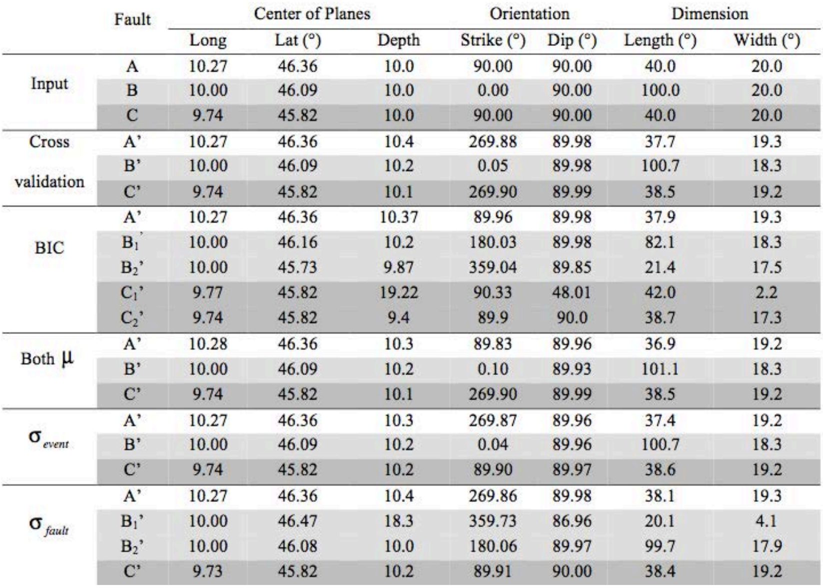
STEP 5

Apply steps 1 to 4 to generate a different segmented fault for each fault of the catalog. Rotate the segmented fault accordingly so that its average strike and dip fit with the original one.

The previous algorithm thus provides a segmentation of the original fault along its strike, but not along its dip where we leave its structure intact. But a similar process can be implemented along that direction too. Alternatively, for the sake of simplicity, we can achieve a 3D structure by extending each subplane to the same depth as the original fault. Their knowledge allows one to locate some events on those segments and build their focal mechanisms. If the total number of events generated on the whole fault is very large, then each segment will feature enough events to be fully identifiable from them. If the number of events is too small, each segment will be undersampled by the synthetic seismicity catalog, resulting in a noisy multiscale subnetwork.

Figure A3 shows an example to generate synthetic multiscale synthetic faults following the approach we discussed above. The original fault shown in Figure A3a is a structure with strike=172°, dip=81° and dimensions 28km×3km. There are 446 events located on it. In order to generate the set of multiscale synthetic faults, we randomly generated at each iteration 2 linear transformation functions to build smaller scale segments. We constrain the linear transformations so that the distribution of small faults is still along the strike direction. After 5 iterations, we finally generated a multiscale set of 32 fault segments. The number of events located on each small fault depends on its size and ranges from 4 to 71.

Table A1: Parameters of the true and the reconstructed fault networks discussed in the text.



	Fault	Center of Planes			Orientation		Dimension	
		Long	Lat (°)	Depth	Strike (°)	Dip (°)	Length (°)	Width (°)
Input	A	10.27	46.36	10.0	90.00	90.00	40.0	20.0
	B	10.00	46.09	10.0	0.00	90.00	100.0	20.0
	C	9.74	45.82	10.0	90.00	90.00	40.0	20.0
Cross validation	A’	10.27	46.36	10.4	269.88	89.98	37.7	19.3
	B’	10.00	46.09	10.2	0.05	89.98	100.7	18.3
	C’	9.74	45.82	10.1	269.90	89.99	38.5	19.2
BIC	A’	10.27	46.36	10.37	89.96	89.98	37.9	19.3
	B_1’	10.00	46.16	10.2	180.03	89.98	82.1	18.3
	B_2’	10.00	45.73	9.87	359.04	89.85	21.4	17.5
	C_1’	9.77	45.82	19.22	90.33	48.01	42.0	2.2
	C_2’	9.74	45.82	9.4	89.9	90.0	38.7	17.3
Both μ	A’	10.28	46.36	10.3	89.83	89.96	36.9	19.2
	B’	10.00	46.09	10.2	0.10	89.93	101.1	18.3
	C’	9.74	45.82	10.1	269.90	89.99	38.5	19.2
σ_{event}	A’	10.27	46.36	10.3	269.87	89.96	37.4	19.2
	B’	10.00	46.09	10.2	0.04	89.96	100.7	18.3
	C’	9.74	45.82	10.2	89.90	89.97	38.6	19.2
σ_{fault}	A’	10.27	46.36	10.4	269.86	89.98	38.1	19.3
	B_1’	10.00	46.47	18.3	359.73	86.96	20.1	4.1
	B_2’	10.00	46.08	10.0	180.06	89.97	99.7	17.9
	C’	9.73	45.82	10.2	89.91	90.00	38.4	19.2

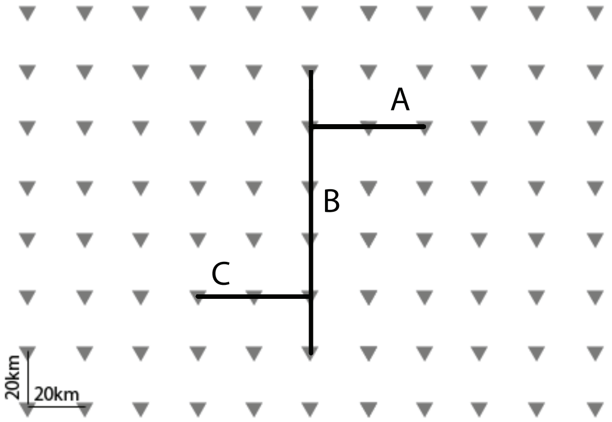


Figure A1: Map view of network design, fault location and earthquake distribution to compute synthetic data. Triangles represent stations (88 in total, 20 km spacing). For each earthquake 11 stations were picked randomly as observations. The lines indicate the fault surface traces along which earthquakes were distributed.

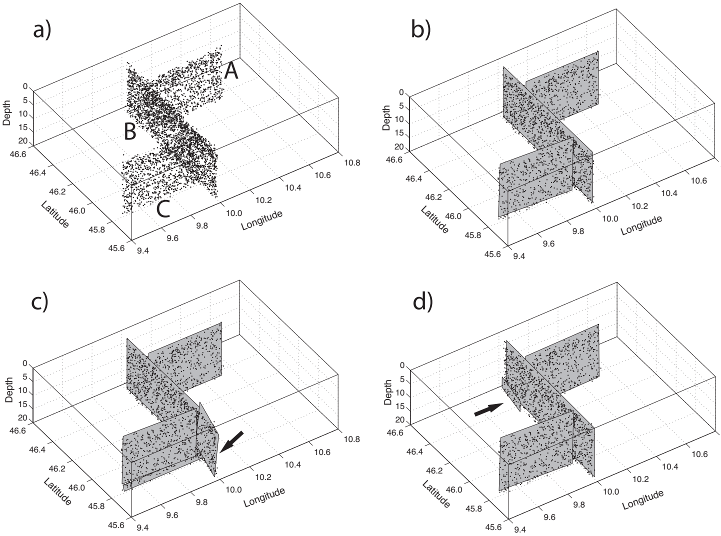


Figure A2: a) Distribution of 4,000 relocated hypocenters located on three vertical faults A, B and C. b) Reconstructed structures from cross validation, using both μ and σ_{event} measures. Three vertical faults are clustered. c) Result from BIC: Fault B is divided into two faults in the southern part. d) Result using σ_{fault}. One small fault is generated at the northern edge of fault B.

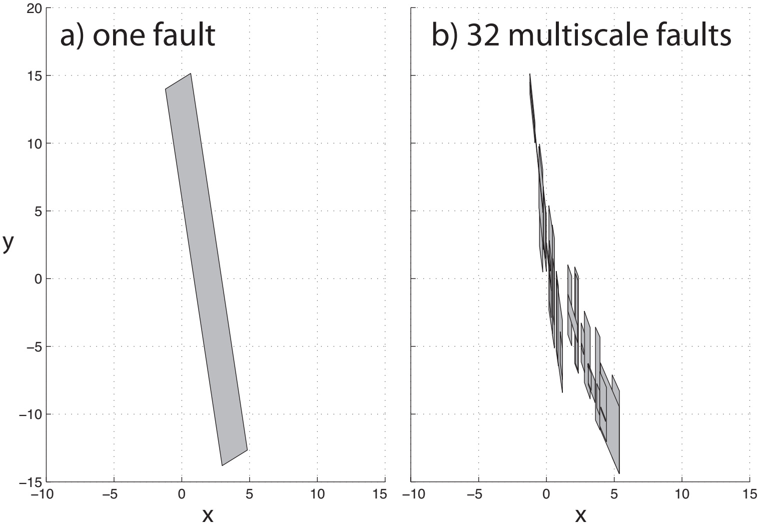


Figure A3: The original fault shown on the left has a strike=172°, dip=81° structure with dimension 28km×3km. Based on it, the multiscale faults structure consisting of 32 small segments (right) was generated following the approach discussed in the text. There are 446 events in total located on the original fault (left). For the multiscale faults, the number of events located on each fault ranges from 4 to 71 events.

References:

Amitrano D (2003) Brittle-ductile transition and associated seismicity: Experimental and numerical studies and relationship with the b value. J Geophys Res-Sol Ea 108 (B1). doi:10.1029/2001jb000680

Arnold R, Townend J (2007) A Bayesian approach to estimating tectonic stress from seismological data. Geophys J Int 170 (3):1336-1356. doi:10.1111/J.1365-246x.2007.03485.X

Arthur D, Vassilvitskii S (2007) k-means plus plus : The Advantages of Careful Seeding. Proceedings of the Eighteenth Annual Acm-Siam Symposium on Discrete Algorithms:1027-1035

Bachmann CE, Wiemer S, Goertz-Allmann BP, Woessner J (2012) Influence of pore-pressure on the event-size distribution of induced earthquakes. Geophys Res Lett 39. doi:10.1029/2012gl051480

Bachmann CE, Wiemer S, Woessner J, Hainzl S (2011) Statistical analysis of the induced Basel 2006 earthquake sequence: introducing a probability-based monitoring approach for Enhanced Geothermal Systems. Geophys J Int 186 (2):793-807. doi:10.1111/J.1365-246x.2011.05068.X

Bahmani B, Moseley B, Vattani A, Kumar R, Vassilvitskii S (2012) Scalable k-means++. Proc VLDB Endow 5 (7):622-633

Barnsley M (1988) Fractals Everywhere. Academic Press, Inc

Basili R., Kastelic V., Demircioglu M. B., Garcia Moreno D., Nemser E. S., Petricca P., Sboras S. P., Besana-Ostman G. M., Cabral J., Camelbeeck T., Caputo R., Danciu L., Domac H., Fonseca J., García-Mayordomo J., Giardini D., Glavatovic B., Gulen L., Ince Y., Pavlides S., Sesetyan K., Tarabusi G., Tiberti M. M., Utkucu M., Valensise G., Vanneste K., Vilanova S., Wössner J. (2013) The European Database of Seismogenic Faults (EDSF) compiled in the framework of the Project SHARE. http://dissrmingvit/share-edsf/. doi:10.6092/INGV.IT-SHARE-EDSF

Bishop CM (2006) Pattern Recognition and Machine Learning. Springer,

Bondár I, McLaughlin K (2009a) Seismic Location Bias and Uncertainty in the Presence of Correlated and Non-Gaussian Travel-Time Errors. B Seismol Soc Am 99 (1):172-193. doi:10.1785/0120080922

Bondár I, McLaughlin KL (2009b) A New Ground Truth Data Set For Seismic Studies. Seismol Res Lett 80 (3):465-472. doi:10.1785/gssrl.80.3.465

Bondár I, Myers SC, Engdahl ER, Bergman EA (2004) Epicentre accuracy based on seismic network criteria. Geophys J Int 156 (3):483-496. doi:10.1111/J.1365-246x.2004.02070.X

Boyd T, Snoke J (1984) Error estimates in some commonly used earthquake location programs. Earthquake Notes 55 (2):3-6

Chatelain JL, Roecker SW, Hatzfeld D, Molnar P (1980) Microearthquake seismicity and fault plane solutions in the Hindu Kush Region and their tectonic implications. Journal of Geophysical Research 85 (B3):1365-1387. doi:10.1029/JB085iB03p01365

Chau M, Cheng R, Kao B, Ng J (2006) Uncertain data mining: An example in clustering location data. Lect Notes Artif Int 3918:199-204

Console R, Digiovambattista R (1987) Local Earthquake Relative Location by Digital Records. Phys Earth Planet In 47:43-49. doi:10.1016/0031-9201(87)90065-3

Courjault-Radé P, Darrozes J, Gaillot P (2009) The M = 5.1 1980 Arudy earthquake sequence (western Pyrenees, France): a revisited multi-scale integrated seismologic, geomorphologic and tectonic investigation. Int J Earth Sci (Geol Rundsch) 98 (7):1705-1719. doi:10.1007/s00531-008-0320-5

Cowie PA, Sornette D, Vanneste C (1995) Multifractal Scaling Properties of a Growing Fault Population. Geophys J Int 122 (2):457-469. doi:10.1111/J.1365-246x.1995.Tb07007.X

Cowie PA, Vanneste C, Sornette D (1993) Statistical Physics Model for the Spatiotemporal Evolution of Faults. J Geophys Res-Sol Ea 98 (B12):21809-21821. doi:10.1029/93jb02223

D. Giardini, J. Woessner, L. Danciu, G. Valensise, G. Grünthal, F. Cotton, S. Akkar, R. Basili, M. Stucchi, A. Rovida, D. Stromeyer, R. Arvidsson, F. Meletti, R. Musson, R., K. Sesetyan, M. B. Demircioglu, H. Crowley, R. Pinho, K. Pitilakis, J. Douglas, J. Fonseca, M. Erdik, A. Campos-Costa, B. Glavatovic, K. Makropoulos, C. Lindholm, T. Cameelbeeck Seismic Hazard Harmonization in Europe (SHARE): Online Data Resource. doi:10.12686/SED-00000001-SHARE, 2013

Deichmann N, Ernst J (2009) Earthquake focal mechanisms of the induced seismicity in 2006 and 2007 below Basel (Switzerland). Swiss J Geosci 102 (3):457-466. doi:10.1007/S00015-009-1336-Y

Deichmann N, Garciafernandez M (1992) Rupture Geometry from High-Precision Relative Hypocenter Locations of Microearthquake Clusters. Geophys J Int 110 (3):501-517. doi:10.1111/J.1365-246x.1992.Tb02088.X

Deichmann N, Giardini D (2009) Earthquakes Induced by the Stimulation of an Enhanced Geothermal System below Basel (Switzerland). Seismol Res Lett 80 (5):784-798. doi:10.1785/Gssrl.80.5.784

Deichmann N, Kraft T, Evans KF (2013) Identification of faults activated during the stimulation of the Basel geothermal project from cluster analysis and focal mechanisms of the larger magnitude events. submitted to Geothermics.

Diehl T, Kissling E, Husen S, Aldersons F (2009) Consistent phase picking for regional tomography models: application to the greater Alpine region. Geophys J Int 176 (2):542-554. doi:10.1111/J.1365-246x.2008.03985.X

Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley,

Faulkner DR, Lewis AC, Rutter EH (2003) On the internal structure and mechanics of large strike-slip fault zones: field observations of the Carboneras fault in southeastem Spain.

Tectonophysics 367 (3-4):235-251. doi:10.1016/S0040-1951(03)00134-3

Field EH, Milner, Kevin R., and the 2007 Working Group on California Earthquake Probabilities (2008) Forecasting California's earthquakes; what can we expect in the next 30 years? US Geological Survey

Frankel AD, Carver DL, Williams RA (2002) Nonlinear and linear site response and basin effects in Seattle for the M 6.8 Nisqually, Washington, earthquake. B Seismol Soc Am 92 (6):2090-2109. doi:10.1785/0120010254

Gabrielov A, KeilisBorok V, Jackson DD (1996) Geometric incompatibility in a fault system. P Natl Acad Sci USA 93 (9):3838-3842. doi:10.1073/Pnas.93.9.3838

Gaillot P, Darrozes J, Courjault-Rade P, Amorese D (2002) Structural analysis of hypocentral distribution of an earthquake sequence using anisotropic wavelets: Method and application. J Geophys Res-Sol Ea 107 (B10). doi:10.1029/2001jb000212

Gerstenberger MC, Wiemer S, Jones LM, Reasenberg PA (2005) Real-time forecasts of tomorrow's earthquakes in California. Nature 435 (7040):328-331. doi:10.1038/Nature03622

Gomberg JS, Shedlock KM, Roecker SW (1990) The effect of S-wave arrival times on the accuracy of hypocenter estimation. B Seismol Soc Am 80 (6A):1605-1628

Gulia L, Wiemer S (2010) The influence of tectonic regimes on the earthquake size distribution: A case study for Italy. Geophys Res Lett 37. doi:10.1029/2010gl043066

Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. B Seismol Soc Am 34 (4):185-188

Hainzl S, Steacy S, Marsan D (2010) Seismicity models based on Coulomb stress calculations. Community Online Resource for Statistical Seismicity Analysis. doi:10.5078/corssa-32035809

Hardebeck JL (2013) Geometry and Earthquake Potential of the Shoreline Fault, Central California. B Seismol Soc Am 103 (1):447-462. doi:10.1785/0120120175

Hardebeck JL, Michael AJ (2006) Damped regional-scale stress inversions: Methodology and examples for southern California and the Coalinga aftershock sequence. Journal of Geophysical

Research: Solid Earth 111 (B11):B11310. doi:10.1029/2005jb004144

Hardebeck JL, Shearer PM (2002) A New Method for Determining First-Motion Focal Mechanisms. B Seismol Soc Am 92 (6):2264-2276. doi:10.1785/0120010200

Häring MO, Schanz U, Ladner F, Dyer BC (2008) Characterisation of the Basel 1 enhanced geothermal system. Geothermics 37 (5):469-495. doi:10.1016/J.Geothermics.2008.06.002

Hauksson E (2010) Spatial Separation of Large Earthquakes, Aftershocks, and Background Seismicity: Analysis of Interseismic and Coseismic Seismicity Patterns in Southern California. Pure Appl Geophys 167 (8-9):979-997. doi:10.1007/S00024-010-0083-3

Hauksson E, Yang WZ, Shearer PM (2012) Waveform Relocated Earthquake Catalog for Southern California (1981 to June 2011). B Seismol Soc Am 102 (5):2239-2244. doi:10.1785/0120120010

Helmstetter A, Kagan YY, Jackson DD (2006) Comparison of short-term and time-independent earthquake forecast models for southern California. B Seismol Soc Am 96 (1):90-106. doi:10.1785/0120050067

Hiemer S, Jackson DD, Wang Q, Kagan YY, Woesner J, Zechar JD, Wiemer S (2013) A stochastic forecast of California earthquakes based on fault slip and smoothed seismicity. B Seismol Soc Am 103 (2A). doi:10.1785/0120120168

Husen S, Bachmann C, Giardini D (2007) Locally triggered seismicity in the central Swiss Alps following the large rainfall event of August 2005. Geophys J Int 171 (3):1126-1134. doi:10.1111/J.1365-246x.2007.03561.X

Husen S, Hardebeck JL (2010) Earthquake location accuracy. Community Online Resource for Statistical Seismicity Analysis. doi:10.5078/corssa-55815573

Husen S, Kissling E, Deichmann N, Wiemer S, Giardini D, Baer M (2003) Probabilistic earthquake location in complex three-dimensional velocity models: Application to Switzerland. J Geophys Res-Sol Ea 108 (B2):2077. doi:10.1029/2002jb001778

Husen S, Kissling E, Flueh ER (2000) Local earthquake tomography of shallow subduction in north Chile: A combined onshore and offshore study. J Geophys Res-Sol Ea 105 (B12):28183-28198. doi:10.1029/2000jb900229

Husen S, Smith RB (2004) Probabilistic Earthquake Relocation in Three-Dimensional Velocity Models for the Yellowstone National Park Region, Wyoming. B Seismol Soc Am 94 (3):880-896. doi:10.1785/0120030170

Hutchinson JE (1981) Fractals and Self Similarity. Indiana U Math J 30 (5):713-747. doi:10.1512/Iumj.1981.30.30055

Jordan TH, Chen YT, Gasparini P, Madariaga R, Main I, Marzocchi W, Papadopoulos G, Sobolev G, Yamaoka K, Zschau J (2011) Operational Earthquake Forecasting: State of Knowledge and Guidelines for Utilization. Ann Geophys-Italy 54 (4):315-391. doi:10.4401/Ag-5350

Kagan YY (2005) Double-couple earthquake focal mechanism: random rotation and display. Geophys J Int 163 (3):1065-1072. doi:10.1111/J.1365-246x.2005.02781.X

Kissling E (1988) Geotomography with local earthquake data. Reviews of Geophysics 26 (4):659-698. doi:10.1029/RG026i004p00659

Klinger Y, Xu XW, Tapponnier P, Van der Woerd J, Lasserre C, King G (2005) High-resolution satellite imagery mapping of the surface rupture and slip distribution of the M-W similar to 7.8, 14 November 2001 Kokoxili Earthquake, Kunlun Fault, northern Tibet, China. B Seismol Soc Am 95 (5):1970-1987. doi:10.1785/0120040233

Lahr JC (1989) HYPOELLIPSE/VERSION 2.0*, a computer program for determining local earthquake hypocentral parameters, magnitude, and first motion pattern. U.S. Geological Survey,

Lee SD, Kao B, Cheng R (2007) Reducing UK-Means to K-Means. Paper presented at the Proceedings of the Seventh IEEE International Conference on Data Mining Workshops,

Lee WHK, Stewart SW (1981) Principles and Applications of Microearthquake Networks. Phys Earth Planet In 31 (2):191-192. doi:10.1016/0031-9201(83)90115-2

Lomax A, Michelini A, Curtis A (2009) Earthquake Location, Direct, Global-Search Methods. In: Meyers RA (ed) Encyclopedia of Complexity and System Science. Springer, New York, pp 2249-2473. doi:10.1007/978-0-387-30440-3

Lomax A, Virieux J, Volant P, Berge-Thierry C (2000) Probabilistic earthquake location in 3D and layered models - Introduction of a Metropolis-Gibbs method and comparison with linear locations. Advances in Seismic Event Location:101-134

Lund B, Townend J (2007) Calculating horizontal stress orientations with full or partial knowledge of the tectonic stress tensor. Geophys J Int 170 (3):1328-1335. doi:10.1111/J.1365-246x.2007.03468.X

Mace CG, Keranen KM (2012) Oblique fault systems crossing the Seattle Basin: Geophysical evidence for additional shallow fault systems in the central Puget Lowland. J Geophys Res-Sol Ea 117. doi:10.1029/2011jb008722

MacQueen JB Some Methods for Classification and Analysis of MultiVariate Observations. In: Cam LML, Neyman J (eds) Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. University of California Press, pp 281-297

Marzocchi W, Zechar JD, Jordan TH (2012) Bayesian Forecast Evaluation and Ensemble Earthquake Forecasting. B Seismol Soc Am 102 (6):2574-2584. doi:10.1785/0120110327

Mignan A, Werner MJ, Wiemer S, Chen CC, Wu YM (2011) Bayesian Estimation of the Spatially Varying Completeness Magnitude of Earthquake Catalogs. B Seismol Soc Am 101 (3):1371-1385. doi:10.1785/0120100223

Moser TJ, Vaneck T, Nolet G (1992) Hypocenter Determination in Strongly Heterogeneous Earth Models Using the Shortest-Path Method. J Geophys Res-Sol Ea 97 (B5):6563-6572. doi:10.1029/91JB03176

Ogata Y (1988) Statistical-Models for Earthquake Occurrences and Residual Analysis for Point-Processes. J Am Stat Assoc 83 (401):9-27. doi:10.2307/2288914

Ogata Y, Zhuang HC (2006) Space-time ETAS models and an improved extension. Tectonophysics 413 (1-2):13-23. doi:10.1016/J.Tecto.2005.10.016

Ouillon G, Castaing C, Sornette D (1996) Hierarchical geometry of faulting. J Geophys Res-Sol Ea 101 (B3):5477-5487. doi:10.1029/95jb02242

Ouillon G, Ducorbier C, Sornette D (2008) Automatic reconstruction of fault networks from seismicity catalogs: Three-dimensional optimal anisotropic dynamic clustering. J Geophys Res-Sol Ea 113 (B1). doi:10.1029/2007jb005032

Ouillon G, Sornette D (2011) Segmentation of fault networks determined from spatial clustering of earthquakes. J Geophys Res-Sol Ea 116. doi:10.1029/2010jb007752

Ouillon G, Sornette D, Castaing C (1995) Organisation of joints and faults from 1-cm to 100-km scales revealed by optimized anisotropic wavelet coefficient method and multifractal analysis. Nonlinear Proc Geoph 2 (3-4):158-177

Page MT, Alderson D, Doyle J (2011) The magnitude distribution of earthquakes near Southern California faults. J Geophys Res-Sol Ea 116. doi:10.1029/2010jb007933

Passchier CW, Trouw RAJ (2005) Microtectonics. Springer,

Pavlis GL (1986) Appraising earthquake hypocenter location errors: A complete, practical approach for single-event locations. B Seismol Soc Am 76 (6):1699-1717

Petersen MD, Cao TQ, Campbell KW, Frankel AD (2007) Time-independent and time-dependent seismic hazard assessment for the State of California: Uniform California Earthquake Rupture Forecast Model 1.0. Seismol Res Lett 78 (1):99-109. doi:10.1785/Gssrl.78.1.99

Pisarenko VF, Sornette A, Sornette D, Rodkin MV (2008) New approach to the characterization of M (max) and of the tail of the distribution of earthquake magnitudes. Pure Appl Geophys 165 (5):847-888. doi:10.1007/S00024-008-0341-9

Pisarenko VF, Sornette D (2003) Characterization of the frequency of extreme earthquake events by the Generalized Pareto Distribution. Pure Appl Geophys 160 (12):2343-2364. doi:10.1007/S00024-003-2397-X

Plesch A, Shaw JH, Benson C, Bryant WA, Carena S, Cooke M, Dolan J, Fuis G, Gath E, Grant L, Hauksson E, Jordan T, Kamerling M, Legg M, Lindvall S, Magistrale H, Nicholson C, Niemi N, Oskin M, Perry S, Planansky G, Rockwell T, Shearer P, Sorlien C, Suss MP, Suppe J, Treiman J, Yeats R (2007) Community fault model (CFM) for southern California. B Seismol Soc Am 97 (6):1793-1802. doi:10.1785/0120050211

Podvin P, Lecomte I (1991) Finite difference computation of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. Geophys J Int 105 (1):271-284. doi:10.1111/j.1365-246X.1991.tb03461.x

Powers PM, Jordan TH (2010) Distribution of seismicity across strike-slip faults in California. J Geophys Res-Sol Ea 115. doi:10.1029/2008jb006234

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical Recipes 3rd Edition: The Art of Scientific Computing. Third Edition edn. Cambridge University Press,

Rhoades DA, Stirling MW (2012) An Earthquake Likelihood Model Based on Proximity to Mapped Faults and Cataloged Earthquakes. B Seismol Soc Am 102 (4):1593-1599. doi:10.1785/0120110326

Richards PG, Waldhauser F, Schaff D, Kim WY (2006) The applicability of modern methods of earthquake location. Pure Appl Geophys 163 (2-3):351-372. doi:10.1007/S00024-005-0019-5

Richards-Dinger K, Dieterich JH (2012) RSQSim Earthquake Simulator. Seismol Res Lett 83 (6):983-990. doi:10.1785/0220120105

Rowe CA, Aster RC, Borchers B, Young CJ (2002) An automatic, adaptive algorithm for refining phase picks in large seismic data sets. B Seismol Soc Am 92 (5):1660-1674. doi:10.1785/0120010224

Scholz CH (2002) The Mechanics of Earthquakes and Faulting. Cambridge University Press,

Schorlemmer D, Wiemer S (2005) Microseismicity data forecast rupture area. Nature 434 (7037):1086-1086. doi:10.1038/4341086a

Schorlemmer D, Wiemer S, Wyss M (2004) Earthquake statistics at Parkfield: 1. Stationarity of b values. J Geophys Res-Sol Ea 109 (B12). doi:10.1029/2004jb003234

Schorlemmer D, Wiemer S, Wyss M (2005) Variations in earthquake-size distribution across different stress regimes. Nature 437 (7058):539-542. doi:10.1038/Nature04094

Schwarz G (1978) Estimating Dimension of a Model. Ann Stat 6 (2):461-464. doi:10.1214/Aos/1176344136

Sornette D (1991) Self-Organized Criticality in Plate-Tectonics. Nato Adv Sci I C-Mat 349:57-106

Sornette D, Miltenberger P, Vanneste C (1994) Statistical Physics of Fault Patterns Self-Organized by Repeated Earthquakes. Pure Appl Geophys 142 (3-4):491-527. doi:10.1007/Bf00876052

Sornette D, Virieux J (1992) Linking Short-Timescale Deformation to Long-Timescale Tectonics. Nature 357 (6377):401-404. doi:10.1038/357401a0

Stegman CE (1989) Nonparametric Statistics for the Behavioral-Sciences, 2nd Edition, Siegel,S, Castellan,Nj. Contemp Psychol 34 (8):773-774

Stein RS (1999) The role of stress transfer in earthquake occurrence. Nature 402 (6762):605-609. doi:10.1038/45144

Stirling MW, Wesnousky SG, Shimazaki K (1996) Fault trace complexity, cumulative slip, and the shape of the magnitude-frequency distribution for strike-slip faults: A global survey. Geophys J Int 124 (3):833-868. doi:10.1111/J.1365-246x.1996.Tb05641.X

Tarantola A, Valette B (1982) Inverse Problems = Quest for Information. Journal of Geophysics 50:159-170

Tchalenko JS (1970) Similarities between Shear Zones of Different Magnitudes. Geol Soc Am Bull 81 (6):1625-&. doi:10.1130/0016-7606(1970)81[1625:Sbszod]2.0.Co;2

Tchalenko JS, Ambraseys NN (1970) Structural Analysis of Dasht-E Bayaz (Iran) Earthquake Fractures. Geol Soc Am Bull 81 (1):41-&. doi:10.1130/0016-7606(1970)81[41:Saotdb]2.0.Co;2

Thurber C, Ritsema J, Schubert G (2007) Theory and Observations - Seismic Tomography and Inverse Methods. In: Treatise on Geophysics. Elsevier, pp 323-360

Thurber CH (1992) Hypocenter-velocity structure coupling in local earthquake tomography. Phys Earth Planet In 75 (1–3):55-62. doi:10.1016/0031-9201(92)90117-e

Tullis TE (2012) Preface to the Focused Issue on Earthquake Simulators. Seismol Res Lett 83 (6):957-958. doi:10.1785/0220120122

Waldhauser F, Ellsworth WL (2000) A double-difference earthquake location algorithm: Method and application to the northern Hayward fault, California. B Seismol Soc Am 90 (6):1353-1368. doi:10.1785/0120000006

Waldhauser F, Schaff DP (2008) Large-scale relocation of two decades of Northern California seismicity using cross-correlation and double-difference methods. J Geophys Res-Sol Ea 113 (B8). doi:10.1029/2007jb005479

Wang Y, Husen S, Woessner J, Ouillon G, Sornette D (2013a) Assessing earthquake location quality using seismic network criteria and its importance in fault network reconstructions. submitted to J. Seismol.

Wang Y, Ouillon G, Woessner J, Sornette D, Husen S (2013b) Automatic reconstruction of fault networks from seismicity catalogs including location uncertainty. submitted to J. Geohys. Res. The manuscript can be downloaded at http://arxivorg/abs/13046912

Werner MJ, Ide K, Sornette D (2011) Earthquake forecasting based on data assimilation: sequential Monte Carlo methods for renewal point processes. Nonlinear Proc Geoph 18 (1):49-70. doi:10.5194/Npg-18-49-2011

Wesson RL, Bakun WH, Perkins DM (2003) Association of earthquakes and faults in the San Francisco Bay area using Bayesian inference. B Seismol Soc Am 93 (3):1306-1332. doi:10.1785/0120020085

Wiemer S, Wyss M (2002) Mapping spatial variability of the frequency-magnitude distribution of earthquakes. Adv Geophys 45:259-302

Wittlinger G, Herquel G, Nakache T (1993) Earthquake location in strongly heterogeneous media. Geophys J Int 115 (3):759-777. doi:10.1111/j.1365-246X.1993.tb01491.x

Woessner J, Christophersen A, Zechar JD, Monelli D (2010) Building self-consistent, short-term earthquake probability (STEP) models: improved strategies and calibration procedures. Ann Geophys-Italy 53 (3):141-154. doi:10.4401/Ag-4812

Woessner J, Hainzl S, Marzocchi W, Werner MJ, Lombardi AM, Catalli F, Enescu B, Cocco M, Gerstenberger MC, Wiemer S (2011) A retrospective comparative forecast test on the 1992 Landers sequence. J Geophys Res-Sol Ea 116. doi:10.1029/2010jb007846

Woessner J, Schorlemmer D, Wiemer S, Mai PM (2006) Spatial correlation of aftershock locations and on-fault main shock properties. J Geophys Res-Sol Ea 111 (B8). doi:10.1029/2005jb003961

Woessner J, Treml M, Wenzel F (2002) Simulation of M-W=6.0 earthquakes in the Upper Rhinegraben using empirical Green functions. Geophys J Int 151 (2):487-500. doi:10.1046/J.1365-246x.2002.01785.X

Woessner J, Wiemer S (2005) Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty. B Seismol Soc Am 95 (2):684-698. doi:10.1785/0120040007

Yang WZ, Hauksson E, Shearer PM (2012) Computing a Large Refined Catalog of Focal Mechanisms for Southern California (1981-2010): Temporal Stability of the Style of Faulting. B Seismol Soc Am 102 (3):1179-1194. doi:10.1785/0120110311

Yang X, Bondár I, Bhattacharyya J, Ritzwoller M, Shapiro N, Antolik M, Ekström G, Israelsson H, McLaughlin K (2004) Validation of Regional and Teleseismic Travel-Time Models by Relocating Ground-Truth Events. B Seismol Soc Am 94 (3):897-919. doi:10.1785/0120030148

Zechar JD, Jordan TH (2008) Testing alarm-based earthquake predictions. Geophys J Int 172 (2):715-724. doi:10.1111/J.1365-246x.2007.03676.X

Zechar JD, Jordan TH (2010) Simple smoothed seismicity earthquake forecasts for Italy. Ann Geophys-Italy 53 (3):99-105. doi:10.4401/Ag-4845

Zechar JD, Schorlemmer D, Werner MJ, Gerstenberger MC, Rhoades DA, Jordan TH (2013) Regional Earthquake Likelihood Models I: First-Order Results. B Seismol Soc Am 103 (2A):787-798. doi:10.1785/0120120186

Zhuang J, Harte D, Werner MJ, Hainzl S, Zhou S (2012) Basic models of seismicity: temporal models. Community Online Resource for Statistical Seismicity Analysis. doi:10.5078/corssa-79905851

Zhuang J, Werner MJ, Hainzl S, Harte D, Zhou S (2011) Basic models of seismicity: spatiotemporal models. Community Online Resource for Statistical Seismicity Analysis. doi:10.5078/corssa-07487583

Acknowledgement – 致谢

It would not have been possible to finish this doctoral thesis without the help of all the group members: Prof. Didier Sornette, Dr. Stephan Husen, Dr. Guy Ouillon and Dr. Jochen Woessner. This thesis benefits from their deep knowledge in earthquake forecasting, statistics, earthquake location, etc.. I appreciate very much for the knowledge I learnt from them in the past years. Furthermore during many profound discussions, they let me realized that there are many doors still open within many different fields. I will definitely benefit from these extraordinary experiences for my whole life.

I would like to thank Professor Stefan Wiemer for his support and encouraging within the last year of my PhD study which made a great part of this thesis possible to finish.

I thank Nicholas Deichmann, Toni Kraft, Egill Hauksson and Anthony Lomax for providing the catalog resources and location program used in this thesis.

I thank Chingyi, Shyam and Yavor. I will miss the “once in a while dinner and beers”. They made me realized I am not alone on the road.

One of the best outcomes from these past years is meeting many good friends with different cultural backgrounds. Cem, Daniel, Fatih, Hendrik, Isa, Matthew, Nico, Qurin, Spitzi, Suna…, there were so many joyful memories we shared together. I am impressed by the diversities in human societies and cultures. I will always cherish these friendships and memories.

过去的五年，是不容易的五年。幸好，一直有 ph=7.18，雅静，小杨，小曾还有李博士的支持和鼓励。一个电话，一声问候，都时常让本人倍感幸福。从苏黎世到北京有 7984km 的距离。瑞士和中国有六或七个小时的时差。但即使这样也丝毫的没有影响到我们亲密的友情。嗯，心与心的距离有时是可以很近的。

非常庆幸我成长在我父母建造的那个的家。他们给与的信任和自由让我一直在自己向往的路上走着。他们一直给予的那无穷无尽的，难以言说的，中国式的爱是我追求理想的强大动力。这篇论文是我献与他们的一个礼物。

Curriculum Vitæ

Yaming Wang

Born on October 06 1981, Shannxi, China

Citizen of China

2008-2013: Doctoral Student, Swiss Seismological Service, Institute of Geophysics, ETH Zurich, Switzerland

2004-2007: Master Student, School of Instrumentation Science & Opto-electronics Engineering, Beijing University of Aeronautics & Astronautics, China

1999-2003: Bachelor Student, School of Astronautics, Beijing University of Aeronautics & Astronautics, China