STATISTICAL ANALYSIS OF SEISMICITY IN INTRAPLATE REGIONS

by

Luc E. Chouinard

M. Ing. Ecole Polytechnique de Montreal (1983),
B.C.L. McGill University (1983),
B. Ing. Ecole Polytechnique de Montreal (1979)

SUBMITTED TO THE DEPARTMENT OF CIVIL ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 1989

Signature of Author _____
Department of Civil Engineering
January 20, 1989

Certified by _____
Professor Daniele Veneziano
Thesis Supervisor

Accepted by _____
Professor Ole S. Madsen
Chairman, Departmental Committee on Graduate Students

Statistical Analysis of Seismicity in Intraplate Regions

by

Luc E. Chouinard

Submitted to the Department of Civil Engineering on January 20, 1989 in partial fulfillment of the requirements for the degree of Doctor of Science.

Abstract

A local estimation procedure is proposed which is based on statistical tests that identify zones of local homogeneity with respect to the seismic process. The resulting estimator preserves significant discontinuities of the recurrence rate of earthquakes. This is an improvement over present day procedures which require the external specification of seismic sources inside which seismicity is assumed constant. In the proposed procedure, seismic sources can optionally be used in the identification of significant features but influence the estimates only if validated by the data.

With respect to the selection of model parameters, two selection procedures are proposed. The first one is based on the method of moments and consists in matching observed statistics to some target values. In the second procedure, which is known as cross-validation, the catalog is divided into non-overlapping estimation and validation samples and optimal parameters are selected on the basis of statistics measuring the goodness-of-fit of the predictions. The first procedure is intuitive and easy to implement, however, it lacks the predictive interpretation of the second procedure, which can be used to simultaneously select several model parameters, and compare competing models.

Finally, a combined estimator of seismic hazard which makes use of both seismic source and historical estimators of seismic hazard is proposed. This estimator is a simple alternative to the previous models of seismicity, is shown to be robust with respect to the specification of source configurations, and is a significant improvement over the seismic source and historical estimators.

Thesis Supervisor: Professor Daniele Veneziano
Title: Professor of Civil Engineering

Dedication

To my parents.

Acknowledgments

I would like to extend my appreciation to :

Professor Daniele Veneziano, for his constant supervision, support, and critical advice during my graduate studies. His contributions to this thesis are too many to mention explicitly here,

My other committee members, Professors Gregory Baecher and Nafi Toksoz, for their useful comments and interest,

All the friends I have made here over the years and which have enriched my everyday life,

My parents, brother and sister for their constant encouragements,,

And specially Heidi for her patience, support, and affection.

I would like to acknowledge the financial support from the National Science and Engineering Council of Canada, the Standard Oil of Ohio Corporation, and the National Center for Earthquake Engineering Research at Buffalo which contributed to the financial support of my graduate studies.

Table of Contents

List of Figures

List of Tables

Chapter 1

Statement of the Problem

Seismicity in intraplate regions is generally assumed to be due to changes in the stress conditions within previously weak portions of the earth's crust as opposed to new faults created during earthquakes. As a consequence, one may expect the level of seismicity to vary as a function of location given the necessity for the simultaneous presence of existing faults and high stresses. However, the spatial variation of seismicity is not well known given the uncertainty with respect to existing stress levels and the location of potentially active faults and one must rely on infomation such as past seismicity and geological and geotectonic characteristics.

It is generally assumed that the occurence of main events can be modelled through a Poisson process which is homogeneous in space within so-called seismogenic provinces and stationary in time. This is the procedure adopted by EPRI (1985), LLNL (1985), YAEC (1983), and the USGS (Algermissen et al. 1982). In all these procedures, seismogenic provinces are typically identified by experts based on an analysis of the historical seismicity and the geological and tectonic setting of the region. There is a lot of uncertainty on the exact configuration of these zones which gives rise to many competing hypotheses.

It is also frequently assumed that main events in each province have exponentially distributed size (macroseismic intensity I, or magnitude m), such that the annual rate of main events with size in the interval ($I-\Delta I/2$, $I+\Delta I/2$) in a unit area of province i is

$$\lambda_i(I) = exp\{a_i - b_i(I - I^*)\} \qquad (1.1)$$

The rate is nil above an upper-bound size I_1, which is independently determined for each seismic source.

Many estimation procedures have been proposed for the estimation of the parameters a and b in Equation 1.1. The most popular ones are least squares on rates, least squares on cumulative rates, maximum likelihood, and maximum penalized likelihood. Maximum likelihood estimation of the exponential rate parameters a and b under the previous assumptions has been studied by (Aki, 1965) and more recently by (Weichert, 1980). The method of Weichert allows for a different period of completeness for each magnitude. The bias effect of magnitude discretization and the uncertainty on the estimation of the slope parameters when using maximum likelihood have been studied in detail by (Bender, 1983). The maximum penalized likelihood, in the version of (Veneziano and VanDyck, 1987), is perhaps the best procedure from a statistical point of view. In this procedure, which is reviewed in section 2.1, incompleteness of the catalog is modelled through a probability of detection which is piecewise continuous in space, time, and magnitude. An innovative feature of the procedure is that imcompleteness and the seismicity parameters are simultaneously estimated. For the estimation of the probability of detection, it is assumed that deviations from the assumptions of stationarity and exponentiality in size distribution are due to incompleteness (section 2.1). Estimates of the probability of detection are further constrained such that monoticity is enforced in the space of increasing time and magnitude. The implementation of the estimation procedure for the seismicity parameters requires that space, and magnitude be discretized. The discretizations in space and magnitude can be different from those used for the definition of the probability of detection. The smoothness level of the esitmates of $a(\underline{x})$ and $b(\underline{x})$ is specified by the user and controlled through additional terms in the likelihood function. In this procedure, the parameters a and b are allowed to vary smoothly within each seismogenic province, but are discontinuous across province boundaries. Optionally, one may decide not to specify provinces and allow the seismicity parameters to vary smoothly in space.

Smoothness is controlled by penalizing deviations between the estimates at a location and the average of the estimates at the immediately neighboring locations. This often leads to undesirable results in regions which exhibit sharp contrasts in the rate of activity, specially for large penalties.

In section 2.2, a statistical procedure is developed which objectively identifies zones of homogeneous seismicity and preserves significant features through the definition of interpolation neighborhoods. The advantage of such a procedure is that significant discontinuities in the recurrence rate are preserved even under large penalties. In section 2.3, the above procedure is modified to allow the inclusion of expert source configurations in the identification of the homogeneous interpolation neighborhoods. The resulting estimates are shown to be robust with respect to source configurations.

An important issue for the family of models considered in this thesis is the selection of the degree of smoothness of the estimates. In Chapter 3, two methods are developped for the objective and optimal selection of the model parameters: we either impose that certain observed statistics equal some predetermined target value or maximize cross-validated measures of goodness-of-fit. For the latter purpose, the catalog is divided into non-overlapping estimation and validation samples and models are classified according to how accurately they predict future events. Therefore, these are attractive statistics for selecting seismicity models to be used in earthquake hazard studies.

The above procedures and models are applied to the Eastern United States using the earthquake catalog compiled by EPRI (1985) and to the New England region using the catalog compiled by (Chiburis, 1981). Goodness-of-fit procedures are also proposed with respect to the assumptions of exponentiality in size distribution, and stationarity in time and in space.

In Chapter 4, an alternative procedure to the local models of Chapters 2 and 3 is proposed for the computation of the seismic hazard at a site. The new estimator

combines two conventioanl estimators of seismic hazard: the seismic-sources and historical estimators of seismic hazard, and is shown to be more accurate than either estimator individually. The estimators are shown to be robust with respect to externally-specified source configurations and to produce estimates similar to those using the local models of seismicity.

Chapter 5 summarizes the proposed methods of analysis and states conclusions and recommendations for future research within the family of models considered in the thesis.

Chapter 2

Local Models of Seismicity

2.1 Introduction

Events in an earthquake catalog can be thought of as points in a multidimensional space ($\underline{x}$,t,m); for earthquake i, $\underline{x}_i$ is the geographical location, t_i is the time of occurence and m_i is a size measure. The problem discussed in this chapter is how to estimate the rate density function $\nu(\underline{x},m)$ from the historical data. This function is defined such that the rate $\nu(\underline{x},m)$ is the expected count of earthquakes in the infinitesimal neighborhood ($d\underline{x}$,dm) around ($\underline{x}$,m). Two basic assumptions are used throughout the chapter: 1. The earthquake sequence is the realization of a Poisson process, i.e. points in ($\underline{x}$,m) space are independently located, 2. Nonstationarity of the observed earthquake sequence is due to incomplete reporting, whereas the seismicity generating process is stationary. Therefore, the yearly rate of events over a unit area can be written as

$$\lambda(\underline{x},t,m) = \nu(\underline{x},m) \bullet P_D(\underline{x},t,m) \tag{2.1}$$

where $P_D(\underline{x},t,m)$ is the probability that an earthquake of size m, and at location $\underline{x}$ and time t is reported. Each event is assumed to be independently reported (detection / no detection of different earthquakes are independent events).

Deviations from the model asumptions can be accomodated through fitting the model locally in time and magnitude. For example, the assumption of stationarity can be relaxed by using only the most recent portion of the catalog for estimating the seismic hazard in the next following years. Similarly, deviations from the assumption of

exponentiality can be accomodated by assigning different weights to each magnitude interval so that the model fits better over a given range in magnitude.

At present, most procedures for the estimation of recurrence rates employ additional asumptions, 1. $\nu(\underline{x},m)$ is considered uniform within given regions (seismogenic provinces) S_i, i.e.

$$\nu(\underline{x},m) = \nu_i(m) \quad for\ \underline{x} \in S_i \tag{2.2}$$

2. the rate density inside province k, ν_k, varies exponentially with m, i.e.

$$ln\ \nu_k(m) = a_k - b_k m \quad for \quad m_o < m < m_{i_k} \tag{2.3}$$

, where a_k and b_k are unknown parameters, m_0 is a lower bound of interest and m_{i_k} is a physical upper bound, which may vary from province to province, 3. inside prescribed regions S_l, which may be different from the previous region, the catalog is complete for magnitude m within the last $t_l(m)$ years (so-called period of completeness) (Stepp, 1972), so that:

$$P_D(\underline{x},t,m)=1.0 \tag{2.4}$$

if $\underline{x} \in S_l$ and $T_0(m)-t_l(m) \leq t \leq T_0(m)$, where $T_0(m)$ is the time of the most recent observation included in the catalog. The seismogenic provinces S_i are not necessarily the same as the completeness regions S_l. Under the above assumptions, estimation of the parameters a_k and b_k in each province is relatively straight-forward if only earthquake data within the periods of completeness are used. The VanDyck (1986) approach differs fundamentally from earlier ones in the sense that the probability of detection $P_D(\underline{x},t,m)$ and the seismicity rate $\nu(\underline{x},m)$ are simultaneously estimated from the data. Doing so allows one to utilize a larger part of the historical data and provides means to objectively quantify the completeness of the catalog.

However, it would be impossible to estimate all of these parameters accurately without introducing some constraints in the estimation procedure, given the sparsity of data in most locations. This is accomplished by introducing prior information on the spatial

variation of the parameters through penalties and prior distributions and by constraining the estimation of the probability of detection to be monotonically increasing in the space defined by time and magnitude.

In section 2.1, the formulation and estimation procedure of the model proposed by VanDyck (1985) is reviewed. In section 2.2, a modification of that procedure is proposed which is based on identifying homogeneous interpolation neighborhoods for each of the seismicity parameters through statistical tests of homogeneity. In section 2.3, the above procedure is extended to allow the inclusion of expert opinion in determining the size and shape of the local interpolation neighborhoods. Finally, in section 2.4, the procedures of section 2.2 are applied to two different regions and earthquake catalogs, first to New England using the Chiburis (1981) catalog and second to the eastern United States using the EPRI catalog. Estimates are also obtained for the Eastern United States using several alternative expert zonations recently proposed in the litterature, some exclusively based on geological and tectonic information and some which follow more closely the contours of the historical seismicity. In section 2.5, a kernel estimator of seismicity is reviewed which allows for the relaxation of the assumptions of exponentiality in magnitude and stationarity in time. Conclusions and recommendations for future research are given in section 2.6.

2.1.1 The VanDyck model

Seismicity is described through a non-homogeneous Poisson process, with parameters a and b that vary as functions of the geographical location vector $\underline{x}$. Stationarity is assumed, at least during the time periods of the data and of the needed earthquake predictions. VanDyck (1985) developed four versions of this model (indexed form A to D). Depending on which of the four models is used, information on P_D is derived only from the non-stationarity and non-exponentiality of the observed recurrence rate

(models C and D) or also from the distribution of the population and seismic instruments in time and space (models A and B). Other differences between the various versions are detailed in VanDyck (1985). Model D, which was the last version developed is the one used in the thesis.

2.1.2 Derivation of the Likelihood

The most convenient way to solve the system of maximum likelihood equations is to discretize space and the magnitude of the events. For a Poisson process with a recurrence rate $\nu(m)$, the probability of observing n(m) earthquakes over a period of observation of T(m) years, in the discretized magnitude interval m $(m-\Delta m/2, m+\Delta/2)$, has Poisson distribution,

$$f_{N(m)}(n(m)) \;\alpha\; [\nu(m)T(m)]^{n(m)}\, e^{-\nu(m)T(m)} \tag{2.5}$$

The likelihood of the counts n(m) over the range (m_o, m_1) depends on the unknown recurrence rate $\nu(m)$ as

$$l(\nu(m) \mid (n(m),T(m))) \;=\; \prod_{m=m_o}^{m_1} f_{N(m)}(n(m)) \tag{2.6}$$

Using Eq. 1.1, the likelihood may be expressed as a function of the parameters a and b,

$$l(a,b \mid n(m),T(m)) \;\alpha\; \prod_{m=m_o}^{m_1} e^{n(m)(a-bm)}\, e^{-\sum_{m=m_o}^{m_1} T(m)exp(a-bm)} \tag{2.7}$$

Taking logs,

$$lnl(a,b \mid \{n(m),T(m)\}) =$$
$$a \sum_{m=m_o}^{m_1} n(m) - b \sum_{m=m_o}^{m_1} mn(m) - \sum_{m=m_o}^{m_1} T(m)e^{a-bm} + cst \tag{2.8}$$

The likelihood depends on the earthquakes only through the total counts N and the total magnitude M,

$$N = \sum_{m=m_o}^{m_1} n(m)$$
$$M = \sum_{m=m_o}^{m_1} mn(m) \tag{2.9}$$

Therefore N and M are sufficient statistics and the log-likelihood function simplifies to,

$$lnl(a,b \mid \{n(m),T(m)\}) = aN - bM - \sum_{m=m_o}^{m_1} T(m)e^{a-bm} + cst \qquad (2.10)$$

The maximum likelihood equations are found by taking partial derivatives of Equation 2.10 with respect to the unknown parameters a and b and setting them equal to zero.

$$N - \sum_{m=m_o}^{m_1} T(m)e^{a-bm} = 0$$

$$-M + \sum_{m=m_o}^{m_1} mT(m)e^{a-bm} = 0 \qquad (2.11)$$

Equation 2.11 implies that expected counts and total magnitude should equal observed counts and total magnitude. These equations can be efficiently solved using Newton's method. At the k^{th} iteration, estimates of a and b are found from,

$$\begin{bmatrix} a^k \\ b^k \end{bmatrix} = \begin{bmatrix} a^{k-1} \\ b^{k-1} \end{bmatrix} - \left[J^{-1}\right] \begin{bmatrix} \Delta f_a^{k-1} \\ \Delta f_b^{k-1} \end{bmatrix} \qquad (2.12)$$

where Δf_a^{k-1}, Δf_b^{k-1} are imbalances at the $(k-1)^{th}$ iteration, in Equations 2.11 and J is the Jacobian of the same system of equations.

For incomplete catalogs, it is also necessary to estimate the probability of detection. A non-parametric form of P_D is preferred in model D and does not consider the mode of detection or the distribution in space of population and instrument. The model of probability of detection is defined over regions that are homogeneous with respect to incompleteness, within the period of time of the analysis. Accordingly, only variation of P_D with t and m within each incompleteness region is considered.

$$P_D = \alpha_{tm} \qquad (2.13)$$

For the estimation of incompleteness, all variables are again discretized.

A likelihood equation can be derived in a similar fashion to the previous ones for the probability of detection. In this case we obtain an additional equation which is added to the system of equations 2.11 (VanDyck 1985), one for each incompleteness region,

$$\frac{n(t,m)}{\alpha_{tm}} - n^*_{tm} = 0 \qquad (2.14)$$

where n^*_{tm} is the expected number of events in the discretized interval (t,m). Estimates of α_{tm} for given n(t,m) are such that the observed count is each (t,m) category is matched. If the recurrence rates are unknown, P_D can be determined only up to a proportionality factor (in addition, one can vary the slope parameter $b(\underline{x})$ and the probabilities α_{tm} such that the likelihood remains the same). Various forms of constraints allow one to estimate P_D. For instance, 1. P_D is typically assumed to be 1.0 above a given magnitude and for recent times, 2. all very large earthquakes are assumed to have been reported over most of the time span of the catalog, and 3. P_D is expected to be smooth and increase monotonically as a function of time and magnitude. In general, constraints are imposed for the highest size measures throughout the entire time span of the catalog, because for strong events the counts are very small and consequently the estimates are unreliable if one does not use additional information. Goodness-of-fit with respect to the estimates of P_D can be checked by comparing the observed and predicted number of events for each cell defined by the discretization of space, time, and magnitude for the estimation of the probability of detection. Smoothness of P_D is imposed through maximum penalized likelihood estimation (MPLE). MPLE is also used to introduce smoothness in the spatial variation of $a(\underline{x})$ and $b(\underline{x})$ in order to reduce the statistical uncertainty on individual estimates.

More, in general, one can classify smooth estimation techniques in two broad classes;

1. Bayes-based methods
2. and Kernel-based methods.

In Bayes-based methods, smoothness is introduced through prior distributions on the seismicity parameters which can be either provided externally or estimated from the data (Berger, 1985). Penalties, in the maximum penalized formulation, can be viewed as priors on the functions $a(\underline{x})$ and $b(\underline{x})$. However, it is perhaps most appropriate to

interpret the technique as a pragmatic way to reduce the number of degreees of freedom of the model. The second general class of methods are the kernel-based methods for which there is an abundant litterature in the context of density estimation (Silverman, 1985) and contingency table analysis (Titterington, 1985), for an example of kernel-based procedures in seismicity see (Shakal and Toksoz, 1977). Generally, the procedure consists in using some of the observations in neighboring cells in the estimation of the parameter at $\underline{x}$. The various kernel estimates differ in the way weights are assigned to observations in neighboring cells. The weights are defined through a kernel function, $k(\underline{x},\underline{y})$, which in its most general form is a function of the relative location of the cells (directional kernel function). Kernel procedures are further discussed in section 2.5 as an alternative procedure to maximum penalized likelihood but are not explicitly implemented.

The basic form of the penalty term used in model D penalizes the local estimates $a(\underline{x})$ and $b(\underline{x})$ from more global estimates $\tilde{a}(\underline{x})$ and $\tilde{b}(\underline{x})$ obtained by local averaging or interpolation.

The penalty term which is added to the log-likelihood is of the following form:

$$Q_{a,b} = -\frac{P_a}{2}\sum_{\underline{x}}(a(\underline{x})-\tilde{a}(\underline{x}))^2 - \frac{P_b}{2}\sum_{\underline{x}}(b(\underline{x})-\tilde{b}(\underline{x}))^2 \qquad (2.15)$$

or

$$Q_{a,b} = -\frac{P_a}{2}[a(\underline{x})]^T[I-H]^T[I-H][a(\underline{x})] - \frac{P_b}{2}[b(\underline{x})]^T[I-H]^T[I-H][b(\underline{x})] \qquad (2.16)$$

where $[a(\underline{x})]$ and $[b(\underline{x})]$ are column vectors, superscript T indicates transposed matrices, vector I is the identity matrix and H is an interpolation matrix such that,

$$[\tilde{a}(\underline{x})]=[H_a][a(\underline{x})] \qquad (2.17)$$

The degree of smoothness of the solution is controlled by P_a and P_b. Notice that in this case, the same interpolation is used for a and b ($[H_a]=[H_b]$).

The likelihood equations one solves at each iteration, if one considers $\tilde{a}(\underline{x})$ an explicit function of $a(\underline{x})$ is,

$$n(\underline{x})-\sum_m T^*(\underline{x},m)e^{a(\underline{x})-b(\underline{x})m}-P_a[W]_{\underline{x}}[a(\underline{x})]=0 \qquad (2.18)$$
$$m\,n(\underline{x})-\sum_m m\,T^*(\underline{x},m)e^{a(\underline{x})-b(\underline{x})m}-P_b[W]_{\underline{x}}[b(\underline{x})]=0$$

where $[W]_{\underline{x}}$ is the x^{th} row of the matrix

$$[W]=[I-H]^T[I-H] \qquad (2.19)$$

A constraint on the interpolators is imposed. It is desirable that the total number of expected and observed counts inside the region being analysed be the same, that is,

$$\sum_{\underline{x}} n(\underline{x})-\sum_{\underline{x}}\sum_m T^*(\underline{x},m)e^{a(\underline{x})-b(\underline{x})m}=0 \qquad (2.20)$$

Under these conditions, interpolation should satisfy the conditions

$$[1]^T[W]_a[a(\underline{x})]=0$$
$$[1]^T[W]_b[b(\underline{x})]=0 \qquad (2.21)$$

The interpolator chosen by VanDyck for model D is the average over a neighborhood of fixed size and shape around cell $\underline{x}$,

$$\bar{a}(\underline{x})=\frac{1}{k(\underline{x})}\sum_{\underline{x}\in N(\underline{x})} a(\underline{x}) \qquad (2.22)$$

where $N(\underline{x})$ is the set of locations that are neighbors of $\underline{x}$ and $k(\underline{x})$ equals the number of neighbors.

Solution of the equations for each $\underline{x}$ proceeds by iteration. One way to solve is to compute the inverse of the Jacobian of the system of equations 2.18 and to use Newton's method. However, this is not very practical because the number of cells is large. Instead, the following iteration scheme is used:

1. initial values for $a(\underline{x})$ and $b(\underline{x})$ are arbitrarily set equal to zero or set equal to values corresponding to a single estimate of a and b for the whole region,
2. the equations 2.18 are solved successively for each location $\underline{x}$, and the equations for sites which are coupled through the matrix [W] to location $\underline{x}$ are immediately updated for the change in $a(\underline{x})$ and $b(\underline{x})$.
3. after solving the equations for the entire region, the total imbalance on the total counts and the total magnitude (Eq. 2.9) are computed and constants Δa and Δb are added to estimates in each cell to restore the balance.
4. return to step 2 until convergence.

Note that the estimates of $a(\underline{x})$ and $b(\underline{x})$ are dependent of each other because the maximum likelihood equations 2.18 are coupled and that penalties on one parameter will affect the estimates of the other. As will be shown in section 2.3, this dependency is typically not very large.

When the size of historical events is not known with certainty, the above procedure is still applicable if the uncertain event is distributed over different magnitude intervals according to its probability density function. Alternatively, deterministic equivalents m^* can be substituted for m, which are defined such that the estimated recurrence rate for m^* is the same as for m (VanDyck 1985, Chapter 2), the converted magnitude is then,

$$m_i^* = m_i - 0.5b(\underline{x}_i)\sigma^2 \tag{2.23}$$

where σ^2 is the uncertainty on the size of the event i and $b(\underline{x}_i)$ is the estimate of b at the location of the i^{th} event. One can then treat m^*as if they were exact in the previous expressions.

In some cases, independent information exists on the value of $b(\underline{x})$. For example, this information may reflect prior knowledge with respect to the distribuion of b for worldwide or regional data. Inclusion of a prior distribution of $b(\underline{x})$ is done by adding the following terms to each maximum likelihood equation (Eq. 2.18(b)),

$$\frac{1}{\sigma_b^2}\ (b(\underline{x})-\bar{b}) \tag{2.24}$$

where $\bar{b}$ is the prior mean value and σ_b^2 is the prior variance. Note that σ_b^2 is the variance of the slope $b(\underline{x})$ averaged within a given neighborhood of $\underline{x}$. If the area of the neighborhood varies, then also σ^2_b should change. If this were not the case, the prior would become very strong compared to information from the data as the area associated with each $\underline{x}$ decreases.

2.2 Local neighborhoods

The model of the previous section can optionally be used with or without the external specification of sources. The effect of the latter are not as severe on the estimates of seismic hazard as in the case of traditional seismic-sources estimates because of the smooth variation of the parameters inside each source. However, the final estimates can be significantly affected by the sources configuration because of the discontinuity of the estimates at the boundaries of the sources. Another undesirable feature of the estimates is that the smoothness is enforced isotropically, without any regard to the spatial pattern of variation of seismicity. With increasing penalties, there is a decrease in the variance of the estimates but also an increase in the bias of the estimates if sources are not properly specified. The bias is greatest in regions where there are sharp spatial contrasts in the rate of activity.

In the present section, a procedure is proposed which identifies local neighborhoods which are local zones of homogeneity with respect to the seismicity parameters. When these local zones of homogeneity are used to define the interpolation neighborhoods of the previous section, significant discontinuities are preserved, even under large penalties. In addition, because of the homogeneity of the neigborhoods, with increasing penaltties, the variance of the estimates is decreased while bias remains small.

Several methods for the objective identification of local zones of homogeneity have been proposed in the context of seismic source identification and in other fields, such as image processing. In the image processing litterature, local neighborhoods are defined as regions which share the same features and are used to smooth the image within homogeneous zones while preserving edges between distinct regions. In the context of image processing, the purpose of smoothing is to eliminate the high

freuqency component of the image, which in general corresponds to noise. In the case of seismicity, one cannot similarly assume that the high frequency component is due to noise, however, smoothing reduces the variance of the estimates. The simplest techniques for the identification of local neighborhoods are tresholding procedures. A group of connected cells is considered to be homogeneous if a feature, or combination of features, does not deviate by more than a fixed quantitiy across all members. In the case of seismicity, such a rule may be to identify the local neighborhoods through tresholding of the observed recurrence rate in each cell.

This is equivalent to contouring procedures proposed by Caputo et al. (1974) and Chiburis (1981) for the identification of seismogenic provinces.

Both are based on the rate $\lambda(m_0)$ of events larger than a given magnitude m_0 and consist of contouring on the plane estimates $\hat{\lambda}_0$ of λ_0. The difference between the two procedures is the way in which the estimates are obtained. Chiburis uses a moving-average estimator with an exponentially decaying kernel function while Caputo and Postpischl find the estimates by low-pass filtering the empirical earthquake counts. Similar contouring procedures are based on other local indices of seismicity, such as the tectonic flux, which is a quantity proportional to the strain release rate per unit area and unit time (Catteneo et al., 1981); the total energy released per unit area and unit time (Bath, 1956), and (St-Amand, 1956); the log-rate of earthquakes with magnitude in a given range (Kaila and Hari Narain, 1971) and (Kaila et al., 1974).

Criteria based on more than one index of seismicity have also been proposed. For example, (Consentino, 1978) suggests to identify homogeneous zones on the basis of the spatial variation of the parameters a and b, the minimum magnitude for which the exponential recurrence law applies, and the upper bound magnitude. Other authors have emphasized that source identification and seismicity parameter estimation should be based on both historic activity and geologic-tectonic parameters. For example, a

functional relationship between a set of tectonic and seismicity indices and maximum magnitude has been used by (Borissoff, 1977) to produce maps of maximum magnitude for Northern Italy. Various zoning procedures that include historic seismicity as well as geologic and tectonic parameters are described and applied to various regions of the USSR in (Medvedev, 1976).

Another alternative, is to treat the feature vector of geologic or tectonic characteristics at a location as a regressor in the estimation of the rate of activity. Smoothness and discontinuities of the estimates become a direct function of the degree of smoothness of the regressors and of their degree of association with the observed seismicity. For example, (Anderson, 1979) in a study of California earthquakes relates the level of seismicity to the strain rate through a regression procedure. Such a procedure, however, is not at the moment applicable in the Eastern United States given the difficulty in associating seismic activity with identifiable features. Barstow et al. (1981) in an extensive study of the Central and Eastern United States analyses geological and geotectonic factors and their association with the level of seismic activity. For this purpose, he identifies 24 seismically active and 24 non-active sites. Active sites are defined as locations which have experienced one or more events with MMI intensity greater or equal to VII. For these 48 sites, 68 separate characteristics are catalogued within a radius of 61 km of the individual sites. Several statistical procedures were applied to the data set to identify the most discriminating characteristics with respect to the level of activity. The statistical procedures which were applied to the data set are, discriminant factor analysis, principal component analysis, factor analysis, and clustering analysis. The most discriminating characteristics were found to be Pre-Triassic rifts, the total number of faults, the number of fault / intrusive intersections, the earthquake frequency, and the cumulative stress release. Most of the results indicate that the detailed investigation of surface

geologic features in the vicinity of a location is of limited usefulness in evaluating future earthquake hazard. Although certain physical anomalies often occur in regions of strong seismicity, earthquake activity is not always present where such anomalies are found. This means that certain physical conditions are necessary to some degree but are not sufficient for intense seismicity to occur. Barstow et al. (1981) conclude that the historical record remains the primary source of information for modeling future earthquake activity. Notice that instead of using discriminant factor analysis, one could have used logistic regression, which is generally considered superior when causal relationships are analysed (Liao 1986).

Another alternative is to use geologic or geotectonic information to form an empirical prior estimate on the seismicity parameters for a given classification. The posterior distribution of the parameters is then computed as a function for the seismicity observed locally (Esteva, 1969). However, this procedure is not useful in a region where geologic and tectonic characteristics are not informative with respect to levels of activity.

The proposed procedure, which is edge-preserving, is to define the local neighborhoods on the basis of a statistical test of homogeneity for each of the seismicity parameters. Appropriate tests for the parameter a are those for the equality of the recurrence rate of Poisson processes. A test for equality of the parameter b (exponential distribution) can be found in Epstein and Tsao (1953). However, this test is based on a ranking of the sample of earthquakes in increasing order of magnitude and is not applicable if the sample is incomplete. In addtion, other difficulties arise because of ties in the ordering of observations because most historical events are reported on a discrete scale. Other tests for the equality of b can be formulated in terms of profile analysis, comparison of distribution functions, non-parametric tests for the equality of medians, and categorical data analysis (Gibbons, 1985) but require large amounts of data in each cell.

Criteria for defining local neighborhoods for both a and b simultaneously can be stated in terms of tests of Poisson homogeneity over different ranges in magnitude. Consider k geographical cells of areas $A_1,...,A_k$ and partition the range of magnitude values into r intervals. The rate of events generated by the i^{th} cell in the j^{th} magnitude interval is denoted by λ_{ij} so that $\underline{\lambda_i}=\{\lambda_{i1},...,\lambda_{ir}\}^T$ is the rate vector for cell i. One is interested in testing

$H_o : \underline{\lambda_i} = A_i \underline{\lambda}$ $\{i=1,...,k\}$ for some vector $\underline{\lambda}$

against

$H_1 : \underline{\lambda_i} \neq A_i \underline{\lambda}$ for at least one i

Two widely-used tests for hypotheses of this type are the Chi-square $\{\chi^2\}$ and likelihood-ratio (LR) tests. In both cases one uses the following quantities:

n_{ij} : number of events in cell i for magnitude range j

$n_{.j}$: $\sum_i n_{ij}$ = total number of events in magnitude range j

A : $\sum_i A_i$ = total area of cells

A_i' : A_i/A = fraction of area in cell i

The Chi-square test consists of calculating the statistic

$$\chi^2 = \sum_{j=1}^{r} \sum_{i=1}^{k} \frac{(n_{ij} - A_i' n_{.j})^2}{A_i' n_{.j}} \qquad (2.25)$$

The null hypothesis is accepted if $\chi^2 < \chi^2_{r(k-1),\alpha}$, where $\chi^2_{n,\alpha}$ is the $(1-\alpha)$-fractile of the Chi-square distribution with n degrees of freedom (see for example (Bhapkar, 1980) page 369). The likelihood-ratio test is based on

$$LR = -2 \sum_{j=1}^{r} \sum_{i=1}^{k} n_{ij} \ln\left(\frac{A_i' n_{.j}}{n_{ij}}\right) \quad for\ n_{ij} \neq 0$$

and consists of accepting H_o if $LR < \chi^2_{r(k-1),\alpha}$. Both tests are approximate, but they are accurate and give very similar results for large values of the products $A_i' n_{.j}$.

A special case of interest is when r=1, i.e. when homogeneity of the cells is evaluated in terms of the total rates λ_i and the hypotheses H_o and H_1 are

H_o: $\lambda_i \propto A_i$
H_1: $\lambda_i \not\propto A_i$

An even more special situation is when r=1 and k=1. In this case there is an exact, uniformly most powerful test based on the binomial distribution (Lehman, 1959) P.140: without loss of generality, the two regions are numbered such that $n_1/A_1 > n_2/A_2$. Knowing the total number of events $n=n_1+n_2$ and the probability under H_o that an event occurs in region 1, $p_1 = A_1/(A_1+A_2)$, one can use the binomial distribution with parameters n and p_1 to calculate the probability P where

$P = P[\text{number of events in } A_1 > n_1 | n, p_1]$

The resulting region of acceptance is illustrated in Figure 2.1 for p=0.5. A point which is clearly brought out by this figure is that it is impossible to detect differences in the rates of cells with small numbers of counts. For a discussion of the power of the test seee (Przyborowski, 1939).

For the application of the test of equality of recurrence rates one must determine the area of each cell. The area is equal to the spatial area of the cell multiplied by the observation time corrected for incompleteness. Because incompleteness varies as a function of time and magnitude, a mean observation time for the total rate of the cell must be defined.

The total equivalent time of observation is defined such that the total expected incomplete rate in each cell is equal to the sum of the expected incomplete rate for each magnitude interval. The resulting equivalent period of observation is,

$$T_{eq}(\underline{x}) = \frac{\sum_{m=m_o}^{m_1} T(\underline{x},m)e^{-b(\underline{x})m}}{\sum_{m=m_o}^{m_1} e^{-b(\underline{x})m}} \qquad (2.26)$$

For the application of the test, $b(\underline{x})$ and $P_D(\underline{x},t,m)$ have to be specified a priori.

Alternatively, one can determine the equivalent period of observation as a function of the observed rates for each magnitude interval,

$$T_{eq}(\underline{x}) = \frac{\sum_{m=m_o}^{m_1} N(\underline{x},m)}{\sum_{m=m_o}^{m_1} \frac{N(\underline{x},m)}{T(\underline{x},m)}} \tag{2.27}$$

The spatial extent of the local neighborhood is defined through the number of rows of cells (M) around cell $\underline{x}$ to which the test is applied. The test is first applied to each cell within the region defined by M and connectivity among the cells which pass the test is then enforced. The above procedure is repeated for each cell of the region and reciprocity required among the local neighborhoods, i.e., cell $\underline{y}$ is in the local neighborhood of $\underline{x}$ ($N_{\underline{x}}$) only if cell $\underline{x}$ is in the local neighborood of cell $\underline{y}$ ($N_{\underline{y}}$).

The interpolation functions are defined as the average of the parameters within the local neighborhoods. In the matrix notation of section 2.1, the only non-null terms in row $\underline{x}$ of the matrix [H] are those corresponding to cells which have been included in the local neighborhood $N_{\underline{x}}$. This implies that many terms of the matrix [H] and [W] are null. The elements of the matrix [H] are defined as

$$h_{\underline{x},\underline{y}} = \begin{cases} \dfrac{k(\underline{x},\underline{y})}{\sum_{\underline{z}\epsilon N_{\underline{x}}} k(\underline{x},\underline{z})} & if\ j \in neighborhood\ of\ \underline{x} \\ 0 & otherwise \end{cases} \tag{2.28}$$

where $k(\underline{x},\underline{y})$ can be any weighting function, in this case $k(\underline{x},\underline{y})=1$ was used.

Great computational savings are obtained by storing only the non-null terms of the smoothing matrix [W]. The terms $w_{\underline{x},\underline{y}}$ are non-null only if there is overlap between the neighborhoods of cells $\underline{x}$ and $\underline{y}$.

These are defined as follows,

$$w_{\underline{x},\underline{y}} = \sum_{\underline{z} \in N_{\underline{x}} \cap \underline{z} \in N_{\underline{y}}} a_1 \bullet a_2 \qquad (2.29)$$

where

$$a_1 = \frac{1}{N_{\underline{z}}} \quad if\ \underline{x} \in N_{\underline{z}}$$

$$a_1 = 1.0 \quad if\ \underline{x} = \underline{z}$$

$$a_2 = \frac{1}{N_{\underline{z}}} \quad if\ \underline{y} \in N_{\underline{z}}$$

$$a_1 = 1.0 \quad if\ \underline{y} = \underline{z}$$

The shape and size of the two dimensional averaging windows are a function of the homogeneity of the cells surrounding each cell. As will be shown in section 2.3 this estimator automatically identifies spatial discontinuites of the recurrence rate and in the limit can generate, when supported by the historical data, earthquake recurrence models of the seismic-source type. Local neighborhoods are used here only for the estimation of a($\underline{x}$). Similar neighborhoods could be defined for b($\underline{x}$) using the previously mentionned procedures, but such neighborhoods would be less useful, due to the high statistical uncertainty on b given the small amount of data in single cells. For the definition of local neighborhoods of b, two simpler options are explored in Chapter 3: the first is to assign the same neighborhood as for a($\underline{x}$) assuming that the same mechanisms which control the spatial variation of a($\underline{x}$) control the spatial variation of b($\underline{x}$), the other is to keep the local neighborhoods fixed in shape and size to the immediately neighboring cells.

2.3 Applications

The procedures developed in the previous section are now applied to two regions. First to New England, using the Chiburis catalog and second, to the Eastern United States using the catalog compiled in the context of EPRI (1985). In both applications,

earthquakes which have been identified as aftershocks in the catalog are removed prior to the analysis.

2.3.1 Chiburis catalog

Only earthquakes with $I_0>4$ are used in the analysis. Category m=0 corresponds to $I_0=4$, category 5 corresponds to the largest intensity found in the catalog, $I_0=8$ (Figure 2.2). It is clear from Figure 2.2 that seismicity is highly non-homogeneous and that the eastern Quebec, Boston and Long Island regions exhibit higher activity. If I_0 is reported, but $\Delta I_0=I_{o,max}-I_{o,min}$ is not zero, the prior distribution of I_0 is assumed to be normal with mean value $(I_{o,min}+I_{o,max})/2$ and $\sigma_{I_o}=0.5$ truncated at $\pm 3\sigma_{I_o}$ and discretized to a mass density function p'_m for different categories m (including m<0) (section 2.1.2). Earthquakes with I_0 not reported are assumed to be normally distributed with mean value (Chiburis 1981),

$$E'[I_o] = \frac{(M-1)}{0.6} \tag{2.30}$$

where M is the reported instrumental size measure and E' refers to the prior expected value of I_0. The standard deviation σ_{I_o} is assumed equal to 0.6 (VanDyck 1985, Chapter 2). Based on a preliminary analysis of the data, the spatial variation of completeness is represented by only two regions, which are the coastal region and the remainder (Figure 2.3). The simplicity of the configuration follows from the sparseness of the earthquake counts in much of the region. For instance, the locations in the Atlantic were not treated as a separate region because the counts are so low that the probability of detection would be impossible to determine. Adding this region to areas over land does not introduce any changes in the estimates of these regions. From a practical point of view, the present choice of only two regions corresponds to assuming that recurrence rates are small in this part of the region. Inclusion of a separate region to account for early settlements around Quebec and Montreal has been

also considered. In this case, it was found that estimates of P_D are very similar to those in the surrounding region. The penalty coefficient P_α which controls the smoothness of the estimates of α_{tm} with time and size in each incompleteness region is set equal to 20. This corresponds to a moderately smooth change of the estimates. For the probability of detection it is assumed that all earthquakes have been reported since 1950, hence,

$$\alpha_{tm} = 1.0 \quad for\ t \geq 1950 \tag{2.31}$$

It is further assumed that, for I_0=7 and 8, the catalog is complete since 1860 and 1625 respectively.

Probabilities of detection for the two incompleteness regions are given in Table 2.I for penalties P_a=5 and P_b=50 . As can be seen, the historical record for the coastal area of New Englnd is much larger than for the remainder.

The test of Poisson homogeneity (section 2.4) was applied to the catalog discretized in one and half degree cells at different significance levels. Figure 2.4 illustrates results obtained for one-degree cells at a significance level of 20% when the test of homogeneity is not limited to immediately neighboring cells. In this figure, the cells for which the neighborhoods are identified are indicated with an asterisk (*) and the neighborhood of homogeneous cells by the symbol (1). Also shown in the figure, are the total counts of events in each cell as well as a decomposition of the number of counts as a function of intensity. The anisotropy of seismicity across the region is clearly indicated. In particular, regions of homogeneous seismicity are identified along the Atlantic coast (region A), in southeastern Quebec (region B), in the Apalachian Mountains (region C) and over the continental shelf (region D).

Following, is a comparison of estimates obtained for two different signifiacance levels of the test of homogeneity (α= 0% and 15%) and for different penalties P_a. Figure 2.5 shows estimates of $a(\underline{x})$ obtained for α=0% (fixed neighborhoods), M=1 (the test is

applied only to immediately neighboring cells), and selected penalties P_a, for a spatial discretization into half degree cells. Note that with an increase or decrease of the significance level α, one can produce nested sequences of local neighborhoods. In these and in later plots, a is the log-rate of events of MM intensity 4 per year per cell of unit equatorial degree ($111.1^2 km^2$). For example, a value of a=-1.0 indicates that earthquakes of intensity I=4 occur in a half-degree cell at lat. 42°N at a rate of $e^{-1}\cos(42^o)/4=0.06$ events/year. For low penalties on $a(\underline{x})$, the estimates are very contrasted and their variance is large. With an increase in the penalty, there is a decrease in the variance of the estimates and an increase in the bias as the solution is pulled towards the average for the entire region. Objective procedures for the selection of the optimal degree of smoothness are proposed in Chapter 3. Figure 2.6 shows the estimates of $b(\underline{x})$ as a function of P_b and α. For the lowest values of P_b, the spatial trend of increasing $b(\underline{x})$ from the southwest to the northeast as well as the local maximum of $b(\underline{x})$ in eastern Massachusetts is clear. Increasing P_b gradually removes these features, first, the local maximum, then the local linear trend. Although there is a slight change in the estimates of $a(\underline{x})$, which counteracts the increase and decrease of $b(\underline{x})$, the global effect at high intensities is to increase the recurrence rates for areas in the central part of the region and to decrease the rates in the northeast corner.

Various contour plots of $a(\underline{x})$ are shown in Fig. 2.7 to illustrate the effect of the significance level α, the number of rings of cells around $\underline{x}$ to which the local neighbornood is confined (M), and the penalty P_a. The contrast in $a(\underline{x})$ between more and less active areas increases with the significance level α. For α=0.01, only few of the neighborhoods have an irregular shape and the estimates $a(\underline{x})$ are similar to those for fixed neighborhoods (compare Fig. 2.5(b) to Fig. 2.7(a)). For α=15%, the function $a(\underline{x})$ displays plateaus of nearly constant activity, in some cases connected by gradual ramps and is relatively insensitive to M and P_a. The estimates $a(\underline{x})$ preserve a high

level of contrast for a high penalty P_a, which is not the case when fixed neighborhoods are used (compare with Fig. 2.5(d)). Features of the seismicity that are accentuated with an increase of α are the seismic activity along Lake Champlain and the Hudson river (74°W 43°N), the plateaus of high activity in eastern Quebec and along the Atlantic coast, and the high peak of activity near Cape Ann and Newburyport.

As was mentionned in the section 2.2, it is not feasible in this application to use a test of homogeneity on the mean magnitude, or of equality of the probability distribution function in magnitude in each cell, to identify local homogeneous neighborhoods for smoothing $b(\underline{x})$. An alternative is to use the same neighborhoods which were identified for the smoothing of $a(\underline{x})$ Figure 2.6(c). These estimates are slightly less smooth than those obtained for fixed neighborhoods. In particular, the procedure identifies an isolated cell in southeastern Quebec where the estimate of b is very different from that in neighboring cells. Contours of constant $b(\underline{x})$ follow those of $a(\underline{x})$ in some areas, for example, in the southwestern corner of the region. Note that in regions with low levels of activity, the estimates are dominated by the prior mean on b (m_b'=1.3) which accounts for the lack of discontinuities in the estimates along the coast (for $a(\underline{x})$, the estimates are set to very small values in the absence of activity).

2.3.2 EPRI catalog

In this section, models of seismicity are estimated for the entire Eastern United States. Models are fitted for different penalties and homogeneous interpolation neighborhoods. The estimation procedure which accounts for expert zonations is then applied for four zonations recently proposed for the Eastern United States. The spatial grid size for this application is one degree cells. The probabilities of detection are fixed to values obatined in the context of EPRI (1985). The following briefly describes, their estimation. Eastern North America is partitioned into several incompleteness regions,

inside which the probability of detection is assumed spatially constant and a fucntion only of time of occurence and magnitude (Figure 2.8). The incompleteness regions have no relationship with seismicity and seismogenic sources. Three pieces of information have been used in the definition of the incompleteness regions: 1. the evolution in time and space of the population and of the seismic instrumentation, 2. results of previous models of incompleteness and 3. the geographical extent of the regional catalogs that have been used in compiling the EPRI catalog.

Magnitude is discretized into 0.6 unit intervals starting from 3.3 to 7.5. Reasons for this discretization of magnitude are documented in EPRI (1985), and are related to the completeness of the historical record and to the accuracy of conversion from Modified Mercalli intensity to body-wave magnitude. Discretization in time is based on demographic history and instrumentation and the availibility of different types of documents (diaries, newspapers, technical publications,...). Maximum magnitude has been set equal to 7.5 everywhere in Eastern North America for the estimation of seismicity, however, it has little influence on the estimates of $a(\underline{x})$ and $b(\underline{x})$ (the last magnitude interval contains the largest event recorded for this region). For all regions and for magnitudes greater than 3.3, the catalog is assumed complete since 1975, which is a time when the instrument network has been improved considerably. A time-magnitude enveloppe which is indicated by a solid line in Table 2.II, identifies the most complete portion of the catalog which is used in fitting the models in the following applications. The earlier portion of the catalog is not used because the uncertainty on the estimates of the probability of detection is very large. The corresponding events are shown in Figure 2.9. In this application, the penalty on $b(\underline{x})$ is set to a very large value to generate almost constant estimates for the region (section 3.3.1).

The estimates of $a(\underline{x})$ are shown for α=0% and α=10% and different penalties P_a in

Figures 2.10 and 2.11. One may notice that when α=10% the contrast in the estimates is preserved where the discontinuities are the most significant even with large penalties. In particular, there appears to be two major regions. The first region is in the northeastern United States and extends to the west to central New York state, and to the south to northern New Jersey. Within this region, one can distinguish three extended areas of higher activity centered around Newburyport, Charlevoix, and the Ottawa River Valley. The second region is in the southeastern United States. Within this region, there are areas of larger activity, this time not as extensive spatially as in the previous zone, centered around Charleston, eastern and western Tennesse, and eastern Virginia. It is interesting to compare contours of $a(\underline{x})$ with source configurations which have been proposed in the litterature for the same region. Fig. 2.12 shows a selection of source zonations from EPRI (1985) and, Fig. 2.13 shows sources proposed by (Barosh, 1986). Some of these were determined from an analysis of past seismicity and clearly follow the contours of $a(\underline{x})$ in Fig. 2.11. However, many sources based on geological information do not show any association with the patterns of historical seismicity. In the following section, a procedure is developed which allows the incorporation of expert opinion on the existence of homogeneous zones of seismicity in the estimation of local neighborhoods.

2.4 Incorporating Expert Opinion in the Local Estimation of Seismicity

As mentionned in section 2.1, there are regions of the eastern and central U.S. where seismic zoning is controversial and earthquake hazard is sensitive to source geometry. The Charleston region is an example. To exemplify the importance of the problem, Fig. 2.12 shows seismic source configurations proposed for the eastern U.S. by various seismologists (EPRI, 1985). Differences are substantial, especially considering the fact that seismologists were provided with the same information, including sets of possible source configurations.

The proposed procedure is a modification of the test from classical statistics used to identify the local neighborhoods (section 2.3). The modification is with respect to the significance level (α_o) of the test for accepting the hypothesis of homogeneity (H_0) (which corresponds to the probability of rejecting H_0 given that the hypothesis is true). Assuming that H_0 is true and given a large sample of pairs of observations (N_1,N_2) from the same Poisson process with parameter λ, the number of times the test is accepted to the number of times the test fails defines an odds ratio R_0 which at the limit tends to

$$\frac{1-\alpha_o}{\alpha_o} \tag{2.32}$$

where α_0 can be interpreted as a misclassification rate (MR_0).

Cells are defined by superimposing the grid of cells used for estimating the seismicity parameters (section 2.1) and the sources (e.g. Fig. 2.14(a)). The cells which are smaller than a certain fraction of their original size are merged with the largest neighboring cell in the same source (otherwise the test of homogeneity is not powerful given the small number of observations). The test is applied to all distinct pairs of cells for the region, and the number of times the test passes at a significance level α_0 is counted. The observed ratio $\hat{R}_0$ of the number of times the null hypothesis is accepted to the number of rejections is a measure of the homogeneity of the region and will be in general smaller than R_0 in Eq. 2.32 (another measure is the observed misclassification rate ($\hat{MR}_0$ which is related to $\hat{R}_0$ through, $\hat{R}_0=(1-\hat{MR}_0)/\hat{MR}_0)$). For a well-specified source configuration, the odds ratio for pairs of cells within the same source ($\hat{R}_{in}$) should be greater than $\hat{R}_0$ while the odds ratio for pairs of cells in neighboring sources ($\hat{R}_{out}$) should be smaller than $\hat{R}_0$. Assuming that the significance level α_0 is adequate for identifying local neighborhoods over a region with an odds ratio $\hat{R}_0$, the significance level of the test is modified depending if two cells are within the same source or not, to preserve the following ratios

$$\frac{\hat{R}_o}{\hat{R}_{in}} = \frac{\frac{1-\hat{MR}_o}{\hat{MR}_o}}{\frac{1-\hat{MR}_{in}}{\hat{MR}_{in}}} = \frac{\frac{1-\alpha_o}{\alpha_o}}{\frac{1-\alpha_{in}}{\alpha_{in}}}$$

Because some sources in a partition may be more informative than others, it is preferable to define the odds ratio for each source ($\hat{R}_{ii}$) and for each pair of neighboring sources ($\hat{R}_{ij}$). The odds ratio $\hat{R}_{ii}$ is then an indicator of the internal homogeneity of source i while $\hat{R}_{ij}$ is a measure of the significance of the boundaries between sources i and j. If $\hat{R}_{ii} > \hat{R}_o$, then the cells within source i belong to a region which is more homogeneous than the unpartitioned region and the adjusted test is more lenient than previoulsy ($\alpha_{ii}<\alpha_o$). If $\hat{R}_{ii} < \hat{R}_o$, the region exhibits more contrasts of seismicity than the original region as a whole and the test becomes more stringent ($\alpha_{ii}>\alpha_o$). The model for the odds ratio is as follows,

$$ln(\hat{R}(k,l)_{ij})=\beta_{o_{ij}}+\beta_{ij}K(k,l) \tag{2.33}$$

where k is a cell in source i, and l is a cell in source j, and K(k,l), is an indicator function:

K(k,l)=0, if i = j
K(k,l)=1, if i ≠ j

$\hat{R}(.,.)_{ij}$ can be interpreted as

$$\frac{P[\textit{cells from source i and j are homogeneous}]}{1-P[\textit{cells from source i and j are homogeneous}]} \tag{2.34}$$

In consequence, if the odds ratio for a source is larger than the odds ratio in the absence of any information on source zonations, that particular source identifies an homogeneous group of cells informative and the significance level of the test for pairs of cells within that source is lowered to allow greater internal smoothing. If the odds ratio for pairs of cells in neighboring sources is smaller than the odds ratio in the absence of any information on source zonations, the boundary between the two sources

identifies a significant discontinuity in the rate of seismicity and the significance level is lowered to decrease the likelihood that local neighborhoods are identified across the boundary. If a boundary is found significant but seismicity is uniform locally, the boundary is ignored by the procedure, so that boundaries are preserved only where they are found to be locally significant. In consequence, the solutions do not necessarily reproduce the seismic-sources estimates even if a source is found to be significantly homogeneous according to the criterion. If an anomaly is found within an hypothesized homogeneous source, the odds ratio for the zone decreases significantly and the anomaly is extracted in the fitting of the model. It is interesting to note that if a source configuration is found to be non-informative, the estimates are identical to those which would be estimated in the absence of the zonation. In consequence, the proceudre is robust with respect to the misspecification of the seismic source configurations.

In a preprocessing step, one may eliminate non-significant boundaries between regions. If two neighboring sources have similar patterns of observed seismicity, the previous procedure will assign the same significance level for tests between cells inside or between the two sources, in effect ignoring the boundary specified by the expert. The removal of unnecessary boundaries before the final estimation reduces the number of split cells and increases the sample size for the combined source. Similarity between pairs of sources is measured through a test of association on a (2X3) contingency table of the number of times the null hypothesis is accepted and rejected within and between the sources (Table 2.III) at the significance level α_0. The test statistic used is (Bishop, 1975)

$$X=\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{2.35}$$

where

$$\hat{n}_{ij}=\frac{n_{i.}n_{.j}}{n_{..}} \tag{2.36}$$

X is distributed as a χ^2_2 and the two sources are merged if the observed statistic is smaller than $\chi^2_{2,1-\alpha}$ where α is a given significance level.

	# of acceptances	# of rejections	total # of tests
both cells in source 1	n_{11}	n_{12}	$n_{1.}$
both cells in source 2	n_{21}	n_{22}	$n_{2.}$
one cell in each	n_{31}	n_{32}	$n_{3.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

Table 2-III: Contingency table for testing the association between sources EPRI(1985)

In the followimg section, the above procedures are applied to a region covering the Atlantic seabord and the Apalachian region. Several source configurations which have been recently suggested in the litterature are used to demonstrate how they affect the estimates of $a(\underline{x})$.

2.5 Application

Four source configurations are being considered for the application of the previous procedure. The first three were proposed by Thenhaus (1987) and are based on information other than seismicity. For all of these sources, it is assumed that the observed seismic activity is due to reactivation of faults formed during prior tectonic regimes. Each of the source configurations is a regional representation of a type of geologic structure or process that has been suggested in the litterature to be responsible for seismicity somewhere in the region. The last configuration was proposed by Woodward-Clyde and Associates within the context of the (EPRI, 1985) project and

was obtained primarily by contouring homogeneous regions of historical seismicity (Figure 2.14).

Seismic sources that are classified according to the structure and tectonic history of a region are defined as structure-based sources. Thenhaus (1987) propose such a zonation based on faults ranging in age from the late Precambrian through early Mesozoic periods (source C, Figure 2.15). The region is divided into two major zones. The first is a zone which follows the Apalachian mountain range and the second is a region that follows the Atlantic coast. Several smaller sources are identified within the second source. Some of these are very small and have a negligible effect on seismic hazard given that in most cases they have been the site of relatively little activity (note that Charleston in not one of these cell).

The remaining two source configurations are based on a classification which relates the structural geologic history of faulting to contemporary faulting and historically observed seismicity. This type of zonation is known as process-based seismic source zones and represents the geographic extent of inferred ongoing geologic processes (crustal uplift and subsidence in this case). Unlike the structure-based zones, specific types of primary crustal structures are unimportant to the definition of the zones. The two zonations are based on two different sources of information on vertical movement. The first (source D, Figure 2.16) is based on geodetic measurements and partitions the regions into areas of (1) rapid uplift, (2) no movement to slow uplift, (3) no movement to slow subsidence and, (4) rapid subsidence. The partition attempts to represent positive and negative movements as a regionally varying continuum throughout the East. The characterization of areas as to positive versus negative movement, or, rapid movement versus no movement has no a-priori implications with regard to seismic potential. The second source (source E, Figure 2.17) integrates information about the regional basement structure, Cenozoic structural framework, and other select geologic information with the observed vertical crustal movements.

In the following, four cases are considered. First, the model is fitted with fixed neighborhoods (α=0%), second, the model is fitted with local neighborhoods at a significance level of 10%, third, the model is fitted with local neighborhoods at a significance level of 10% when sources are introduced, and finally, estimates are obtained for the case when the significance level is modified according to the sources to which belong the cells. These different cases are considered to separate effects from the local neighborhoods, the modified grid of cells from the introduction of the sources, and the effect of the modified significance levels. The effect of the procedure is evaluated for two penalties on $a(\underline{x})$ corresponding to low and intermediate smoothing (P_a=5, 50).

The earthquake catalog used in this sutdy is the catalog from (EPRI, 1985). For the purpose of this section, the penalty on $b(\underline{x})$ is fixed to a very high value which results in an almost constant b value (P_b=1000) (Figure 2.18(d)), and the local neighborhood for interpolation is limited to only immediately neighboring cells. The grid size for the discretization in space is one square degree cells. Magnitude is discretized from 3.3 to 7.5 in 0.6 intervals. The time-magnitude envelopes used for estimating the model and the probabilities of detection correspond to portions of the catalog for which $P_D \geq 0.5$ (Table 2.II). Split cells smaller than 0.2 square degrees at the boundaries between sources are eliminated by merging them to the largest neighboring cells within the same source.

Figures 2.19, 2.20, 2.21, and 2.22, parts a and b, show estimates obtained with isotropic smoothing (i.e. α=0%) and anisotropic smoothing (α=10%) in the absence of any external information about source configuration. The influence of the penalty for fixed neighborhoods (α=0%) is such that for large penalties (P_a=50) the high seismic activity around Charleston and Boston is dissipated throughout the neighboring cells. The increase in the significance level leads to more contrasts in the seismicity estimates, which are preserved with increases in the penalty on $a(\underline{x})$.

Part c of the same figures shows the effect of introducing a source configuration and therefore splitting some of the cells. Estimates of $a(\underline{x})$ and $b(\underline{x})$ are obtained with a test of Poisson homogeneity at a significance level of 10%. Note that because the interpolation neighborhoods include only the immediaely neighboring cells, and the size of some of the cells is reduced by the introduction of the source boundaries, the estimates of $a(\underline{x})$ can be locally different from part b of the same figures even if the significance level has not been modified.

Tables 2.IV through 2.VII summarise the results for fitting the model of equation 2.33 to the four source configurations considered. For the estimation of $R(i,j)_{kl}$, only pairs of cells with non-zero total observations were used for the testing. For example, considering the one provided by Woodward-Clyde (Figure 2.14(c)), the odds ratio for the whole region without any information about sources ($\hat{R}_o$) is equal to 2.47. For source 18, the internal odds ratio ($\hat{R}_{18,18}$) is equal to 12.50 (the total number of tests is equal to 270 while the total number of failed tests is equal to 20. The coefficients β_o and β_1 are defined in Eq. 2.33. In this case, the modified significance level for cells within source 18 is $\alpha_{18,18}$=0.021, and for pairs of cells, one from source 18 and one from source 12, $\alpha_{18,12}$=0.134, indicating that source 18 is internally homogeneous and that its level of activity differs from source 12. The last column corresponds to the test statistic for the degree of association between sources.

For source configuration C, the odds ratios for cells within the same source ($\hat{R}_{in}$=2.90) and for cells in different sources ($\hat{R}_{out}$=2.46) are both smaller than the odds ratio for the whole region without any partition ($\hat{R}_o$=4.24), indicating that the partition is not informative as a whole with respect to the spatial distribution of seismicity (Figure 2.15). The odds ratios for individual sources ($\hat{R}_{ii}$, e .g. $\hat{R}_{11}$=3.18, $\hat{R}_{22}$=2.57) are also lower than $\hat{R}_o$ except for the smaller sources where no test is performed internally (sources 4,6,7,and 8). The pattern of seismicity inside the individual sources is more

variable than when the whole region is considered as a single source. As a consequence, the significance level of tests between cells is increased but the effect on the estimates of $a(\underline{x})$ is negligible (compare Figures 2.19(c) and (d)). The main effects of including this source configuration are the identification of the seismicity around Charleston as an anomaly , and to locally modify the estimates in the middle of the region.

For source configuration D, three pairs of sources are found not to be significantly different and are merged (1 and 3, 4 and 6, 9 and 10). The final partition results in sources that are in general slightly less homogeneous internally than the whole region as a whole ($\hat{R}_{in}=2.51 < \hat{R}_o=2.88$). The boundaries between sources appear to be well defined however ($\hat{R}_{out}=2.21 < \hat{R}_o=2.88$). The main effect of this source configuration is again to extract the seismic anomaly centered on Charleston.

For source configuration E, the partition results in sources which are not on average as homogeneous as the original region ($\hat{R}_o=3.71$, $\hat{R}_{in}=3.30$, $\hat{R}_{out}=2.73$). Individual sources which are merged are 3-4, 5-6 and 10-12, and sources which are found to be individually more homogeneous than the original region are 1,3,5, and 6. Again, in the case of the source which contains Charleston, the odds ratio is very small and the modified level of significance is efficient in extracting the anomaly.

For the source configuration suggested by Woodward-Clyde, the odds ratios indicate that the zones of activity are well delimited ($\hat{R}_o=2.47$, $\hat{R}_{in}=6.28$, $\hat{R}_{out}=1.22$). Sources which are being merged are (2-4-5), (8-22), (13-14), (16-17), and (18-20). Not surprisingly, most boundaries between remaining sources are found to be significant ($\hat{R}_{ij} < \hat{R}_o$) and are enforced in the estimation of $a(\underline{x})$. For example, sources (3-4-5-6) and (16-17) result in more homogeneous estimates of $a(\underline{x})$ (compare Figures 2.22(c) and (d)).

In conclusion, few of the specified sources are validated by the actual distribution of

earthquakes. Of the three configurations proposed by (Thenhaus et al., 1987), configuration E appears to be the most informative with respect to the rate density of events. However, the procedure has the advantage of being robust with respect to possibly bad source configurations. Notice that although sources C and D did not have much influence on the estimates of $a(\underline{x})$, these configurations might be informative for other aspects of the seismicity, such as the maximum magnitude, and characteristic events, which are not considered here.

2.6 Kernel Estimation of Seismicity Parameters

In this section, an alternative model is presented for estimating seismicity in the EUS. Smoothness is again a function of local neighborhoods but is imposed through a kernel function instead of maximum penalized likelihood. A kernel function is essentially a weighting function which is a function of the distance between the location at which the estimate is required and the locations of the observations. In its most general form, the kernel function can be defined in the space of $(\underline{x},t,m)$ and generate a completely non-parametric formulation of seismicity. Such a model may be valuable in identifying migrations of seismicity and other deviations from the usual model assumptions.

As was pointed out in section 2.1, $N(\underline{x})$ and $M(\underline{x})$ are sufficient statistics for the estimation of the seismicity parameters $a(\underline{x})$ and $b(\underline{x})$. In consequence, to implement a kernel estimation procedure, one can define kernel function for $N(\underline{x})$ and $M(\underline{x})$. If one wishes to specify two independent kernel functions for $a(\underline{x})$ and $b(\underline{x})$ one needs to first isolate a and b in the likelihood equation.

Eliminating $a(\underline{x})$ from equation 2.11, one obtains;

$$-M(\underline{x})+N(\underline{x})\frac{\sum_m T(\underline{x},m)me^{-b(\underline{x})m}}{\sum_m T(\underline{x},m)e^{-b(\underline{x})m}} = 0 \qquad (2.37)$$

which can be solved for b($\underline{x}$).

From the above equation, it is clear that spatial smoothness of b($\underline{x}$) is related to smoothness of M($\underline{x}$), N($\underline{x}$) and T($\underline{x}$,m). In consequence, smooth estimates of b($\underline{x}$) can be found by replacing M($\underline{x}$), N($\underline{x}$) and T($\underline{x}$,m) in the previous equations with smoothed values as follows,

$$M^b(\underline{x})=\sum_{\underline{y}} K_b(\underline{x}-\underline{y})M(\underline{x})$$

$$N^b(\underline{x})=\sum_{\underline{y}} K_b(\underline{x}-\underline{y})N(\underline{x})$$

$$T^b(\underline{x},m)=\sum_{\underline{y}} K_b(\underline{x}-\underline{y})T(\underline{x},m) \qquad (2.38)$$

A similar procedure cannot be applied to the determination of a($\underline{x}$) because b($\underline{x}$) is initially unknown. However, one may proceed by imposing smoothness on the cumulative counts, $\sum_m T(\underline{x},m)e^{a(\underline{x})-b(\underline{x})m}$. In this case, a different kernel function K_a can be used to allow different smoothness of a($\underline{x}$) and b($\underline{x}$). The estimate of the a-parameter is then found from,

$$N^a(\underline{x})-\sum_m T^a(\underline{x},m)e^{a(\underline{x})-b(\underline{x})m}=0. \qquad (2.39)$$

Equations 2.38 and 2.39 are solved numerically for a($\underline{x}$) and b($\underline{x}$).

2.7 Conclusions

The main conclusions of this chapter are that:

1. A convenient procedure for preserving the anisotropic nature of the earthquake generating process in regions of intraplate seismicity is to define local homogeneous interpolation neighborhoods for each of the parameters to be fitted. For the parameter a($\underline{x}$), a test for the equality of the recurrence rate in neighboring cells is recommended.

The resulting model preserves significant discontinuities of the recurrence rate of earthquakes, and does not require the external specification of sources. The most significant discontinuities are preserved even for large penalties which proves to be an improvement over previous procedures. The procedure at the limit reproduces the seismic-source estimates of seismicity if warranted by the data.

2. The previous procedure can be easily extended to include information provided by experts with respect to possible zones of homogeneous seismicity. The proposed partitions are enforced only if they are validated with respect to the historical seismicity and insures that the procedure is robust with respect to bad source configurations.

Incompleteness region : 1

Int.	Time period 1	2	3	4	5
4	0.	0.095	0.428	0.744	1.
5	0.	0.095	0.428	0.947	1.
6	0.	0.31	0.743	0.947	1.
7	0.598	0.907	1.	1.	1.
8	1.	1.	1.	1.	1.

Time period

1 : 1627-1780
2 : 1780-1860
3 : 1860-1910
4 : 1910-1950
5 : 1950-1980

Incompleteness region : 2

Int.	Time period 1	2	3	4	5
4	0.12	0.317	0.721	0.93	1.
5	0.12	0.317	0.91	1.	1.
6	0.301	0.745	0.93	1.	1.
7	0.921	0.973	1.	1.	1.
8	1.	1.	1.	1.	1.

Table 2-I: Probabilities of detection and periods of observation for the Chiburis catalog

REGION 1

MB*	1	2	3	4	5	6
3.6	0.00	0.00	0.03	0.39	0.71	1.00
4.2	0.00	0.00	0.09	0.85	+1.00	1.00
4.8	0.00	0.00	0.27	-0.86	1.00	1.00
5.4	0.00	0.00	0.28	0.95	1.00	1.00
6.0	0.00	0.00	0.70	1.00	1.00	1.00
6.6	0.00	0.02	1.00	1.00	1.00	1.00

REGION 2

1	2	3	4	5	6
0.00	0.00	0.10	0.51	0.63	-1.00
0.00	0.00	0.16	0.90	+1.00	1.00
0.00	0.00	0.24	0.98	1.00	1.00
0.00	0.00	0.24	0.98	1.00	1.00
0.00	0.00	0.70	1.00	1.00	1.00
0.00	0.01	1.00	1.00	1.00	1.00

REGION 3

1	2	3	4	5	6
0.00	0.02	0.18	0.49	0.76	1.00
0.00	0.05	0.62	+1.00	>1.00	<1.00
0.00	0.23	0.72	1.00	1.00	1.00
0.00	-0.23	0.96	1.00	-1.00	1.00
0.00	0.44	0.98	1.00	1.00	1.00
0.00	0.59	1.00	1.00	1.00	1.00

REGION 4

MB*	1	2	3	4	5	6
3.6	0.00	0.05	0.32	+0.75	<0.75	<1.00
4.2	0.00	0.16	0.85	>1.00	1.00	1.00
4.8	0.00	0.43	>1.00	+1.00	-1.00	1.00
5.4	0.00	0.72	+1.00	-1.00	1.00	1.00
6.0	0.57	0.93	1.00	1.00	1.00	1.00
6.6	0.90	1.00	1.00	1.00	1.00	1.00

REGION 5

1	2	3	4	5	6
0.05	0.16	0.36	0.74	0.89	-1.00
0.05	0.36	0.71	>1.00	+1.00	<1.00
0.33	0.69	+0.97	<1.00	1.00	1.00
0.88	0.96	1.00	1.00	1.00	1.00
1.00	1.00	1.00	1.00	1.00	1.00
1.00	1.00	1.00	1.00	1.00	1.00

REGION 6

1	2	3	4	5	6
0.05	0.29	0.86	0.99	+1.00	<1.00
0.05	0.65	1.00	+1.00	+1.00	1.00
0.41	0.94	1.00	1.00	1.00	1.00
0.81	0.96	1.00	1.00	1.00	1.00
0.96	0.99	1.00	1.00	1.00	1.00
0.99	1.00	1.00	1.00	1.00	1.00

REGION 7

MB*	1	2	3	4	5	6
3.6	0.00	0.00	0.31	-0.53	0.95	1.00
4.2	0.00	0.19	+0.55	0.92	+1.00	1.00
4.8	0.14	0.47	0.93	0.98	1.00	1.00
5.4	0.88	0.96	0.99	1.00	1.00	1.00
6.0	0.99	1.00	1.00	1.00	1.00	1.00
6.6	1.00	1.00	1.00	1.00	1.00	1.00

REGION 8

1	2	3	4	5	6
0.00	0.00	0.00	0.04	0.49	1.00
0.00	0.00	0.00	0.04	-0.81	1.00
0.00	0.00	0.00	0.15	0.96	1.00
0.00	0.17	0.18	0.81	+1.00	1.00
0.00	0.17	0.76	0.98	1.00	1.00
0.00	0.17	0.88	1.00	1.00	1.00

REGION 9

1	2	3	4	5	6
0.00	0.02	0.08	0.26	0.56	1.00
0.00	0.02	0.12	0.77	>1.00	1.00
0.00	0.02	0.12	0.96	1.00	1.00
0.00	0.02	0.65	0.96	1.00	1.00
0.00	0.03	0.91	1.00	1.00	1.00
0.00	0.10	1.00	1.00	1.00	1.00

REGION 10

MB*	1	2	3	4	5	6
3.6	0.00	0.04	0.20	0.39	0.85	1.00
4.2	0.01	0.11	0.35	+1.00	>1.00	1.00
4.8	0.05	0.20	<0.67	1.00	1.00	1.00
5.4	0.21	0.81	0.95	1.00	1.00	1.00
6.0	-0.83	1.00	+1.00	1.00	1.00	1.00
6.6	1.00	1.00	1.00	1.00	1.00	1.00

REGION 11

1	2	3	4	5	6
0.00	0.00	0.00	0.00	0.76	1.00
0.00	0.00	0.00	0.15	1.00	1.00
0.00	0.00	0.00	0.49	+1.00	1.00
0.00	0.00	0.00	0.71	1.00	1.00
0.00	0.00	0.03	0.94	1.00	1.00
0.00	0.00	0.11	1.00	1.00	1.00

REGION 12

1	2	3	4	5	6
0.01	0.13	0.30	0.75	1.00	1.00
0.01	0.44	0.88	1.00	1.00	1.00
0.33	0.82	0.99	+1.00	1.00	1.00
0.76	0.94	0.99	1.00	1.00	1.00
0.95	0.99	1.00	1.00	1.00	1.00
0.98	1.00	1.00	1.00	1.00	1.00

REGION 13

MB*	1	2	3	4	5	6
3.6	0.00	0.00	0.24	0.71	0.88	1.00
4.2	0.00	0.00	0.24	0.77	0.95	1.00
4.8	0.00	0.00	0.30	0.92	0.99	1.00
5.4	0.00	0.03	0.69	0.99	1.00	1.00
6.0	0.11	0.54	0.98	1.00	1.00	1.00
6.6	0.51	0.90	1.00	1.00	1.00	1.00

time category	corresponding period
1	1625-1779 (155 years)
2	1780-1859 (80 years)
3	1860-1909 (50 years)
4	1910-1949 (40 years)
5	1950-1974 (25 years)
6	1975-1983 (8 years)

Table 2-II: Probabilities of detection and periods of completeness for the EPRI catalog

	magnitude interval						
	3.3 3.9	3.9 4.5	4.5 5.1	5.1 5.7	5.7 6.3	6.3 6.9	6.9 7.5
zone							
1	42.414	67.917	81.437	86.307	108.993	123.988	123.988
2	50.004	77.357	84.997	84.997	109.093	123.988	123.988
3	56.659	100.193	128.683	140.833	157.923	171.028	171.028
4	73.884	116.293	158.548	181.828	198.468	203.980	203.980
5	91.104	137.873	229.123	336.333	358.964	358.964	358.964
6	139.647	176.148	262.808	326.728	352.378	357.120	357.120
7	69.404	98.017	157.607	200.103	203.668	203.980	203.980
8	21.124	29.174	33.074	66.197	72.997	73.993	73.993
9	33.269	64.677	78.227	99.977	119.693	123.988	123.988
10	56.159	91.493	123.473	218.368	332.785	358.964	358.964
11	27.899	33.997	53.637	62.557	71.517	73.993	73.993
12	79.307	152.647	239.778	316.558	350.508	356.500	356.500
13	71.709	75.909	85.374	108.157	123.043	123.988	123.988

Equivalent periods of observation (years)
$T_{eq}(\underline{x},m)$

ICELL 222

ALPHA	ITIMES	IOUT	IFAIL	IFOUT	IFIN
0.100	15096	10201	4204	2950	1254

R-OUT	R-IN	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT
2.458	2.904	0.899	0.167	0.146	0.172	0.140	0.161

REGIONS : ITEST	IFAIL	ITEST-OUT	IFAIL-OUT	R	R2	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT	U
IGT, IFT 24531		4684										
R0 4.237190												
1	3363	804	7731	2182	2.543	3.183	0.933	0.224	0.133	0.167	0.129	0.156	
2	3363	804	4923	1211	3.065	3.183	1.120	0.038	0.133	0.138	0.129	0.133	9.012
3	3363	804	1348	380	2.547	3.183	0.935	0.223	0.133	0.166	0.129	0.156	19.709
7	3363	804	87	36	1.417	3.183	0.348	0.809	0.133	0.299	0.129	0.249	999.000
9	3363	804	494	235	1.102	3.183	0.097	1.061	0.133	0.384	0.129	0.299	130.457
10	3363	804	879	320	1.747	3.183	0.558	0.600	0.133	0.243	0.129	0.212	72.465
2	1353	379	7358	1976	2.724	2.570	1.002	-0.058	0.165	0.156	0.155	0.147	
1	1353	379	4923	1211	3.065	2.570	1.120	-0.176	0.165	0.138	0.155	0.133	9.012
3	1353	379	746	247	2.020	2.570	0.703	0.241	0.165	0.210	0.155	0.189	10.026
4	1353	379	22	4	4.500	2.570	1.504	-0.560	0.165	0.094	0.155	0.095	999.000
5	1353	379	95	6	14.833	2.570	2.697	-1.753	0.165	0.029	0.155	0.031	21.779
6	1353	379	73	14	4.214	2.570	1.438	-0.495	0.165	0.101	0.155	0.100	999.000
7	1353	379	73	39	0.872	2.570	-0.137	1.081	0.165	0.486	0.155	0.351	999.000
8	1353	379	22	0	999.000	2.570	6.907	-5.963	0.165	0.000	0.155	0.000	999.000
9	1353	379	387	211	0.834	2.570	-0.181	1.125	0.165	0.508	0.155	0.361	98.574
10	1353	379	490	189	1.593	2.570	0.465	0.479	0.165	0.266	0.155	0.228	29.574
11	1353	379	227	41	4.537	2.570	1.512	-0.568	0.165	0.093	0.155	0.094	12.420
12	1353	379	300	14	20.429	2.570	3.017	-2.073	0.165	0.021	0.155	0.023	78.972
3	99	39	2094	627	2.340	1.538	0.850	-0.419	0.275	0.181	0.234	0.168	
1	99	39	1348	380	2.547	1.538	0.935	-0.504	0.275	0.166	0.234	0.156	19.709
2	99	39	746	247	2.020	1.538	0.703	-0.272	0.275	0.210	0.234	0.189	10.026
4	0	0	22	4	4.500	999.000	1.504	999.000	0.000	0.424	0.000	0.095	
2	0	0	22	4	4.500	999.000	1.504	999.000	0.000	0.094	0.000	0.095	999.000
5	1	0	95	6	14.833	999.000	2.697	999.000	0.000	0.424	0.000	0.031	
2	1	0	95	6	14.833	999.000	2.697	999.000	0.000	0.029	0.000	0.031	21.779
6	0	0	73	14	4.214	999.000	1.438	999.000	0.000	0.424	0.000	0.100	
2	0	0	73	14	4.214	999.000	1.438	999.000	0.000	0.101	0.000	0.100	999.000
7	0	0	167	77	1.169	999.000	0.156	999.000	0.000	0.424	0.000	0.287	
1	0	0	87	36	1.417	999.000	0.348	999.000	0.000	0.299	0.000	0.249	999.000
2	0	0	73	39	0.872	999.000	-0.137	999.000	0.000	0.486	0.000	0.351	999.000
8	0	0	1	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	999.000
9	0	0	6	2	2.000	999.000	0.693	999.000	0.000	0.212	0.000	0.191	999.000
8	0	0	28	1	27.000	999.000	3.296	999.000	0.000	0.424	0.000	0.017	
2	0	0	22	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	999.000
7	0	0	1	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	999.000
9	0	0	5	1	4.000	999.000	1.386	999.000	0.000	0.106	0.000	0.105	999.000

Table 2-IV: Odds ratios for source configuration C

9	15	9	892	449	0.987	0.667	-0.013	-0.392	0.636	0.429	0.414	0.323	
1	15	9	494	235	1.102	0.667	0.097	-0.503	0.636	0.384	0.414	0.299	130.457
2	15	9	387	211	0.834	0.667	-0.181	-0.224	0.636	0.508	0.414	0.361	98.574
7	15	9	6	2	2.000	0.667	0.693	-1.099	0.636	0.212	0.414	0.191	999.000
8	15	9	5	1	4.000	0.667	1.386	-1.792	0.636	0.106	0.414	0.105	999.000
10	42	23	1369	509	1.690	0.826	0.524	-0.716	0.513	0.251	0.363	0.218	
1	42	23	879	320	1.747	0.826	0.558	-0.749	0.513	0.243	0.363	0.212	72.465
2	42	23	490	189	1.593	0.826	0.465	-0.656	0.513	0.266	0.363	0.228	29.574
11	7	0	250	41	5.098	999.000	1.629	999.000	0.000	0.424	0.000	0.085	
2	7	0	227	41	4.537	999.000	1.512	999.000	0.000	0.093	0.000	0.094	12.420
12	7	0	23	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	12.420
12	15	0	323	14	22.071	999.000	3.094	999.000	0.000	0.424	0.000	0.021	
2	15	0	300	14	20.429	999.000	3.017	999.000	0.000	0.021	0.000	0.023	78.972
11	15	0	23	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	78.972

ICELL 157

ALPHA	ITIMES	IOUT	IFAIL	IFOUT	IFIN
0.100	9073	4690	2710	1462	1248

R-OUT	R-IN	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT
2.208	2.512	0.792	0.129	0.115	0.130	0.113	0.127

REGIONS : ITEST	IFAIL	ITEST-OUT	IFAIL-OUT	R	R2	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT	U
IGT, IFT	12246	3157										
R0 2.878999												
1 4059	1140	4566	1422	2.211	2.561	0.793	0.147	0.112	0.130	0.111	0.126	
2 4059	1140	475	160	1.969	2.561	0.677	0.263	0.112	0.146	0.111	0.140	8.728
3 4059	1140	1230	350	2.514	2.561	0.922	0.018	0.112	0.115	0.111	0.113	1.173
5 4059	1140	285	69	3.130	2.561	1.141	-0.201	0.112	0.092	0.111	0.093	3.149
7 4059	1140	256	55	3.655	2.561	1.296	-0.356	0.112	0.079	0.111	0.080	6.392
8 4059	1140	322	46	6.000	2.561	1.792	-0.852	0.112	0.048	0.111	0.051	30.653
10 4059	1140	1998	742	1.693	2.561	0.526	0.414	0.112	0.170	0.111	0.159	55.563
2 10	5	475	160	1.969	1.000	0.677	-0.677	0.288	0.146	0.242	0.140	
1 10	5	475	160	1.969	1.000	0.677	-0.677	0.288	0.146	0.242	0.140	8.728
3 75	17	1230	350	2.514	3.412	0.922	0.305	0.084	0.115	0.086	0.113	
1 75	17	1230	350	2.514	3.412	0.922	0.305	0.084	0.115	0.086	0.113	1.173
4 0	0	12	5	1.400	999.000	0.336	999.000	0.000	0.288	0.000	0.186	
6 0	0	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.173
10 0	0	12	5	1.400	999.000	0.336	999.000	0.000	0.206	0.000	0.186	999.000
5 3	0	285	69	3.130	999.000	1.141	999.000	0.000	0.288	0.000	0.093	
1 3	0	285	69	3.130	999.000	1.141	999.000	0.000	0.092	0.000	0.093	3.149
6 0	0	12	5	1.400	999.000	0.336	999.000	0.000	0.288	0.000	0.186	
4 0	0	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	3.149
10 0	0	12	5	1.400	999.000	0.336	999.000	0.000	0.206	0.000	0.186	999.000
7 3	0	256	55	3.655	999.000	1.296	999.000	0.000	0.288	0.000	0.080	
1 3	0	256	55	3.655	999.000	1.296	999.000	0.000	0.079	0.000	0.080	6.392
8 5	0	322	46	6.000	999.000	1.792	999.000	0.000	0.288	0.000	0.051	
1 5	0	322	46	6.000	999.000	1.792	999.000	0.000	0.048	0.000	0.051	30.653
9 6	2	100	30	2.333	2.000	0.847	-0.154	0.144	0.123	0.138	0.121	
10 6	2	100	30	2.333	2.000	0.847	-0.154	0.144	0.123	0.138	0.121	1.864
10 222	84	2122	782	1.714	1.643	0.539	-0.042	0.175	0.168	0.163	0.157	
1 222	84	1998	742	1.693	1.643	0.526	-0.030	0.175	0.170	0.163	0.159	55.563
4 222	84	12	5	1.400	1.643	0.336	0.160	0.175	0.206	0.163	0.186	999.000
6 222	84	12	5	1.400	1.643	0.336	0.160	0.175	0.206	0.163	0.186	999.000
9 222	84	100	30	2.333	1.643	0.847	-0.351	0.175	0.123	0.163	0.121	1.864

Table 2-V: Odds ratios for source configuration D (unmerged sources)

ICELL 151

ALPHA	ITIMES	IOUT	IFAIL	IFOUT	IFIN
0.100	9276	3946	2915	1282	1633

R-OUT	R-IN	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT
2.078	2.264	0.731	0.086	0.116	0.126	0.114	0.123

REGIONS :	ITEST	IFAIL	ITEST-OUT	IFAIL-OUT	R	R2	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT	U
IGT, IFT		11325		3126									
R0 2.622841													
1	5041	1526	3918	1270	2.085	2.303	0.735	0.100	0.114	0.126	0.112	0.123	
2	5041	1526	535	188	1.846	2.303	0.613	0.222	0.114	0.142	0.112	0.136	7.119
3	5041	1526	321	82	2.915	2.303	1.070	-0.235	0.114	0.090	0.112	0.091	4.497
5	5041	1526	285	62	3.597	2.303	1.280	-0.446	0.114	0.073	0.112	0.075	10.629
6	5041	1526	356	58	5.138	2.303	1.637	-0.802	0.114	0.051	0.112	0.054	999.000
7	5041	1526	2421	880	1.751	2.303	0.560	0.274	0.114	0.150	0.112	0.143	30.715
2	10	5	535	188	1.846	1.000	0.613	-0.613	0.262	0.142	0.226	0.136	
1	10	5	535	188	1.846	1.000	0.613	-0.613	0.262	0.142	0.226	0.136	7.119
3	3	0	321	82	2.915	999.000	1.070	999.000	0.000	0.262	0.000	0.091	
1	3	0	321	82	2.915	999.000	1.070	999.000	0.000	0.090	0.000	0.091	4.497
4	0	0	28	12	1.333	999.000	0.288	999.000	0.000	0.262	0.000	0.179	
7	0	0	28	12	1.333	999.000	0.288	999.000	0.000	0.197	0.000	0.179	999.000
5	3	0	285	62	3.597	999.000	1.280	999.000	0.000	0.262	0.000	0.075	
1	3	0	285	62	3.597	999.000	1.280	999.000	0.000	0.073	0.000	0.075	10.629
6	0	0	356	58	5.138	999.000	1.637	999.000	0.000	0.262	0.000	0.054	
1	0	0	356	58	5.138	999.000	1.637	999.000	0.000	0.051	0.000	0.054	999.000
7	273	102	2449	892	1.746	1.676	0.557	-0.040	0.156	0.150	0.148	0.143	
1	273	102	2421	880	1.751	1.676	0.560	-0.044	0.156	0.150	0.148	0.143	30.715
4	273	102	28	12	1.333	1.676	0.288	0.229	0.156	0.197	0.148	0.179	999.000

Source configuration D (similar sources merged)

ICELL 219

ALPHA	ITIMES	IOUT	IFAIL	IFOUT	IFIN
0.100	8057	5817	2170	1635	535

R-OUT	R-IN	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT
2.558	3.187	0.939	0.220	0.117	0.146	0.115	0.140

REGIONS : ITEST	IFAIL	ITEST-OUT	IFAIL-OUT	R	R2	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT	U
IGT, IFT	23871	5038										
R0 3.738190												
1	33	0	334	122	1.738	999.000	0.553	999.000	0.000	0.374	0.000	0.193	
5	33	0	210	93	1.258	999.000	0.230	999.000	0.000	0.297	0.000	0.248	29.815
2	33	0	124	29	3.276	999.000	1.187	999.000	0.000	0.114	0.000	0.113	12.811
2	63	20	410	117	2.504	2.150	0.918	-0.153	0.174	0.149	0.162	0.142	
1	63	20	124	29	3.276	2.150	1.187	-0.421	0.174	0.114	0.162	0.113	12.811
5	63	20	190	70	1.714	2.150	0.539	0.226	0.174	0.218	0.162	0.195	3.661
3	63	20	96	18	4.333	2.150	1.466	-0.701	0.174	0.086	0.162	0.087	12.974
3	30	0	372	83	3.482	999.000	1.248	999.000	0.000	0.374	0.000	0.107	
2	30	0	96	18	4.333	999.000	1.466	999.000	0.000	0.086	0.000	0.087	12.974
5	30	0	150	57	1.632	999.000	0.490	999.000	0.000	0.229	0.000	0.203	18.282
4	30	0	126	8	14.750	999.000	2.691	999.000	0.000	0.025	0.000	0.027	1.969
4	119	7	931	285	2.267	16.000	0.818	1.954	0.023	0.165	0.025	0.155	
3	119	7	126	8	14.750	16.000	2.691	0.081	0.023	0.025	0.025	0.027	1.969
5	119	7	273	141	0.936	16.000	-0.066	2.839	0.023	0.399	0.025	0.307	81.404
7	119	7	532	136	2.912	16.000	1.069	1.704	0.023	0.128	0.025	0.125	31.905
5	121	32	1110	431	1.575	2.781	0.455	0.568	0.134	0.237	0.130	0.209	
1	121	32	210	93	1.258	2.781	0.230	0.793	0.134	0.297	0.130	0.248	29.815
2	121	32	190	70	1.714	2.781	0.539	0.484	0.134	0.218	0.130	0.195	3.661
3	121	32	150	57	1.632	2.781	0.490	0.533	0.134	0.229	0.130	0.203	18.282
4	121	32	273	141	0.936	2.781	-0.066	1.089	0.134	0.399	0.130	0.307	81.404
6	121	32	202	45	3.489	2.781	1.250	-0.227	0.134	0.107	0.130	0.106	3.059
7	121	32	0	0	999.000	2.781	6.907	-5.884	0.134	0.000	0.130	0.000	999.000
12	121	32	85	25	2.400	2.781	0.875	0.147	0.134	0.156	0.130	0.148	0.932
6	76	12	950	220	3.318	5.333	1.199	0.475	0.070	0.113	0.072	0.111	
5	76	12	202	45	3.489	5.333	1.250	0.424	0.070	0.107	0.072	0.106	3.059
10	76	12	614	131	3.687	5.333	1.305	0.369	0.070	0.101	0.072	0.101	20.297
11	76	12	134	44	2.045	5.333	0.716	0.958	0.070	0.183	0.072	0.169	7.206
7	570	90	2488	562	3.427	5.333	1.232	0.442	0.070	0.109	0.072	0.108	
4	570	90	532	136	2.912	5.333	1.069	0.605	0.070	0.128	0.072	0.125	31.905
8	570	90	46	11	3.182	5.333	1.157	0.517	0.070	0.117	0.072	0.115	999.000
9	570	90	107	29	2.690	5.333	0.989	0.685	0.070	0.139	0.072	0.134	12.803
10	570	90	1803	386	3.671	5.333	1.300	0.374	0.070	0.102	0.072	0.102	50.945
8	0	0	98	34	1.882	999.000	0.633	999.000	0.000	0.374	0.000	0.181	
7	0	0	46	11	3.182	999.000	1.157	999.000	0.000	0.117	0.000	0.115	999.000
10	0	0	52	23	1.261	999.000	0.232	999.000	0.000	0.296	0.000	0.248	999.000
9	3	2	107	29	2.690	0.500	0.989	-1.683	0.748	0.139	0.454	0.134	
7	3	2	107	29	2.690	0.500	0.989	-1.683	0.748	0.139	0.454	0.134	12.803

Table 2-VI: Odds ratios for source configuration E (unmerged sources)

10	1155	347	3542	929	2.813	2.329	1.034	-0.189	0.161	0.133	0.151	0.129	
7	1155	347	1803	386	3.671	2.329	1.300	-0.455	0.161	0.102	0.151	0.102	50.945
8	1155	347	52	23	1.261	2.329	0.232	0.613	0.161	0.296	0.151	0.248	999.000
6	1155	347	614	131	3.687	2.329	1.305	-0.460	0.161	0.101	0.151	0.101	20.297
11	1155	347	501	196	1.556	2.329	0.442	0.403	0.161	0.240	0.151	0.211	13.849
12	1155	347	260	72	2.611	2.329	0.960	-0.115	0.161	0.143	0.151	0.137	1.075
13	1155	347	312	121	1.579	2.329	0.456	0.389	0.161	0.237	0.151	0.208	13.965
11	45	12	635	240	1.646	2.750	0.498	0.513	0.136	0.227	0.131	0.202	
6	45	12	134	44	2.045	2.750	0.716	0.296	0.136	0.183	0.131	0.169	7.206
10	45	12	501	196	1.556	2.750	0.442	0.569	0.136	0.240	0.131	0.211	13.849
12	10	4	345	97	2.557	1.500	0.939	-0.533	0.249	0.146	0.217	0.140	
5	10	4	85	25	2.400	1.500	0.875	-0.470	0.249	0.156	0.217	0.148	0.932
10	10	4	260	72	2.611	1.500	0.960	-0.554	0.249	0.143	0.217	0.137	1.075
13	15	9	312	121	1.579	0.667	0.456	-0.862	0.561	0.237	0.384	0.208	
10	15	9	312	121	1.579	0.667	0.456	-0.862	0.561	0.237	0.384	0.208	13.965

ICELL 213

ALPHA	ITIMES	IOUT	IFAIL	IFOUT	IFIN
0.100	10614	7976	2753	2139	614

R-OUT	R-IN	R0	R1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT
2.729	3.296	1.004	0.189	0.113	0.136	0.111	0.131

IGT, IFT 22578 4790

R0 3.713570

REGIONS	ITEST	IFAIL	ITEST-OUT	IFAIL-OUT	R	R2	R0	R1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT	U
1	33	0	436	138	2.159	999.000	0.770	999.000	0.000	0.371	0.000	0.160	
2	33	0	124	29	3.276	999.000	1.187	999.000	0.000	0.113	0.000	0.112	12.811
4	33	0	312	109	1.862	999.000	0.622	999.000	0.000	0.199	0.000	0.181	23.683
2	63	20	692	187	2.701	2.150	0.993	-0.228	0.173	0.138	0.161	0.133	
1	63	20	124	29	3.276	2.150	1.187	-0.421	0.173	0.113	0.161	0.112	12.811
3	63	20	272	65	3.185	2.150	1.158	-0.393	0.173	0.117	0.161	0.115	41.763
4	63	20	296	93	2.183	2.150	0.781	-0.015	0.173	0.170	0.161	0.159	5.441
3	265	16	1694	509	2.328	15.563	0.845	1.900	0.024	0.160	0.026	0.151	
2	265	16	272	65	3.185	15.563	1.158	1.587	0.024	0.117	0.026	0.115	41.763
4	265	16	632	265	1.385	15.563	0.326	2.419	0.024	0.268	0.026	0.230	123.191
5	265	16	790	179	3.413	15.563	1.228	1.517	0.024	0.109	0.026	0.108	39.759
4	312	73	3748	1069	2.506	3.274	0.919	0.267	0.113	0.148	0.112	0.141	
1	312	73	312	109	1.862	3.274	0.622	0.564	0.113	0.199	0.112	0.181	23.683
2	312	73	296	93	2.183	3.274	0.781	0.405	0.113	0.170	0.112	0.159	5.441
3	312	73	632	265	1.385	3.274	0.326	0.860	0.113	0.268	0.112	0.230	123.191
5	312	73	916	166	4.518	3.274	1.508	-0.322	0.113	0.082	0.112	0.084	7.845
8	312	73	1324	347	2.816	3.274	1.035	0.151	0.113	0.132	0.112	0.128	6.219
9	312	73	268	89	2.011	3.274	0.699	0.487	0.113	0.185	0.112	0.170	6.941
5	570	90	3831	782	3.899	5.333	1.361	0.313	0.070	0.095	0.072	0.096	
3	570	90	790	179	3.413	5.333	1.228	0.446	0.070	0.109	0.072	0.108	39.759
4	570	90	916	166	4.518	5.333	1.508	0.166	0.070	0.082	0.072	0.084	7.845
6	570	90	46	11	3.182	5.333	1.157	0.517	0.070	0.117	0.072	0.115	999.000
7	570	90	107	29	2.690	5.333	0.989	0.685	0.070	0.138	0.072	0.133	12.803
8	570	90	1972	397	3.967	5.333	1.378	0.296	0.070	0.094	0.072	0.094	57.319
6	0	0	101	34	1.971	999.000	0.678	999.000	0.000	0.371	0.000	0.173	
5	0	0	46	11	3.182	999.000	1.157	999.000	0.000	0.117	0.000	0.115	999.000
8	0	0	55	23	1.391	999.000	0.330	999.000	0.000	0.267	0.000	0.229	999.000
7	3	2	107	29	2.690	0.500	0.989	-1.683	0.743	0.138	0.452	0.133	
5	3	2	107	29	2.690	0.500	0.989	-1.683	0.743	0.138	0.452	0.133	12.803
8	1332	392	4213	1104	2.816	2.398	1.035	-0.161	0.155	0.132	0.147	0.128	
9	1332	392	532	206	1.583	2.398	0.459	0.416	0.155	0.235	0.147	0.207	15.690
4	1332	392	1324	347	2.816	2.398	1.035	-0.161	0.155	0.132	0.147	0.128	6.219
5	1332	392	1972	397	3.967	2.398	1.378	-0.503	0.155	0.094	0.147	0.094	57.319
6	1332	392	55	23	1.391	2.398	0.330	0.544	0.155	0.267	0.147	0.229	999.000
10	1332	392	330	131	1.519	2.398	0.418	0.457	0.155	0.244	0.147	0.214	18.460
9	45	12	800	295	1.712	2.750	0.538	0.474	0.135	0.217	0.130	0.194	
4	45	12	268	89	2.011	2.750	0.699	0.313	0.135	0.185	0.130	0.170	6.941
8	45	12	532	206	1.583	2.750	0.459	0.553	0.135	0.235	0.130	0.207	15.690
10	15	9	330	131	1.519	0.667	0.418	-0.824	0.557	0.244	0.382	0.214	
8	15	9	330	131	1.519	0.667	0.418	-0.824	0.557	0.244	0.382	0.214	18.460

Source configuration E (similar sources merged)

ICELL 291

ALPHA	ITIMES	IOUT	IFAIL	IFOUT	IFIN
0.100	21993	16809	8281	7569	712

R-OUT	R-IN	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT
1.221	6.281	0.199	1.638	0.039	0.202	0.042	0.183

REGIONS : ITEST	IFAIL	ITEST-OUT	IFAIL-OUT	R	R2	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT	U
IGT, IFT	42195	12173										
R0 2.466278												
1	3	0	51	28	0.821	999.000	-0.197	999.000	0.000	0.247	0.000	0.250	
2	3	0	51	28	0.821	999.000	-0.197	999.000	0.000	0.300	0.000	0.250	8.197
2	100	34	1582	513	2.084	1.941	0.734	-0.071	0.127	0.118	0.124	0.116	
22	100	34	1438	452	2.181	1.941	0.780	-0.117	0.127	0.113	0.124	0.112	320.046
1	100	34	51	28	0.821	1.941	-0.197	0.860	0.127	0.300	0.124	0.250	8.197
4	100	34	68	27	1.519	1.941	0.418	0.246	0.127	0.162	0.124	0.153	0.593
5	100	34	25	6	3.167	1.941	1.153	-0.489	0.127	0.078	0.124	0.080	1.392
3	9	0	507	139	2.647	999.000	0.974	999.000	0.000	0.247	0.000	0.094	
22	9	0	487	129	2.775	999.000	1.021	999.000	0.000	0.089	0.000	0.090	83.907
4	9	0	20	10	1.000	999.000	0.000	999.000	0.000	0.247	0.000	0.215	6.890
4	6	2	644	372	0.731	2.000	-0.313	1.006	0.123	0.337	0.121	0.273	
22	6	2	548	331	0.656	2.000	-0.422	1.115	0.123	0.376	0.121	0.295	813.001
3	6	2	20	10	1.000	2.000	0.000	0.693	0.123	0.247	0.121	0.215	6.890
2	6	2	68	27	1.519	2.000	0.418	0.275	0.123	0.162	0.121	0.153	0.593
5	6	2	8	4	1.000	2.000	0.000	0.693	0.123	0.247	0.121	0.215	1.111
5	1	0	210	76	1.763	999.000	0.567	999.000	0.000	0.247	0.000	0.135	
4	1	0	8	4	1.000	999.000	0.000	999.000	0.000	0.247	0.000	0.215	1.111
2	1	0	25	6	3.167	999.000	1.153	999.000	0.000	0.078	0.000	0.080	1.392
6	1	0	2	2	0.000	999.000	0.000	999.000	0.000	0.247	0.000	0.215	999.000
22	1	0	175	64	1.734	999.000	0.551	999.000	0.000	0.142	0.000	0.136	93.112
6	0	0	139	125	0.112	999.000	-2.189	999.000	0.000	0.247	0.000	0.710	
5	0	0	2	2	0.000	999.000	0.000	999.000	0.000	0.247	0.000	0.215	999.000
22	0	0	137	123	0.114	999.000	-2.173	999.000	0.000	2.167	0.000	0.707	999.000
7	1	1	294	274	0.073	0.000	-2.617	*******	1.000	3.379	1.000	0.790	
22	1	1	274	257	0.066	0.000	-2.716	*******	1.000	3.728	1.000	0.806	*******
8	1	1	12	12	0.000	0.000	*******	*******	1.000	0.247	1.000	0.000	18.000
9	1	1	8	5	0.600	0.000	-0.511	*******	1.000	0.411	1.000	0.314	2.109
8	5	0	339	49	5.918	999.000	1.778	999.000	0.000	0.247	0.000	0.044	
22	5	0	327	37	7.838	999.000	2.059	999.000	0.000	0.031	0.000	0.034	0.737
7	5	0	12	12	0.000	999.000	0.000	999.000	0.000	0.247	0.000	0.215	18.000
9	6	2	457	254	0.799	2.000	-0.224	0.917	0.123	0.309	0.121	0.255	
22	6	2	449	249	0.803	2.000	-0.219	0.912	0.123	0.307	0.121	0.254	585.387
7	6	2	8	5	0.600	2.000	-0.511	1.204	0.123	0.411	0.121	0.314	2.109
10	88	29	1860	810	1.296	2.034	0.260	0.451	0.121	0.190	0.119	0.175	
22	88	29	1621	710	1.283	2.034	0.249	0.461	0.121	0.192	0.119	0.176	755.438
12	88	29	70	34	1.059	2.034	0.057	0.653	0.121	0.233	0.119	0.206	5.503
11	88	29	66	33	1.000	2.034	0.000	0.710	0.121	0.247	0.119	0.215	999.000
13	88	29	103	33	2.121	2.034	0.752	-0.042	0.121	0.116	0.119	0.114	11.402
11	0	0	294	65	3.523	999.000	1.259	999.000	0.000	0.247	0.000	0.072	

Table 2-VII: Odds ratios for Woodward-Clyde's source configuration (unmerged sources)

11	0	0	294	65	3.523	999.000	1.259	999.000	0.000	0.247	0.000	0.072	
22	0	0	228	32	6.125	999.000	1.812	999.000	0.000	0.040	0.000	0.043	999.000
10	0	0	66	33	1.000	999.000	0.000	999.000	0.000	0.247	0.000	0.215	999.000
12	10	6	925	434	1.131	0.667	0.123	-0.529	0.370	0.218	0.291	0.195	
10	10	6	70	34	1.059	0.667	0.057	-0.463	0.370	0.233	0.291	0.206	5.503
13	10	6	40	18	1.222	0.667	0.201	-0.606	0.370	0.202	0.291	0.183	18.474
18	10	6	130	47	1.766	0.667	0.569	-0.974	0.370	0.140	0.291	0.134	62.018
22	10	6	685	335	1.045	0.667	0.044	-0.449	0.370	0.236	0.291	0.208	597.634
13	25	0	1295	260	3.981	999.000	1.381	999.000	0.000	0.247	0.000	0.064	
10	25	0	103	33	2.121	999.000	0.752	999.000	0.000	0.116	0.000	0.114	11.402
22	25	0	799	98	7.153	999.000	1.968	999.000	0.000	0.034	0.000	0.037	3.505
14	25	0	32	2	15.000	999.000	2.708	999.000	0.000	0.016	0.000	0.018	2.001
15	25	0	40	32	0.250	999.000	-1.386	999.000	0.000	0.987	0.000	0.523	40.946
16	25	0	106	55	0.927	999.000	-0.076	999.000	0.000	0.266	0.000	0.228	23.737
18	25	0	175	22	6.955	999.000	1.939	999.000	0.000	0.035	0.000	0.038	6.071
12	25	0	40	18	1.222	999.000	0.201	999.000	0.000	0.202	0.000	0.183	18.474
14	6	0	600	148	3.054	999.000	1.116	999.000	0.000	0.247	0.000	0.082	
13	6	0	32	2	15.000	999.000	2.708	999.000	0.000	0.016	0.000	0.018	2.001
15	6	0	20	12	0.667	999.000	-0.405	999.000	0.000	0.370	0.000	0.291	8.315
22	6	0	548	134	3.090	999.000	1.128	999.000	0.000	0.080	0.000	0.081	69.209
15	10	7	875	596	0.468	0.429	-0.759	-0.088	0.575	0.527	0.390	0.369	
22	10	7	685	489	0.401	0.429	-0.914	0.067	0.575	0.615	0.390	0.406	*******
14	10	7	20	12	0.667	0.429	-0.405	-0.442	0.575	0.370	0.390	0.291	8.315
16	10	7	70	36	0.944	0.429	-0.057	-0.790	0.575	0.261	0.390	0.225	1.328
13	10	7	40	32	0.250	0.429	-1.386	0.539	0.575	0.987	0.390	0.523	40.946
21	10	7	60	27	1.222	0.429	0.201	-1.048	0.575	0.202	0.390	0.183	2.972
16	90	46	1978	1151	0.719	0.957	-0.331	0.286	0.258	0.343	0.223	0.276	
15	90	46	70	36	0.944	0.957	-0.057	0.013	0.258	0.261	0.223	0.225	1.328
17	90	46	82	46	0.783	0.957	-0.245	0.201	0.258	0.315	0.223	0.259	1.397
22	90	46	1720	1014	0.696	0.957	-0.362	0.318	0.258	0.354	0.223	0.282	*******
13	90	46	106	55	0.927	0.957	-0.076	0.031	0.258	0.266	0.223	0.228	23.737
17	15	10	805	662	0.216	0.500	-1.532	0.839	0.493	1.142	0.354	0.559	
16	15	10	82	46	0.783	0.500	-0.245	-0.448	0.493	0.315	0.354	0.259	1.397
22	15	10	723	616	0.174	0.500	-1.750	1.057	0.493	1.420	0.354	0.612	*******
18	270	20	2950	696	3.239	12.500	1.175	1.351	0.020	0.076	0.021	0.078	
22	270	20	2473	609	3.061	12.500	1.119	1.407	0.020	0.081	0.021	0.082	209.858
19	270	20	56	10	4.600	12.500	1.526	1.000	0.020	0.054	0.021	0.056	9.784
20	270	20	116	8	13.500	12.500	2.603	-0.077	0.020	0.018	0.021	0.020	0.501
13	270	20	175	22	6.955	12.500	1.939	0.586	0.020	0.035	0.021	0.038	6.071
12	270	20	130	47	1.766	12.500	0.569	1.957	0.020	0.140	0.021	0.134	62.018

19	2	1	278	131	1.122	1.000	0.115	-0.115	0.247	0.220	0.215	0.196	
18	2	1	56	10	4.600	1.000	1.526	-1.526	0.247	0.054	0.215	0.056	9.784
20	2	1	9	3	2.000	1.000	0.693	-0.693	0.247	0.123	0.215	0.121	3.106
22	2	1	213	118	0.805	1.000	-0.217	0.217	0.247	0.306	0.215	0.254	326.962
20	6	0	490	35	13.000	999.000	2.565	999.000	0.000	0.247	0.000	0.021	
18	6	0	116	8	13.500	999.000	2.603	999.000	0.000	0.018	0.000	0.020	0.501
19	6	0	9	3	2.000	999.000	0.693	999.000	0.000	0.123	0.000	0.121	3.106
22	6	0	365	24	14.208	999.000	2.654	999.000	0.000	0.017	0.000	0.019	9.906
21	66	27	1605	1263	0.271	1.444	-1.306	1.674	0.171	0.911	0.159	0.503	
15	66	27	60	27	1.222	1.444	0.201	0.167	0.171	0.202	0.159	0.183	2.972
22	66	27	1545	1236	0.250	1.444	-1.386	1.754	0.171	0.987	0.159	0.523	*******
22	4465	527	15440	7057	1.188	7.472	0.172	1.839	0.033	0.208	0.035	0.187	
2	4465	527	1438	452	2.181	7.472	0.780	1.231	0.033	0.113	0.035	0.112	320.046
3	4465	527	487	129	2.775	7.472	1.021	0.991	0.033	0.089	0.035	0.090	83.907
4	4465	527	548	331	0.656	7.472	-0.422	2.433	0.033	0.376	0.035	0.295	813.001
5	4465	527	175	64	1.734	7.472	0.551	1.461	0.033	0.142	0.035	0.136	93.112
6	4465	527	137	123	0.114	7.472	-2.173	4.184	0.033	2.167	0.035	0.707	999.000
7	4465	527	274	257	0.066	7.472	-2.716	4.727	0.033	3.728	0.035	0.806	*******
8	4465	527	327	37	7.838	7.472	2.059	-0.048	0.033	0.031	0.035	0.034	0.737
9	4465	527	449	249	0.803	7.472	-0.219	2.230	0.033	0.307	0.035	0.254	585.387
10	4465	527	1621	710	1.283	7.472	0.249	1.762	0.033	0.192	0.035	0.176	755.438
11	4465	527	228	32	6.125	7.472	1.812	0.199	0.033	0.040	0.035	0.043	999.000
12	4465	527	685	335	1.045	7.472	0.044	1.967	0.033	0.236	0.035	0.208	597.634
13	4465	527	799	98	7.153	7.472	1.968	0.044	0.033	0.034	0.035	0.037	3.505
14	4465	527	548	134	3.090	7.472	1.128	0.883	0.033	0.080	0.035	0.081	69.209
15	4465	527	685	489	0.401	7.472	-0.914	2.925	0.033	0.615	0.035	0.406	*******
16	4465	527	1720	1014	0.696	7.472	-0.362	2.373	0.033	0.354	0.035	0.282	*******
17	4465	527	723	616	0.174	7.472	-1.750	3.762	0.033	1.420	0.035	0.612	*******
18	4465	527	2473	609	3.061	7.472	1.119	0.893	0.033	0.081	0.035	0.082	209.858
19	4465	527	213	118	0.805	7.472	-0.217	2.228	0.033	0.306	0.035	0.254	326.962
20	4465	527	365	24	14.208	7.472	2.654	-0.643	0.033	0.017	0.035	0.019	9.906
21	4465	527	1545	1236	0.250	7.472	-1.386	3.398	0.033	0.987	0.035	0.523	*******

ICELL 287

ALPHA	ITIMES	IOUT	IFAIL	IFOUT	IFIN							
0.100	22609	16828	8594	7592	1002							
R-OUT	R-IN	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT					
1.217	4.769	0.196	1.366	0.049	0.193	0.052	0.177					
REGIONS : ITEST	IFAIL	ITEST-OUT	IFAIL-OUT	R	R2	B0	B1	ALPHA1-IN	ALPHA1-OUT	ALPHA2-IN	ALPHA2-OUT	U
IGT, IFT	41041	12238										
R0 2.353571												
1	3	0	72	34	1.118	999.000	0.111	999.000	0.000	0.235	0.000	0.190	
2	3	0	72	34	1.118	999.000	0.111	999.000	0.000	0.211	0.000	0.190	8.939
2	231	69	2546	993	1.564	2.348	0.447	0.406	0.100	0.150	0.100	0.143	
1	231	69	72	34	1.118	2.348	0.111	0.742	0.100	0.211	0.100	0.190	8.939
3	231	69	100	18	4.556	2.348	1.516	-0.663	0.100	0.052	0.100	0.054	8.343
4	231	69	24	16	0.500	2.348	-0.693	1.547	0.100	0.471	0.100	0.343	999.000
15	231	69	2350	925	1.541	2.348	0.432	0.421	0.100	0.153	0.100	0.145	535.886
3	9	0	598	157	2.809	999.000	1.033	999.000	0.000	0.235	0.000	0.085	
2	9	0	100	18	4.556	999.000	1.516	999.000	0.000	0.052	0.000	0.054	8.343
15	9	0	498	139	2.583	999.000	0.949	999.000	0.000	0.091	0.000	0.092	59.354
4	0	0	164	140	0.171	999.000	-1.764	999.000	0.000	0.235	0.000	0.604	
2	0	0	24	16	0.500	999.000	-0.693	999.000	0.000	0.471	0.000	0.343	999.000
15	0	0	140	124	0.129	999.000	-2.048	999.000	0.000	1.824	0.000	0.670	999.000
5	1	1	280	257	0.089	0.000	-2.414	*******	1.000	2.630	1.000	0.745	
15	1	1	280	257	0.089	0.000	-2.414	*******	1.000	2.630	1.000	0.745	*******
6	6	2	459	255	0.800	2.000	-0.223	0.916	0.118	0.294	0.116	0.246	
15	6	2	459	255	0.800	2.000	-0.223	0.916	0.118	0.294	0.116	0.246	462.749
7	88	29	2341	910	1.573	2.034	0.453	0.258	0.116	0.150	0.114	0.143	
8	88	29	66	33	1.000	2.034	0.000	0.710	0.116	0.235	0.114	0.207	999.000
9	88	29	70	34	1.059	2.034	0.057	0.653	0.116	0.222	0.114	0.198	5.503
10	88	29	148	40	2.700	2.034	0.993	-0.283	0.116	0.087	0.114	0.088	21.787
15	88	29	1657	729	1.273	2.034	0.241	0.469	0.116	0.185	0.114	0.170	602.954
12	88	29	400	74	4.405	2.034	1.483	-0.773	0.116	0.053	0.114	0.056	28.790
8	0	0	300	67	3.478	999.000	1.246	999.000	0.000	0.235	0.000	0.070	
7	0	0	66	33	1.000	999.000	0.000	999.000	0.000	0.235	0.000	0.207	999.000
15	0	0	234	34	5.882	999.000	1.772	999.000	0.000	0.040	0.000	0.043	999.000
9	10	6	285	117	1.436	0.667	0.362	-0.767	0.353	0.164	0.282	0.154	
10	10	6	55	24	1.292	0.667	0.256	-0.661	0.353	0.182	0.282	0.168	34.526
7	10	6	70	34	1.059	0.667	0.057	-0.463	0.353	0.222	0.282	0.198	5.503
12	10	6	160	59	1.712	0.667	0.538	-0.943	0.353	0.137	0.282	0.133	62.207
10	54	0	2057	473	3.349	999.000	1.209	999.000	0.000	0.235	0.000	0.072	
9	54	0	55	24	1.292	999.000	0.256	999.000	0.000	0.182	0.000	0.168	34.526
7	54	0	148	40	2.700	999.000	0.993	999.000	0.000	0.087	0.000	0.088	21.787
15	54	0	1338	241	4.552	999.000	1.516	999.000	0.000	0.052	0.000	0.054	18.217
14	54	0	196	139	0.410	999.000	-0.891	999.000	0.000	0.574	0.000	0.389	92.374
11	54	0	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	92.374
12	54	0	320	29	10.034	999.000	2.306	999.000	0.000	0.023	0.000	0.025	6.162
11	168	99	2696	1744	0.546	0.697	-0.605	0.244	0.338	0.431	0.273	0.324	
14	168	99	339	167	1.030	0.697	0.030	-0.391	0.338	0.229	0.273	0.202	9.516

Source configuration - Woodward Clyde Co. (similar sources merged)

15	168	99	2357	1577	0.495	0.697	-0.704	0.343	0.338	0.476	0.273	0.346	*******
12	376	39	3808	901	3.226	8.641	1.171	0.985	0.027	0.073	0.029	0.075	
13	376	39	64	16	3.000	8.641	1.099	1.058	0.027	0.078	0.029	0.080	13.107
9	376	39	160	59	1.712	8.641	0.538	1.619	0.027	0.137	0.029	0.133	62.207
7	376	39	400	74	4.405	8.641	1.483	0.674	0.027	0.053	0.029	0.056	28.790
10	376	39	320	29	10.034	8.641	2.306	-0.150	0.027	0.023	0.029	0.025	6.162
15	376	39	2864	723	2.961	8.641	1.086	1.071	0.027	0.079	0.029	0.081	146.447
13	2	1	282	134	1.104	1.000	0.099	-0.099	0.235	0.213	0.207	0.191	
12	2	1	64	16	3.000	1.000	1.099	-1.099	0.235	0.078	0.207	0.080	13.107
15	2	1	218	118	0.847	1.000	-0.166	0.166	0.235	0.278	0.207	0.236	235.049
14	153	64	2954	2093	0.411	1.391	-0.888	1.218	0.169	0.572	0.158	0.389	
15	153	64	2419	1787	0.354	1.391	-1.039	1.369	0.169	0.665	0.158	0.425	*******
10	153	64	196	139	0.410	1.391	-0.891	1.221	0.169	0.574	0.158	0.389	92.374
11	153	64	339	167	1.030	1.391	0.030	0.300	0.169	0.229	0.158	0.202	9.516
15	4680	692	14814	6909	1.144	5.763	0.135	1.617	0.041	0.206	0.043	0.186	
2	4680	692	2350	925	1.541	5.763	0.432	1.319	0.041	0.153	0.043	0.145	535.886
3	4680	692	498	139	2.583	5.763	0.949	0.803	0.041	0.091	0.043	0.092	59.354
4	4680	692	140	124	0.129	5.763	-2.048	3.799	0.041	1.824	0.043	0.670	999.000
5	4680	692	280	257	0.089	5.763	-2.414	4.165	0.041	2.630	0.043	0.745	*******
6	4680	692	459	255	0.800	5.763	-0.223	1.975	0.041	0.294	0.043	0.246	462.749
7	4680	692	1657	729	1.273	5.763	0.241	1.510	0.041	0.185	0.043	0.170	602.954
8	4680	692	234	34	5.882	5.763	1.772	-0.020	0.041	0.040	0.043	0.043	999.000
10	4680	692	1338	241	4.552	5.763	1.516	0.236	0.041	0.052	0.043	0.054	18.217
11	4680	692	2357	1577	0.495	5.763	-0.704	2.455	0.041	0.476	0.043	0.346	*******
12	4680	692	2864	723	2.961	5.763	1.086	0.666	0.041	0.079	0.043	0.081	146.447
13	4680	692	218	118	0.847	5.763	-0.166	1.917	0.041	0.278	0.043	0.236	235.049
14	4680	692	2419	1787	0.354	5.763	-1.039	2.791	0.041	0.665	0.043	0.425	*******

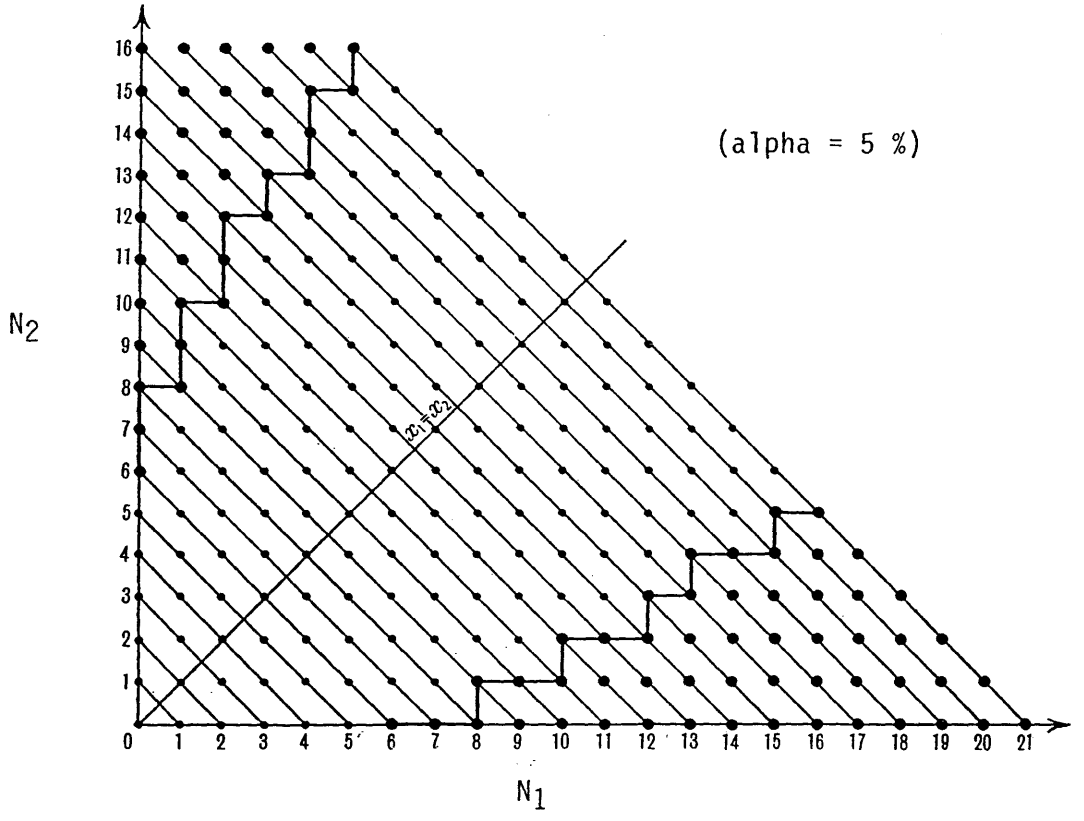


Figure 2-1: Regions of rejection (•) / acceptance (·) for the binomial test of equality of the recurrence rate

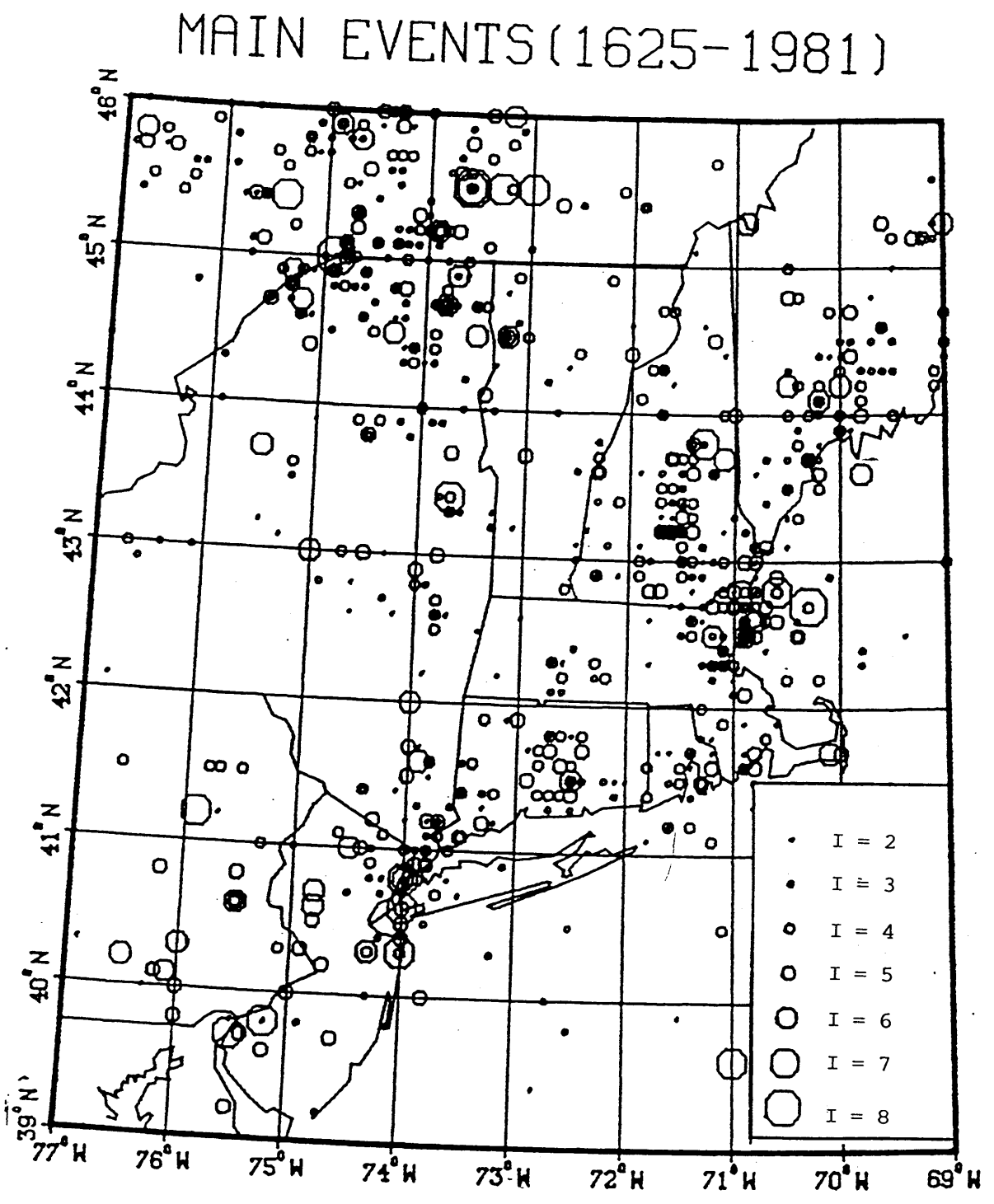


Figure 2-2: Main events for the Chiburis catalog (1625-1981)

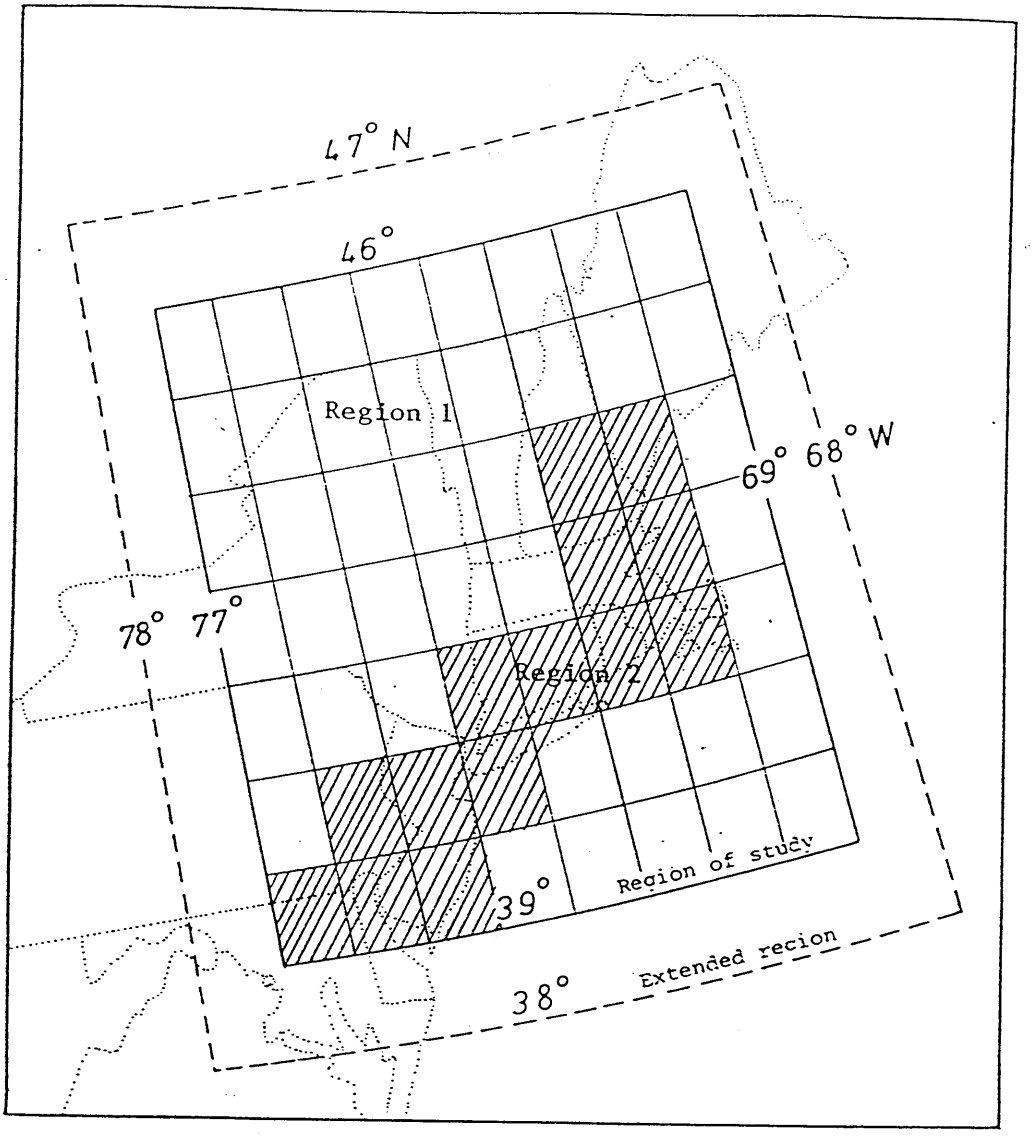


Figure 2-3: Incompleteness regions for the Chiburis catalog

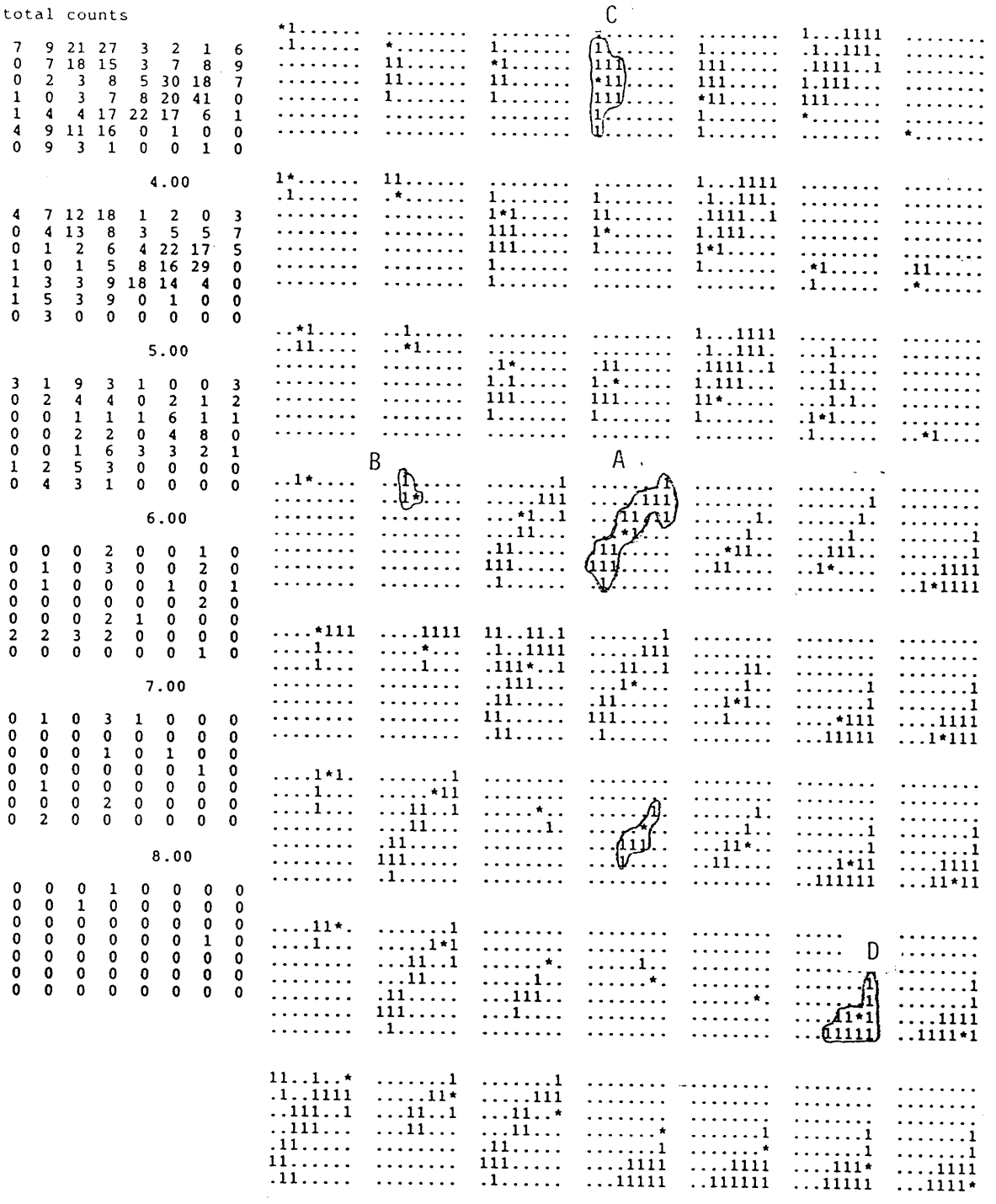


Figure 2-4: Total counts and results of local tests of homogeneity for the Chiburis catalog discretized at one square degree cells.

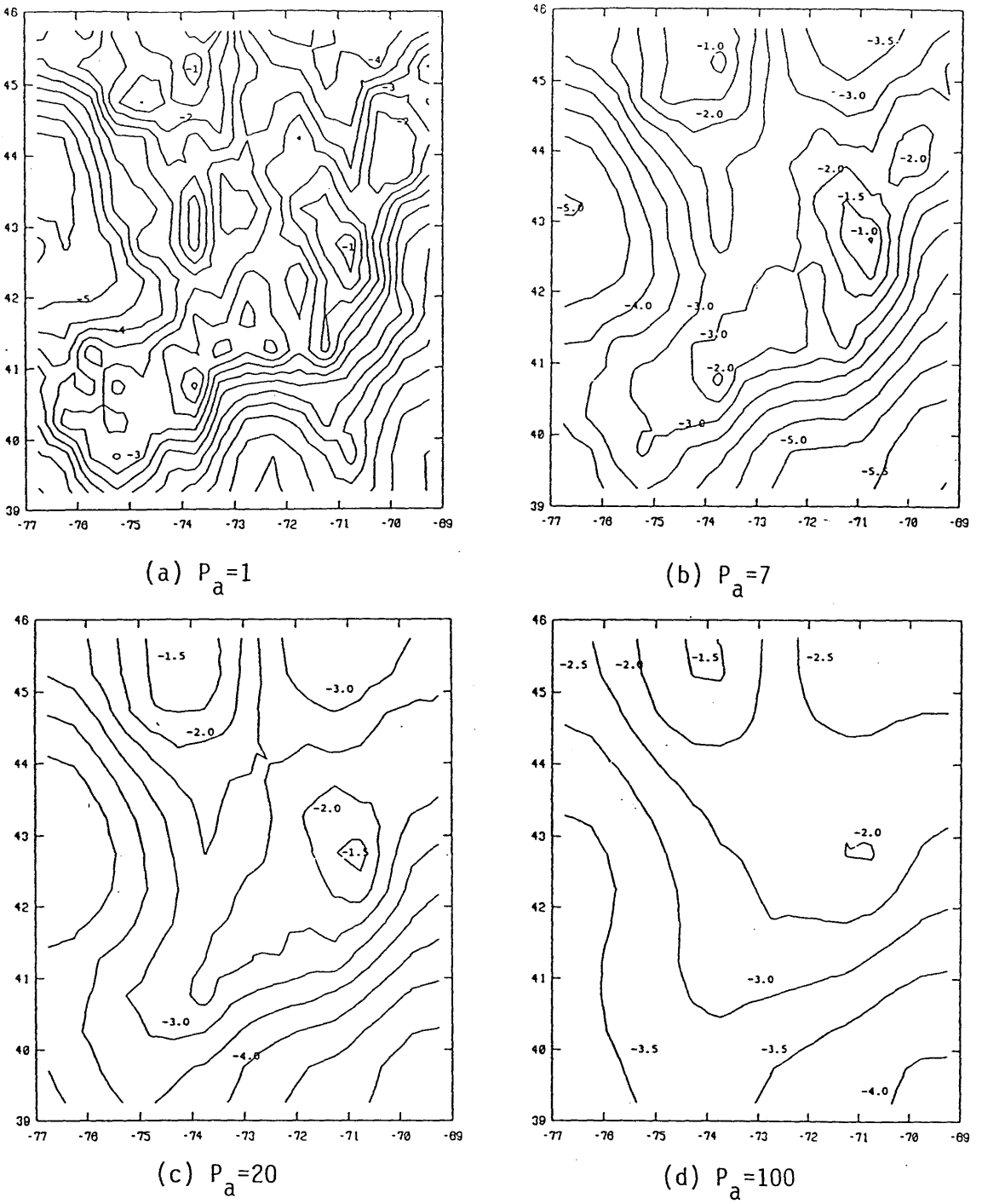
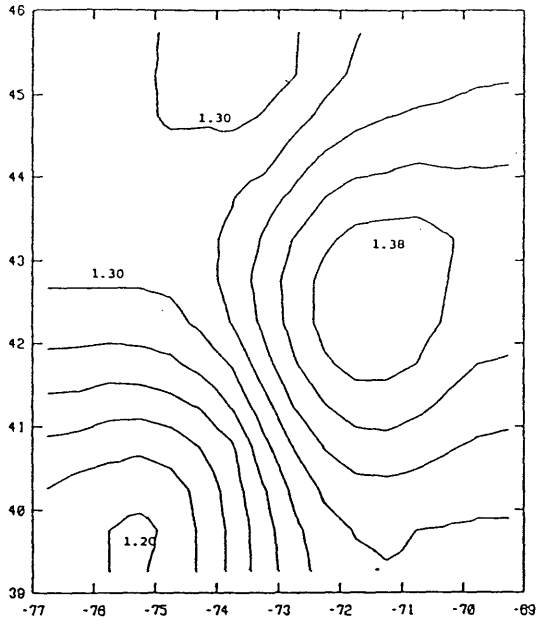
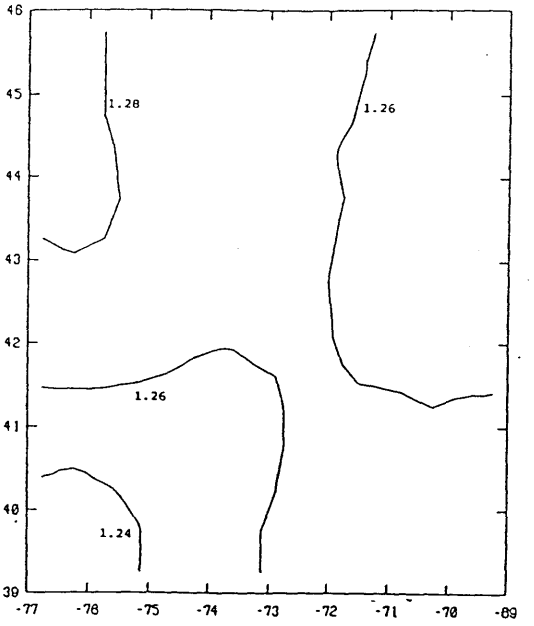


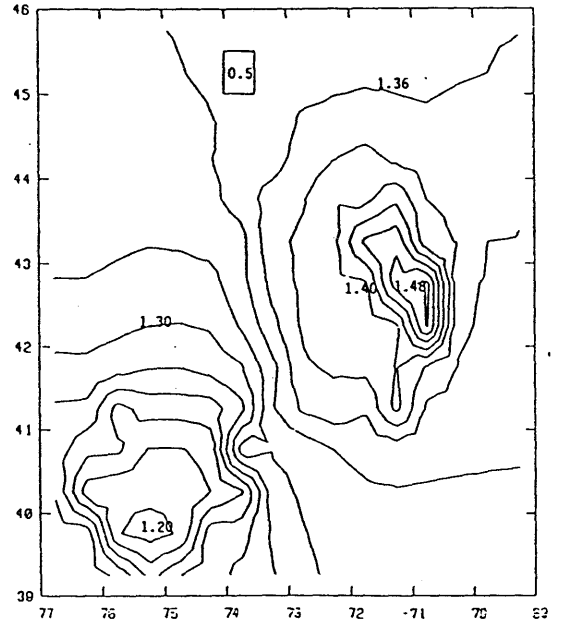
Figure 2-5: Estimates of a(x) for the Chiburis catalog as a function of the penalty on a(x)



(a) P_b=100 (fixed, M=1)



(b) P_b=1000 (fixed, M=1)



(c) P_b=100, alpha=20% (M=1)

Figure 2-6: Estimates of b(x) as a function of P_b and of α. for the Chiburis catalog

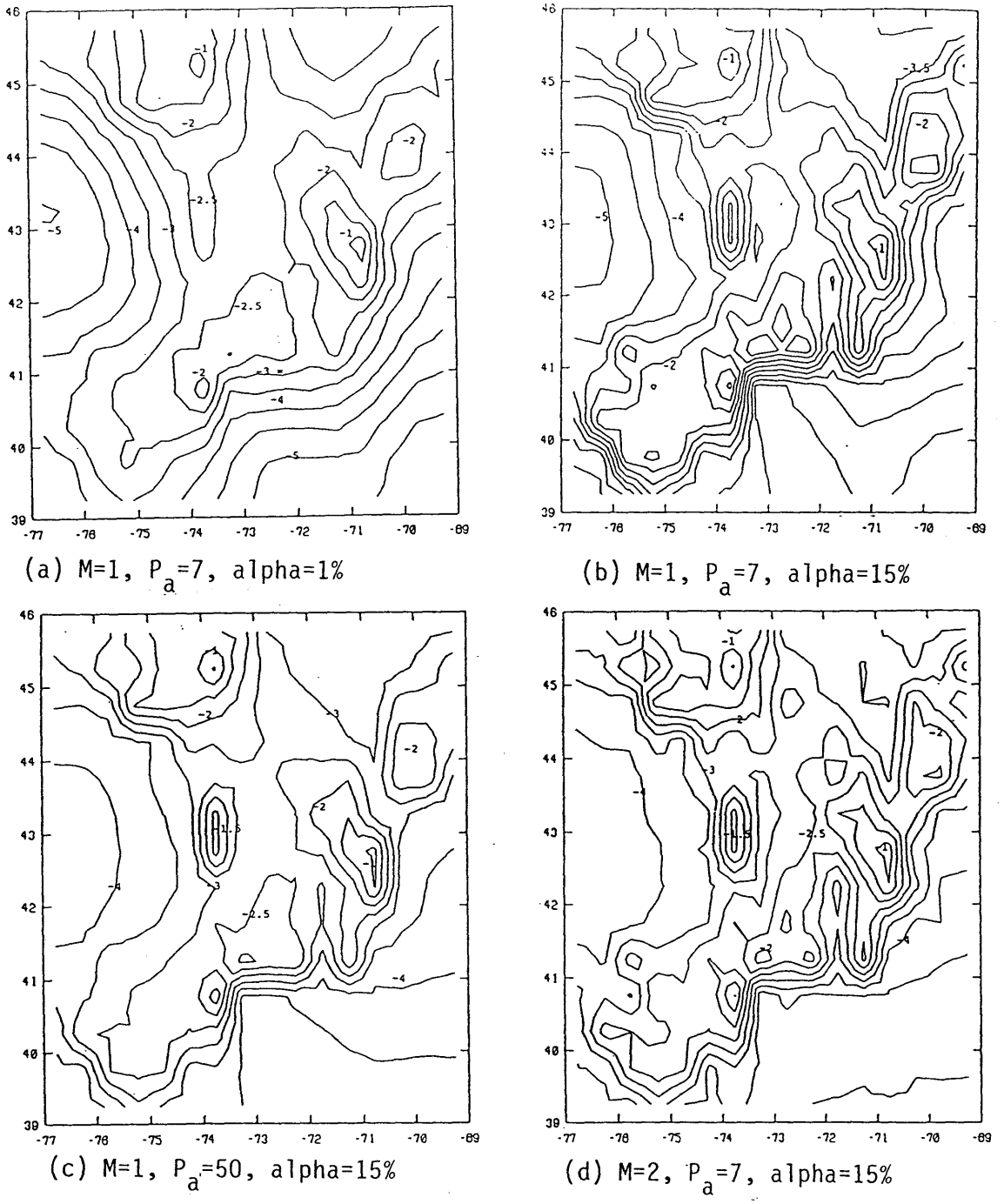


Figure 2-7: Estimates of $a(\underline{x})$ as a function of M, α, and P_a for the Chiburis catalog.

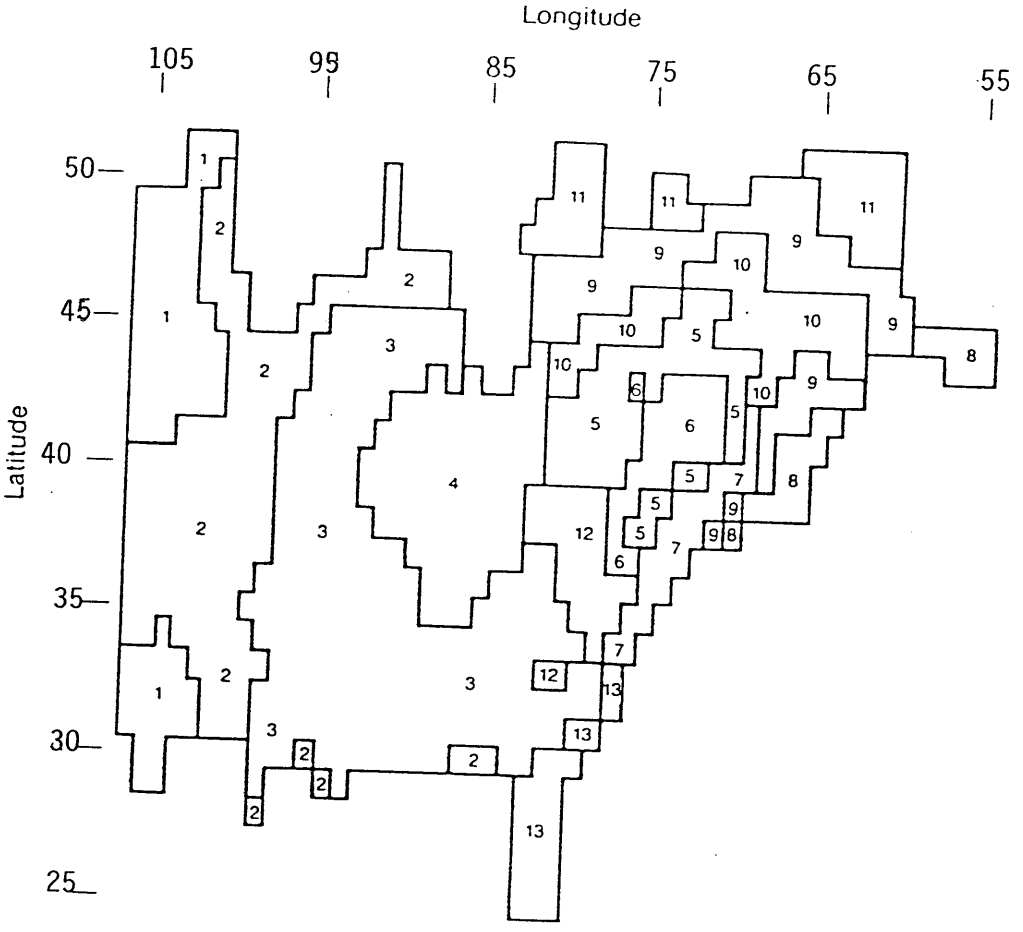


Figure 2-8: Incompleteness regions for the EPRI catalog

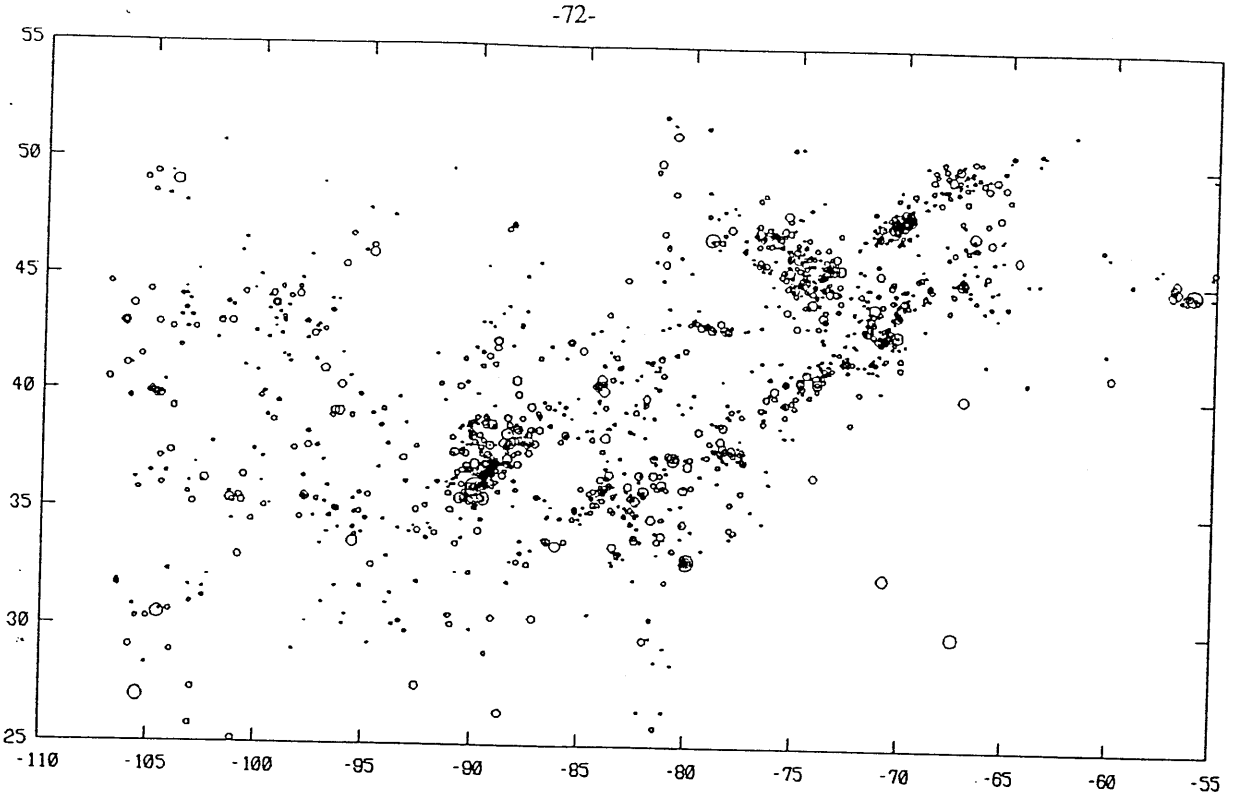


Figure 2-9: Main events for the EPRI catalog

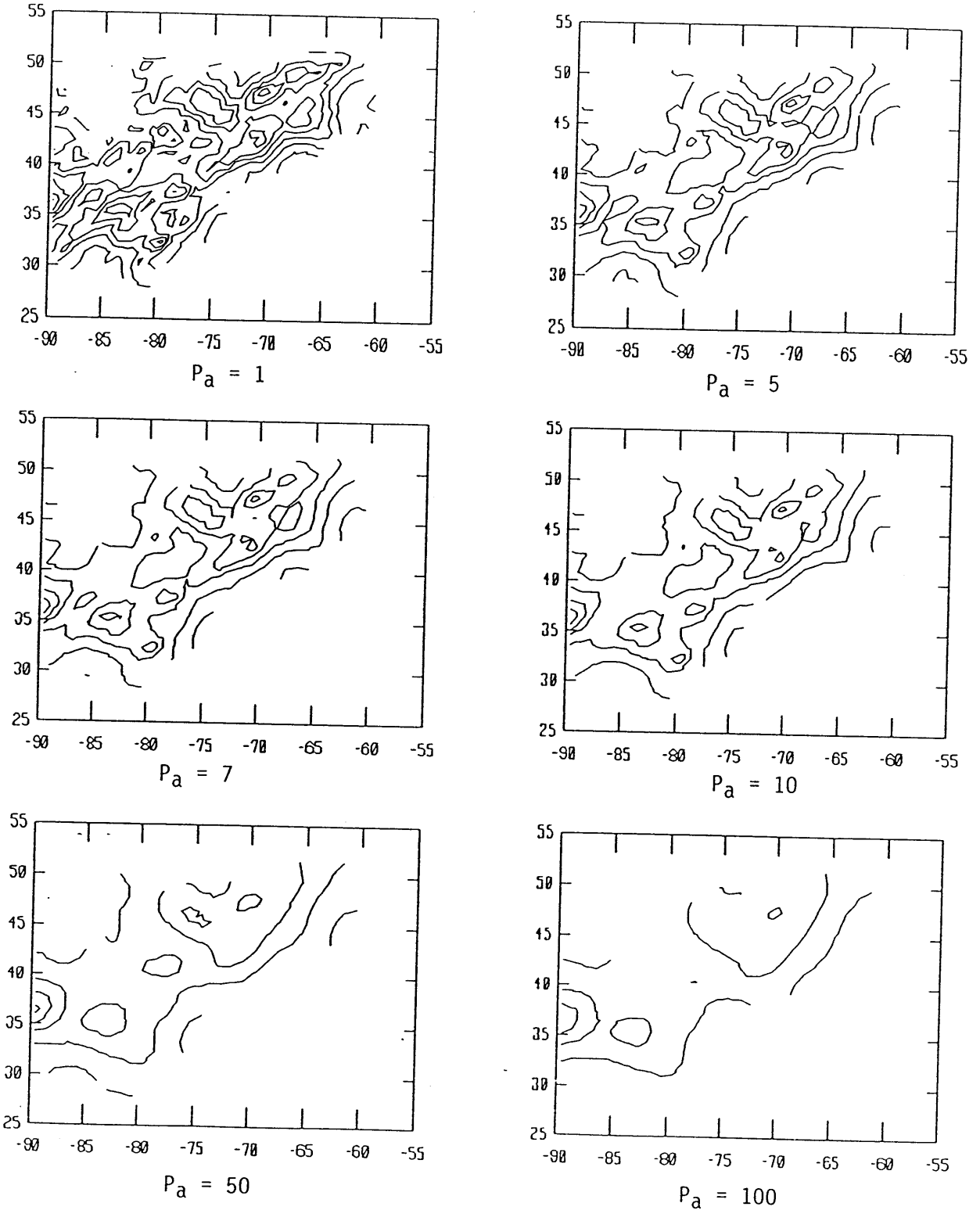


Figure 2-10: Estimates of a(x) for the EPRI catalog as a function of the penalty on a(x) (α=0%)

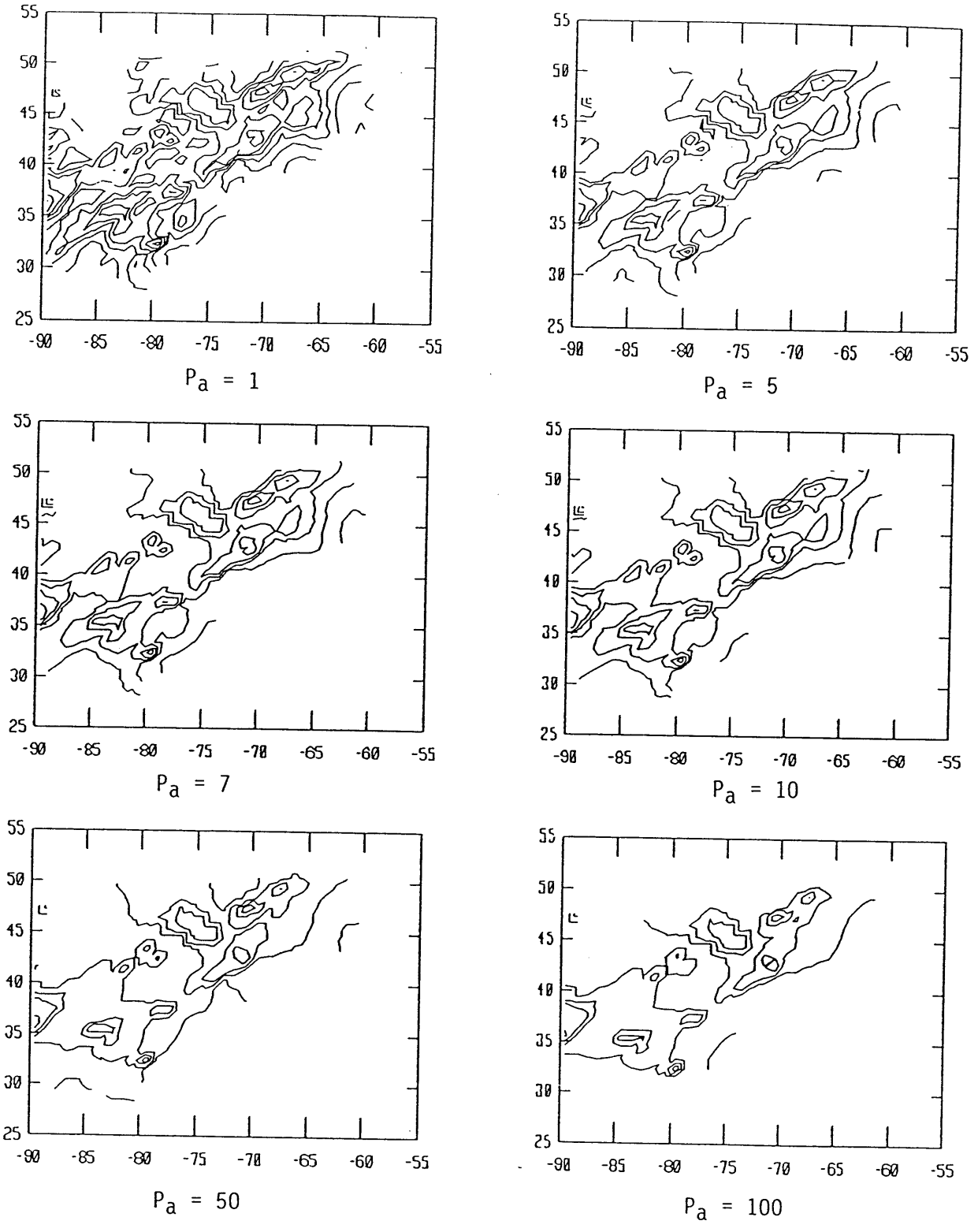


Figure 2-11: Estimates of a(x) for the EPRI catalog as a function of the penalty on a(x) (α=10%)

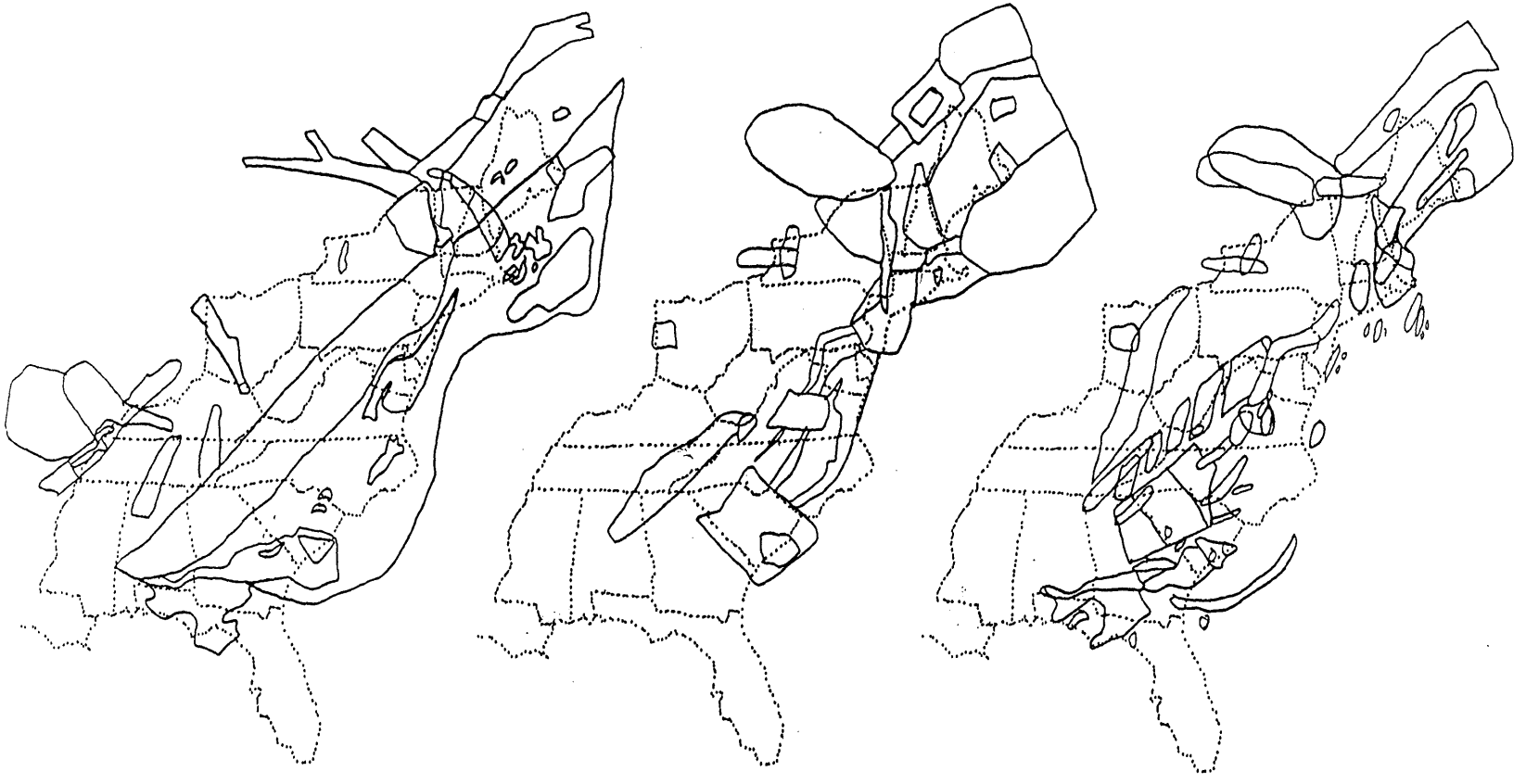


Figure 2-12: Alternative source configurations for the Eastern United States (EPRI 1985)

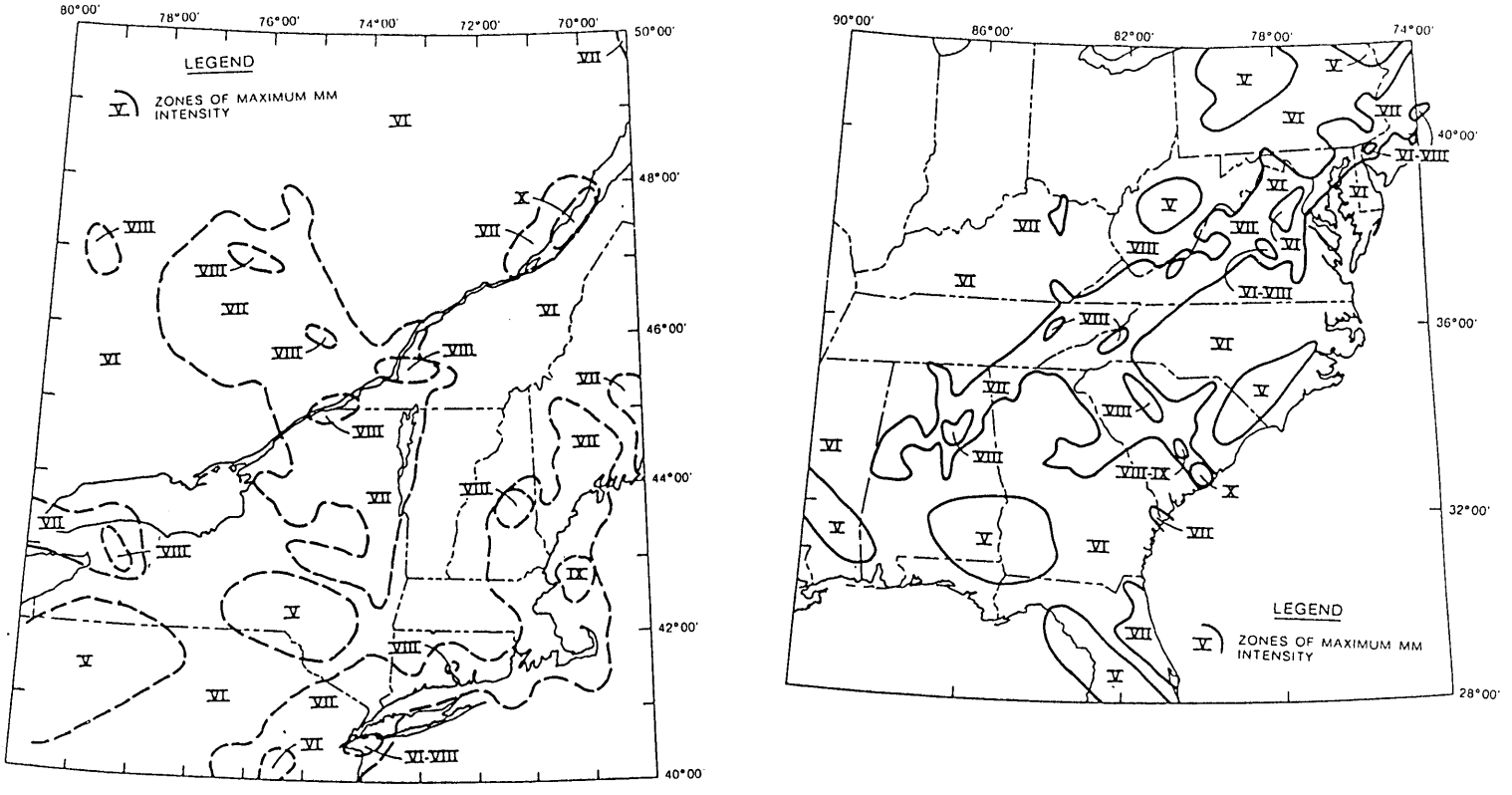


Figure 2-13: Alternative source configurations for the Eastern United States (Barosh 1986)

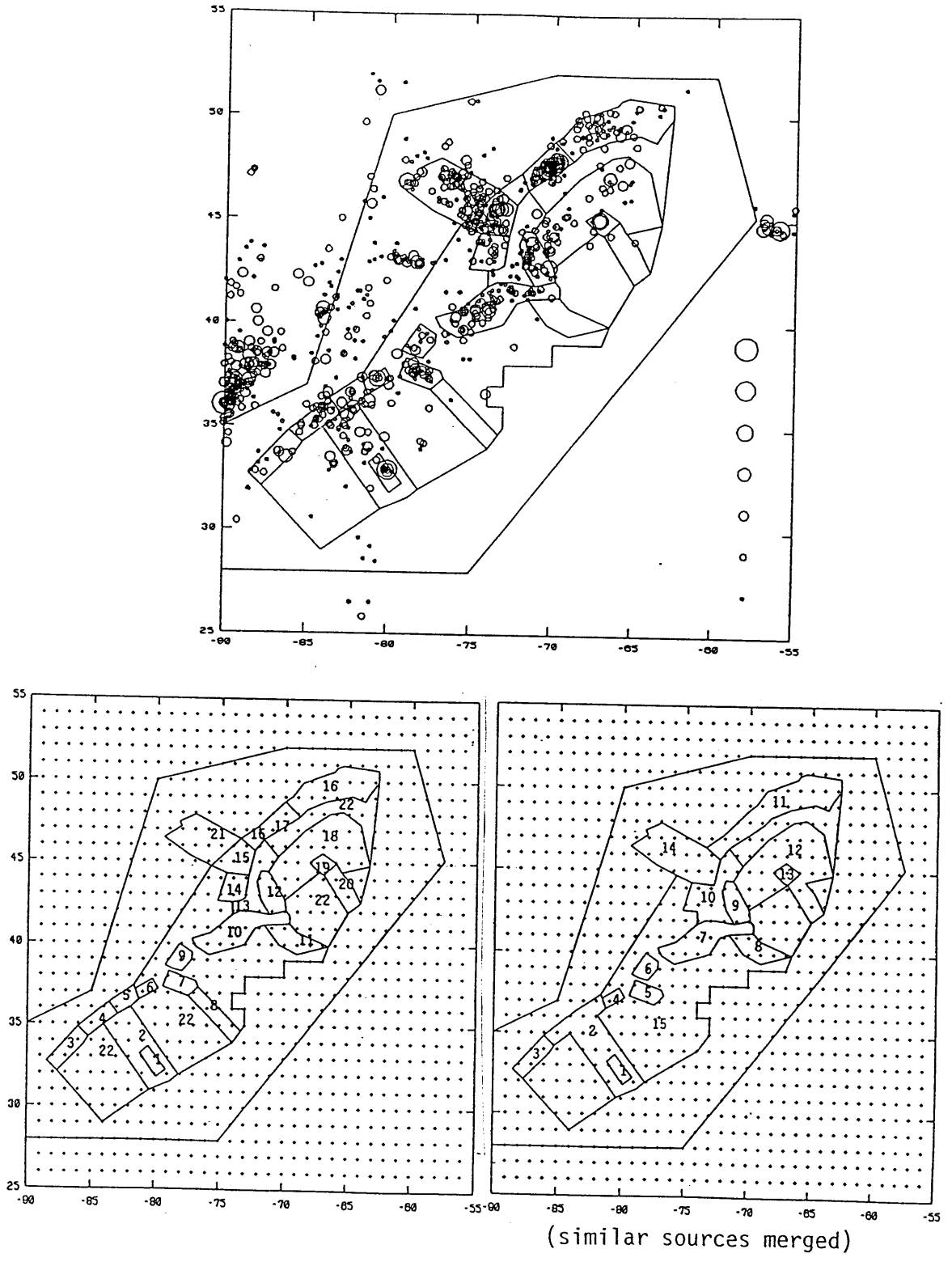


Figure 2-14: Source configuration suggested by Woodward-Clyde (EPRI 1985)

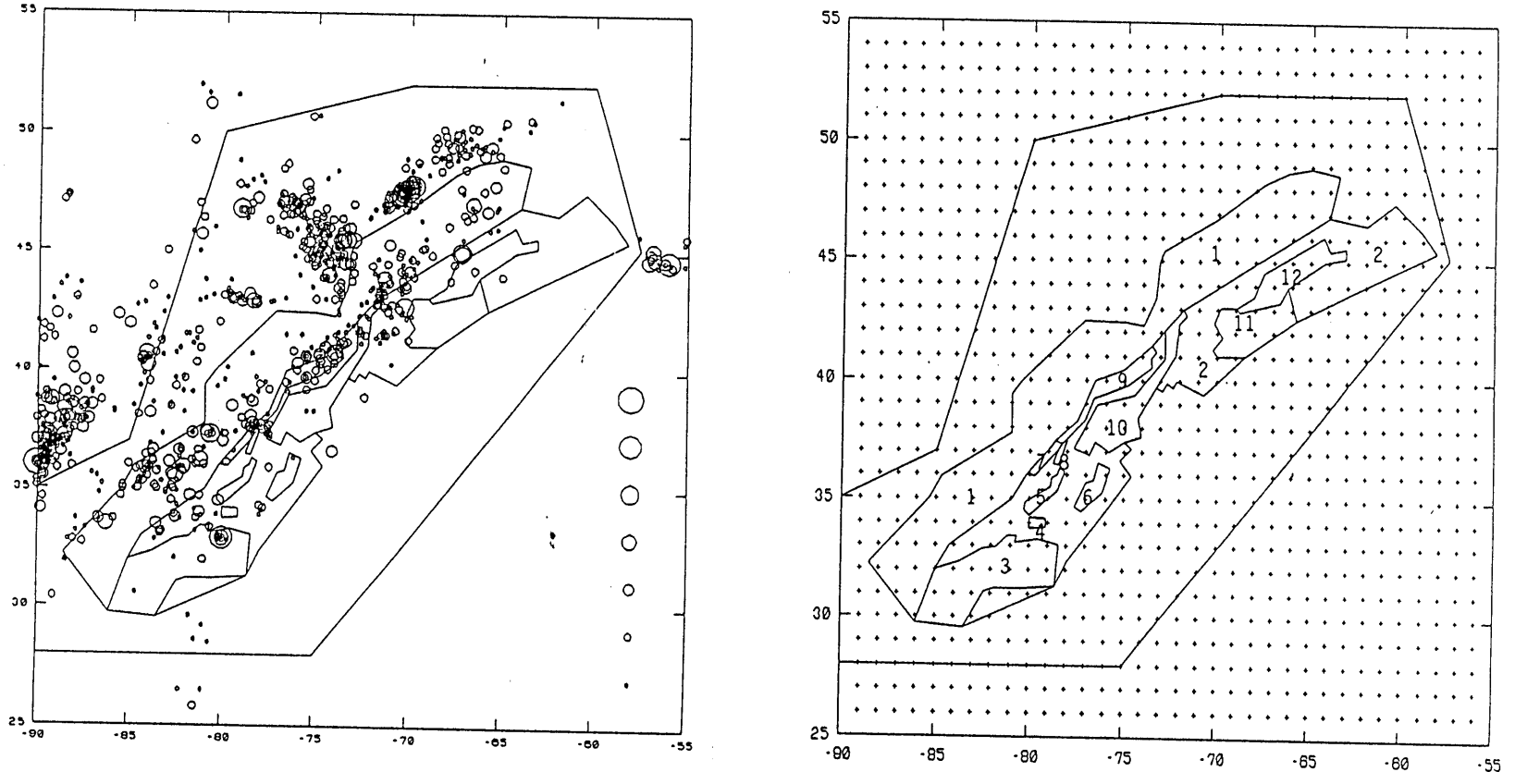


Figure 2-15: Source configuration C

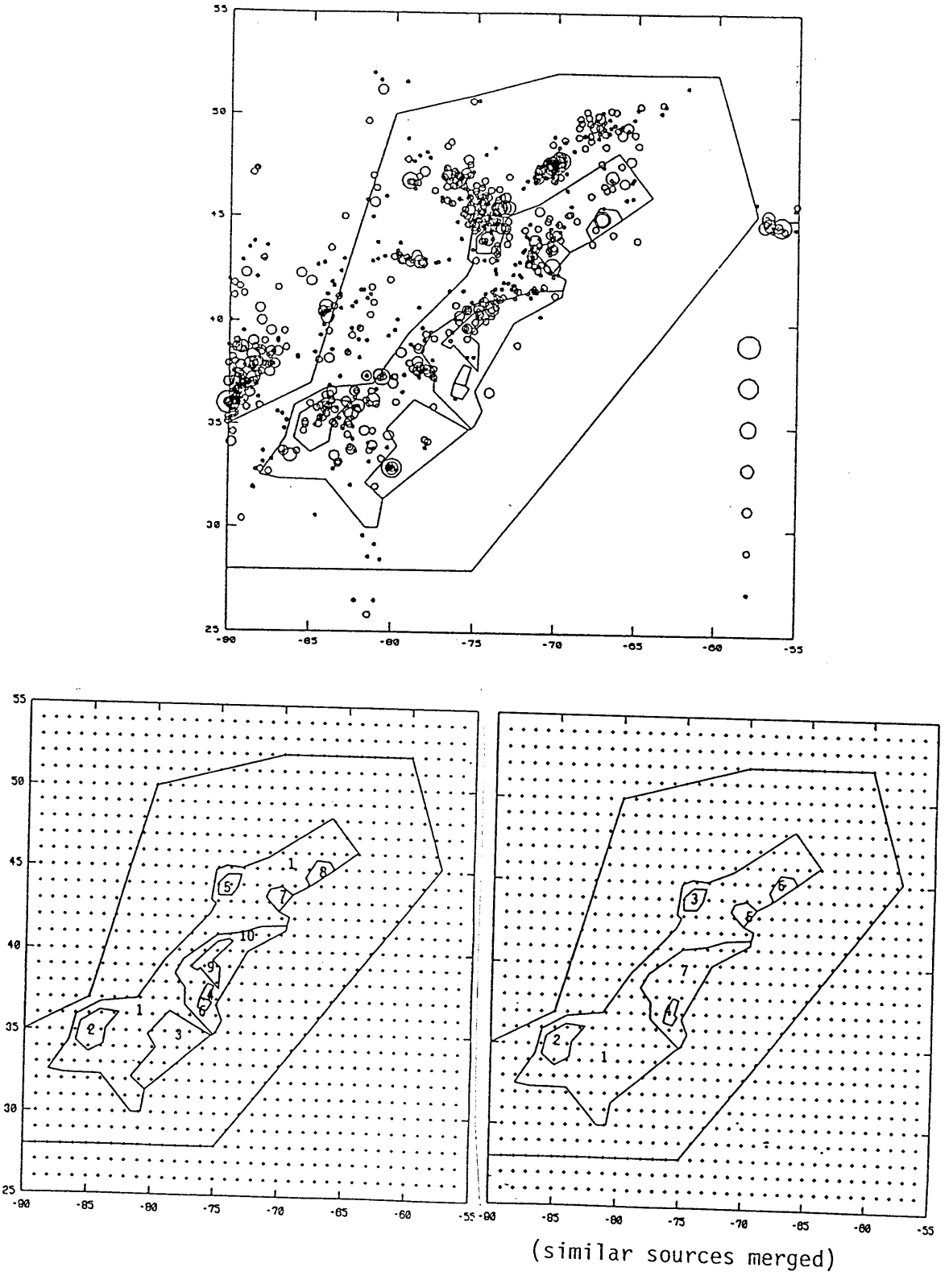


Figure 2-16: Source configuration D

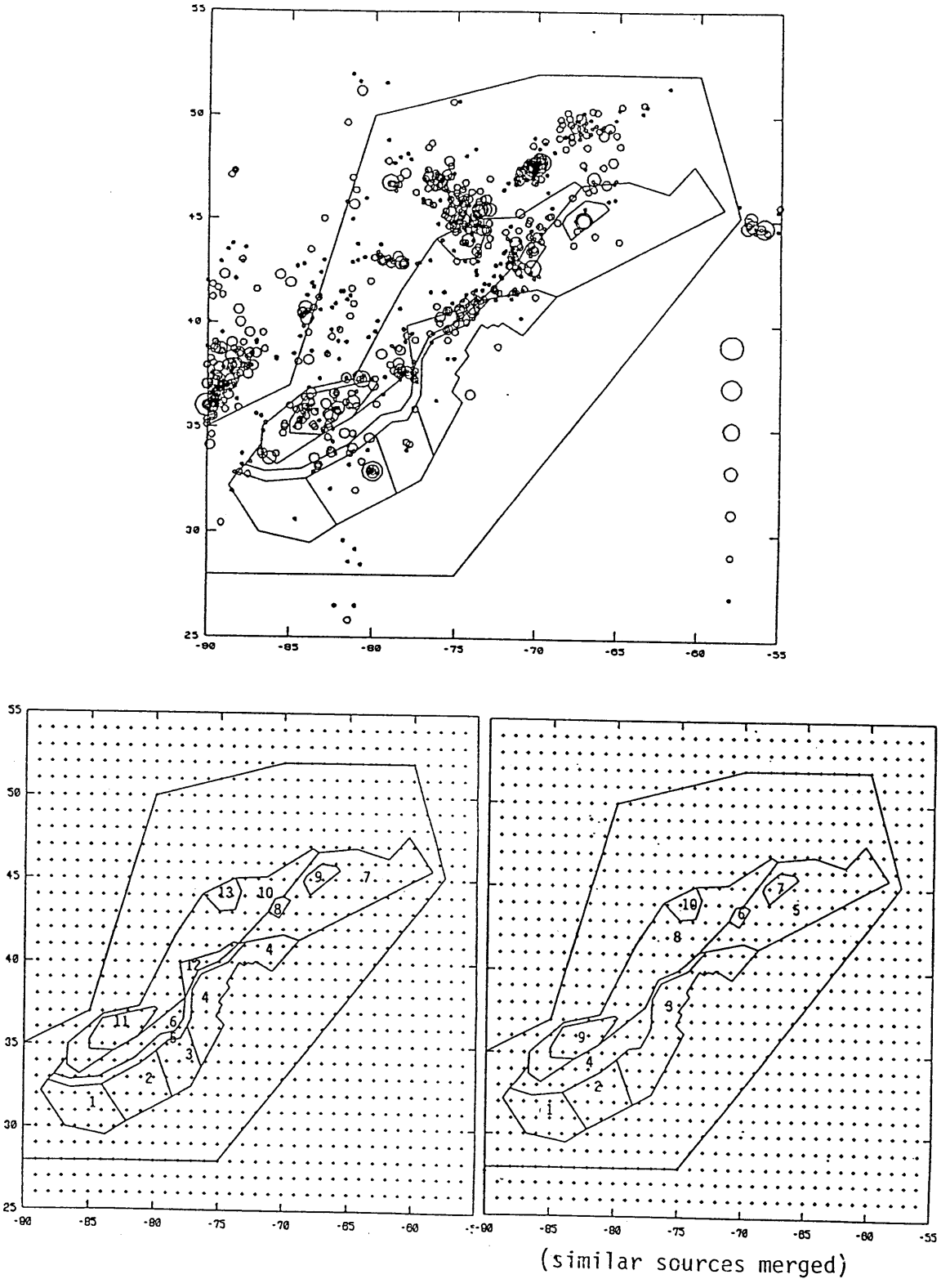


Figure 2-17: Source configuration E

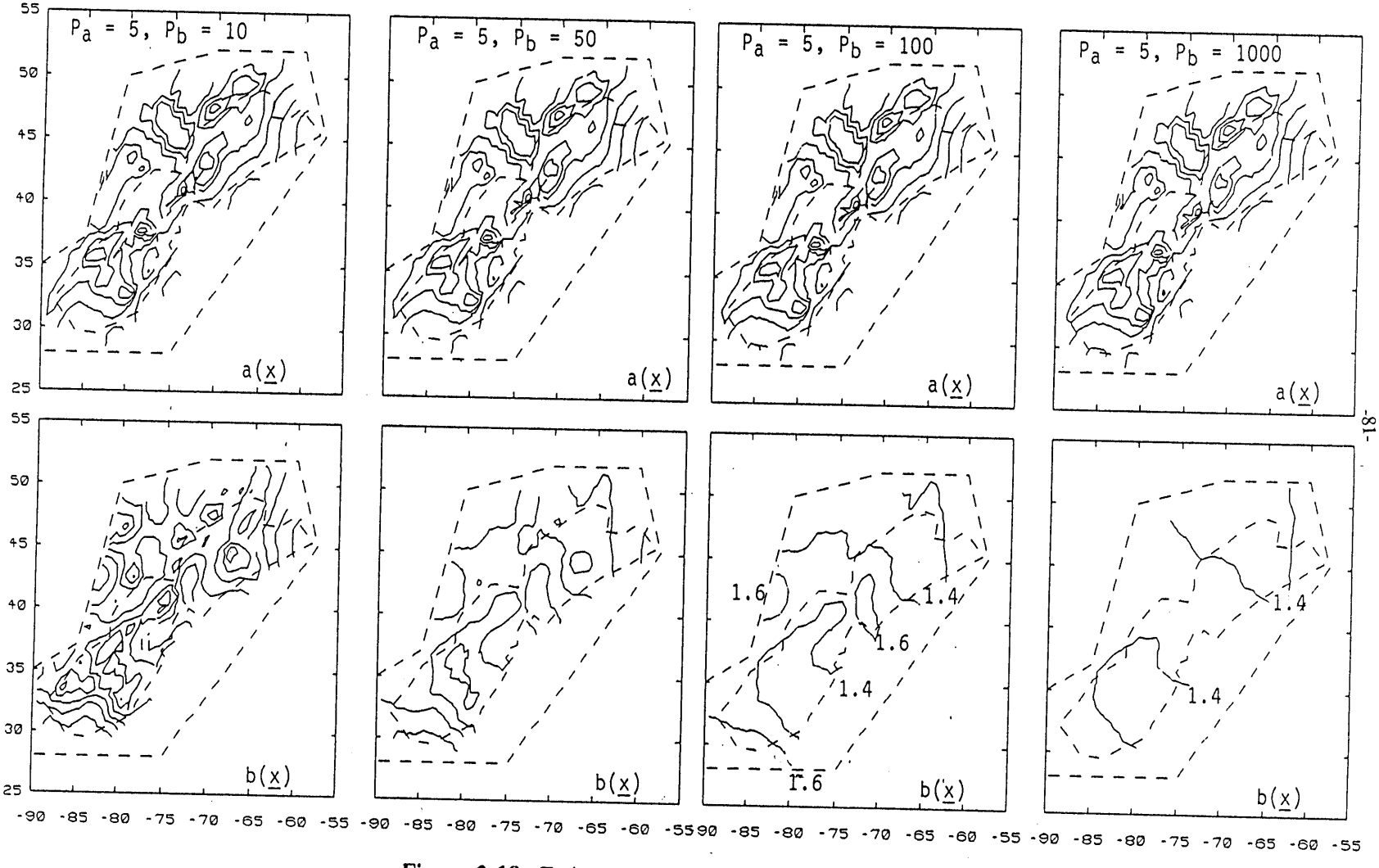


Figure 2-18: Estimates of a(x) and b(x), source configuration C

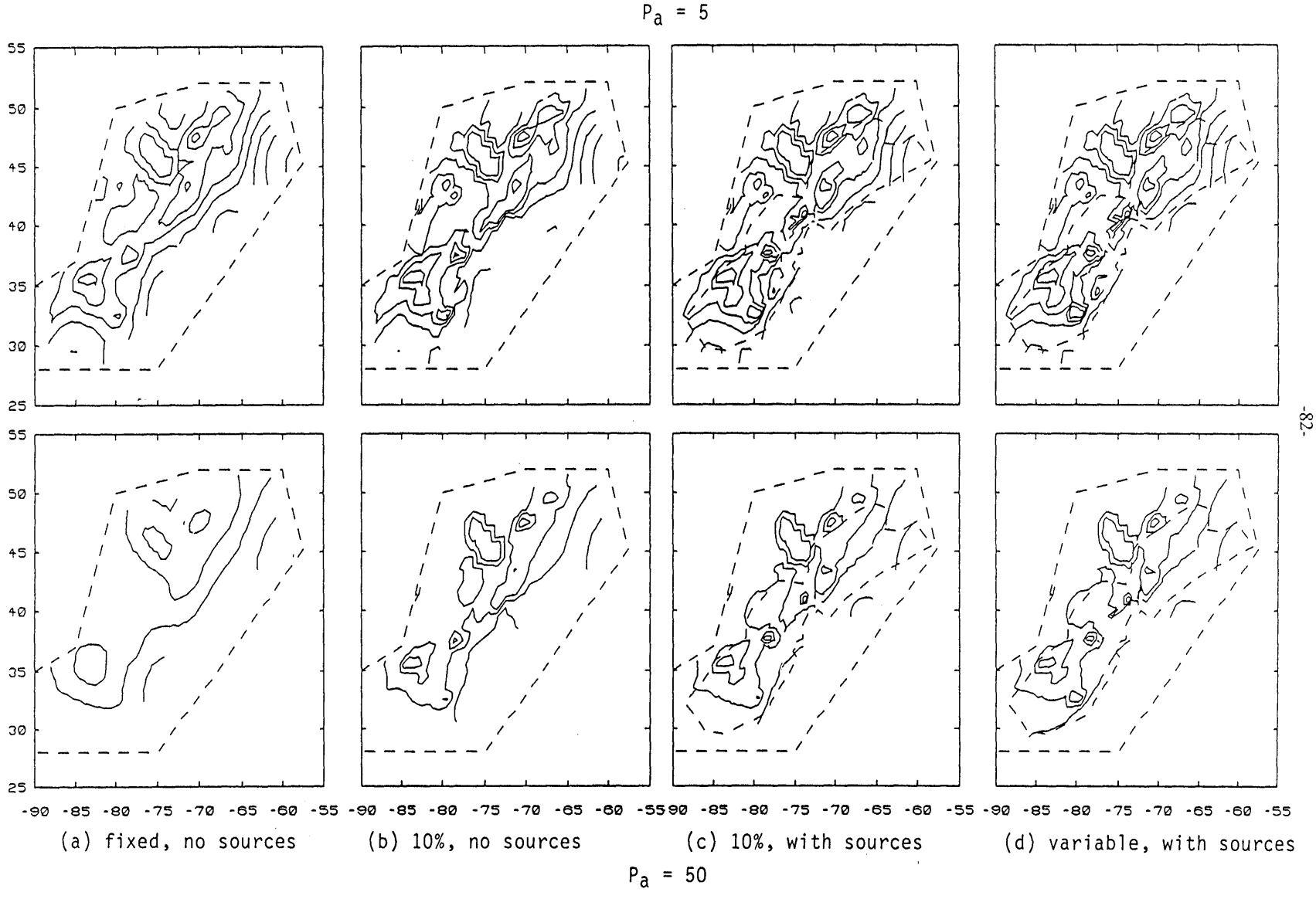


Figure 2-19: Estimates of a(x), source configuration C

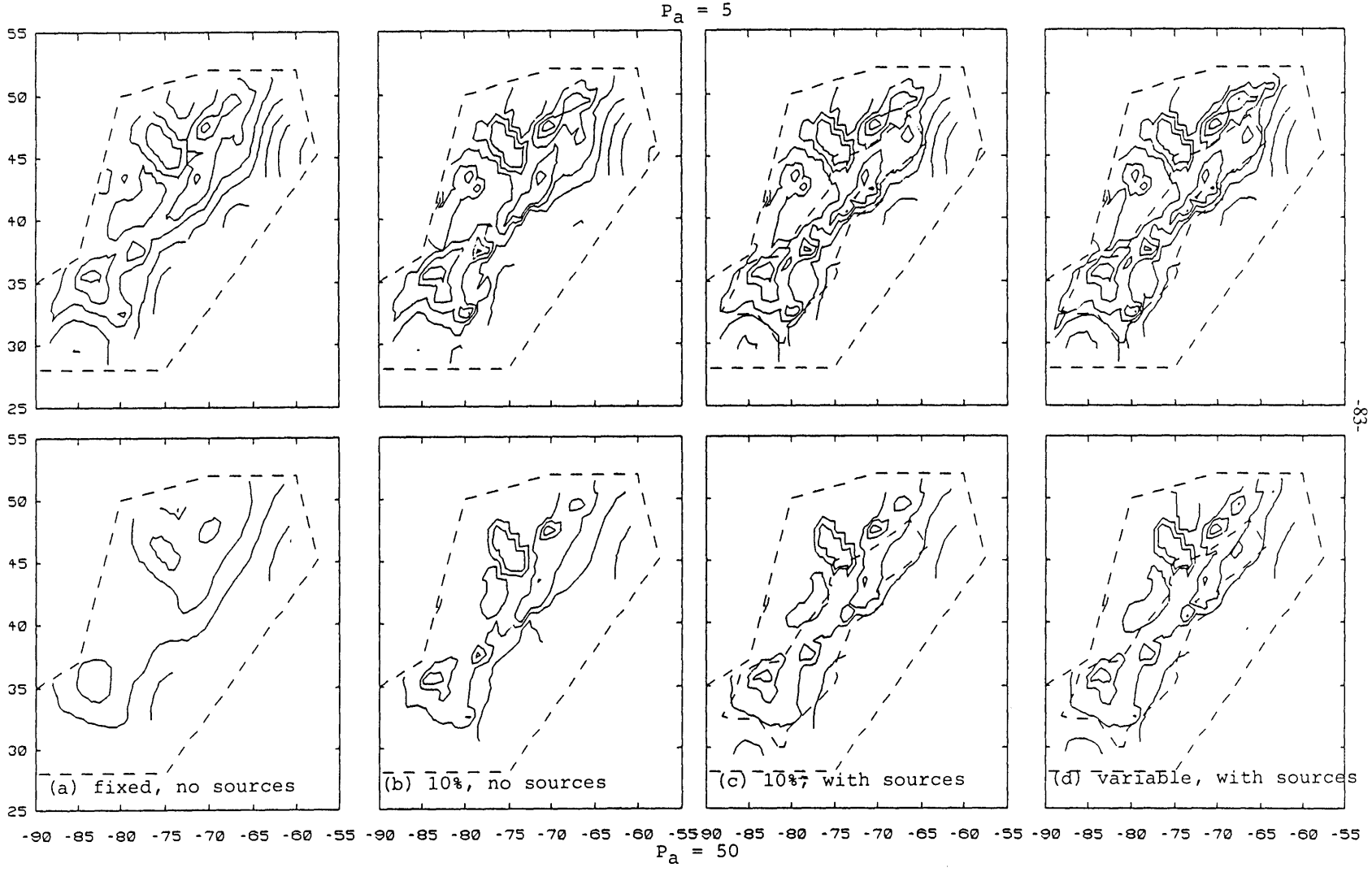


Figure 2-20: Estimates of a(x), source configuration D

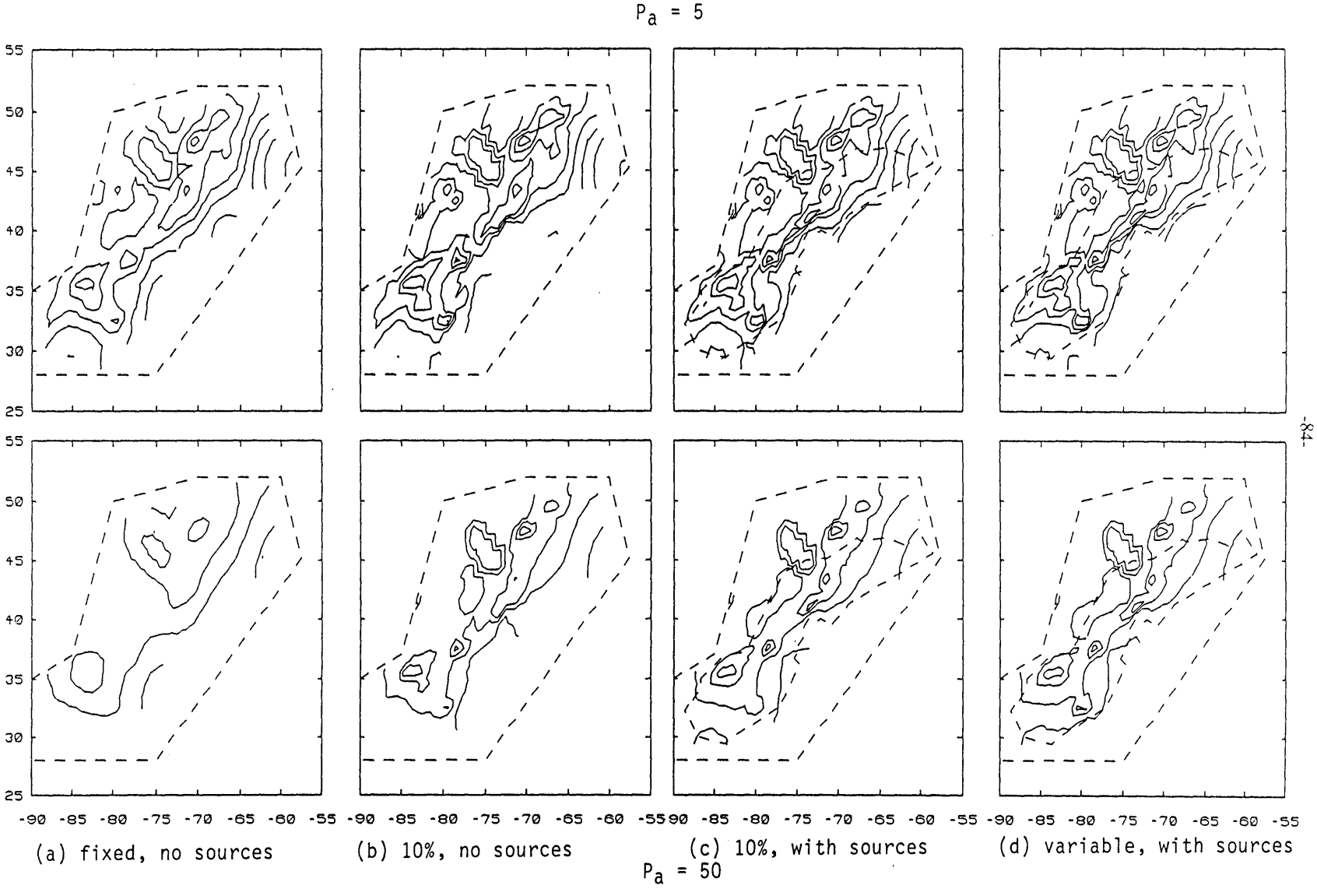


Figure 2-21: Estimates of a(x), source configuration E

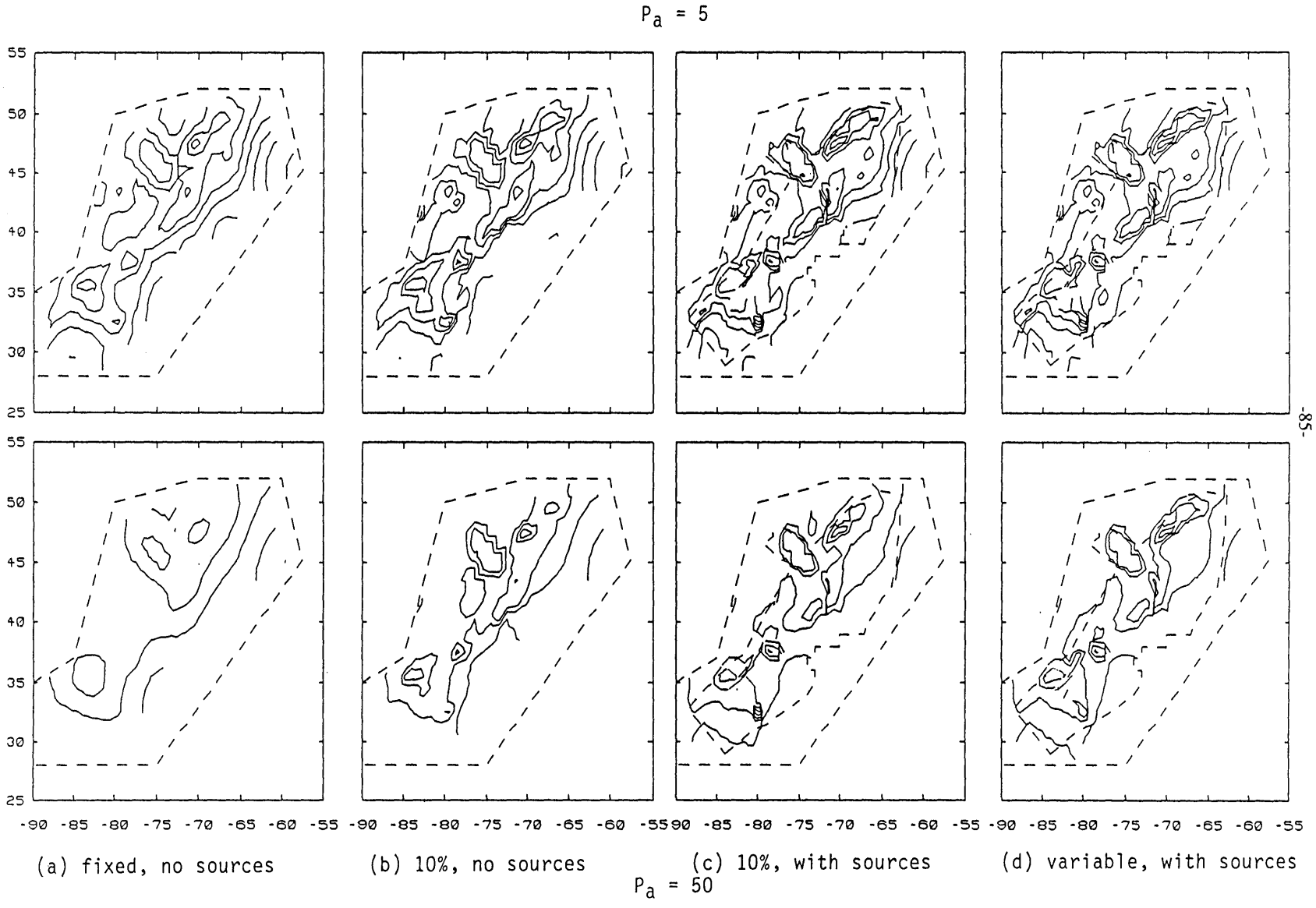


Figure 2-22: Estimates of a(x), Woodward-Clyde source configuration

Chapter 3

Selection of Seismicity Models

3.1 Introduction

In implementing the models of Chapter 2, the analyst must specify the following parameters:

- the discretization in ($\underline{x}$,t,m) for the probability of detection,
- the discretization in space for a($\underline{x}$) and b($\underline{x}$),
- the penalty on a and b (P_a,P_b),
- the time-magnitude envelope for the estimation data set
- and the significance level (α) and spatial extent (M) for the local neighborhoods of the models of section 2.3

Some of these parameters can be selected from considerations exterior to the model. For example, the discretization of the probability of detection is mainly a function of the evolution of the reporting capability as a function of space and time. The choice of other parameters, such as the penalties (P_a,P_b), the size of the spatial cells, the time-magnitude envelope and the parameters for the identification of local neighborhoods are a function of the characteristics of the point process in the mutidimensional space of ($\underline{x}$,t,m).

Two types of optimality criteria are compared for alternative estimators of a($\underline{x}$) and b($\underline{x}$). Either certain observed statistics are set equal to their mean or median values under the model, or cross-validated measures of goodness-of-fit such as the likelihood or negative squared error are maximized. Bayesian procedures are discarted because they are either computationally too demanding (if they require calculation of the posterior distributions of a($\underline{x}$) and b($\underline{x}$) for all $\underline{x}$) or inferior to cross-validation alternatives (if one wants only the a-posteriori most likely values of a($\underline{x}$) and b($\underline{x}$)).

One should distinguish between the maximum-penalized-likelihood method, described in the previous chapter, by which $a(\underline{x})$ and $b(\underline{x})$ are estimated under a given set of conditions (spatial discretization, interpolators $\tilde{a}(\underline{x})$ and $\tilde{b}(\underline{x})$, penalty coefficients P_a and P_b, etc.) and the procedure to optimally select such conditions. The cross-validated likelihood and squared error are defined so that they measure how accurately the model predicts future events. Therefore, these are attractive statistics for selecting seismicity models to be used in earthquake hazard studies.

The optimality criteria can be applied to the entire catalog or to any partition of the catalog in time and magnitude. For example, the model may be fitted to the entire catalog and the selection of the optimal penalties can be based on a comparison of the predictions and observations for the entire catalog or with respect to only the most recent events with large size measure. This may be important if the assumptions of stationarity or exponentiality of the size distribtuion are violated for the period covering the whole catalog but are acceptable locally within short periods of time typical of seismic hazard predictions.

The estimation procedures are applied to the Northeastern United States using the earthquake catalog compiled by (Chiburis, 1981) and to the EUS using the catalog compiled for EPRI (1985). Goodness-of-fit of the optimal models is assessed with respect to the distribution of the earthquakes in space, time and magnitude.

3.2 Target-Statistics Method

Let $\underline{\theta}=[\theta_1,\theta_2,...,\theta_n]$ be the vector of parameters on which the seismicity estimates $\hat{a}(\underline{x})$ and $\hat{b}(\underline{x})$ depend ($\underline{\theta}$ includes the cell size, the penalty coefficients P_a and P_b, etc.). A way to select $\underline{\theta}$ is to choose a set of statistics S_1, ...,S_n that measure in different ways the degree to which the model fits the data and then solve for $\underline{\theta}$ the equations

$$S_i^*(\underline{\theta})=s_i(\underline{\theta}),\ i=1,...n \tag{3.1}$$

where the $s_i(\underline{\theta})$ are the empirically observed statistics and $S_i^*(\underline{\theta})$ are target values for the case when $a(\underline{x})=(\hat{a}(\underline{x})|\underline{\theta})$ and $b(\underline{x})=(\hat{b}(\underline{x})|\underline{\theta})$. For example, one might choose S_i^* to be the mean or the median of $[S_i \mid \hat{a}(\underline{x})|\underline{\theta},\hat{b}(\underline{x})|\underline{\theta}]$, as proposed respectively by (Titterington, 1985) and (Good and Gaskins, 1980). Skilling (1979) uses the 95% fractile of the same distribution. The idea behind the method is that the statistics $s_i(\underline{\theta})$ should be neither excessively good (an indication of overfitting) nor excessively bad (an indication of underfitting).

In our case, goodness-of-fit statistics such as the Chi-square (χ^2), the Kolmogorov-Smirnov statistic, and the likelihood are possible choices. Following is a quick description of each statistic as well as their implementation in the context of seismicity.

3.2.1 Kolmogorov-Smirnov statistic

For testing H_0: $F_X=F_{X_o}$ against the two-sided alternative H_1: $F_X \neq F_{X_o}$, the Kolmogorov-Smirnov statistic is

$$D = \sup_{all\ x} |\hat{F}_X(x) - F_{X_o}(x)| \tag{3.2}$$

where $F_{X_o}(x)$ and $\hat{F}_X(x)$ are respectively the hypothetical true and empirical cumulative distribution functions. Under H_0, this statistic has a distribution that does not depend on the true CDF, F_{X_o}; hence D is a distribution-free statistic. The critical value D_α can be modified for the case when the true distribution is unknown and parameters are estimated from the data. The modification depends on the form of F_{X_o} and is usually approximated through Monte Carlo simulation. In this application, the true distribution is assumed to be known and equal to the estimated distribution.

To apply the test to a two dimensional process, one needs to build an equivalent one-dimensional representation of the process. The procedure used in this application is to build an histogram of the number of events in each cell by joining successive rows of cells as suggested by Skilling (1979) (Figure 3.1). The ordering of the cells may

influence the outcome of the test in some cases. The variance of the empirical CDF is largest for P[x≤X] near 0.5 and decreases to zero as P[x≤X] → 0 or P[x≤X]→ 1.0. This means that we can expect fairly large differences between the theoretical and empirical CDF near the center of the distribution, and much smaller differences in the tails of the distribution. Whether or not an observed difference between the number of observations or expected events in a group of cells is significant will thus depend upon whether it occurs near the center of the distribution or in the tails of the equivalent one-dimensional distribution. However, the test involves only maximum separation of the curves, without regard to where it occurs. A test based on this statistic may well fail to detect substantial departures from the model if they occur in the tails of the distributions, while exaggerating the importance of departures in the middle of the distribution (i.e. the ordering of the cells may turn out to be important).

To implement the test, the estimated model is assumed to be the true model, and the cumulative distribution function is defined as

$$F_{X_o}(\underline{x}_j) = \frac{\sum_{i=1}^{j} \hat{N}(\underline{x}_i)}{\sum_{i=1}^{N_{cell}} \hat{N}(\underline{x}_i)} \tag{3.3}$$

where $\hat{N}(\underline{x}_i)$ is the expected number of events in the i^{th} ordered cell and N_{cell} is the total number of cells. The empirically observed cumulative distribution is similarly defined with $N(\underline{x}_j)$, the number of observations in the j^{th} cell replacing $\hat{N}(\underline{x}_j)$ in the previous expression.

3.2.2 The Chi-square test

Another convenient way to evaluate the goodness-of-fit of a probability density function is to compare the probabilities associated with k non-overlapping intervals covering the range of variation of the random variables with the observed frequencies. Then, the goodness-of-fit problem takes the form,

$$H_o: \; p_1=p_{1o}, \; p_2=p_{2o}, \; \cdots \; , \; p_k=p_{ko} \tag{3.4}$$

$$H_1: \; p_i \neq p_{io} \; \textit{for at least one } i.$$

The most common statistic for testing this hypothesis is Pearson's Chi square (χ^2) test. The test requires that the n observations be grouped into k non-overlapping cells and that p_{io}, i=1,...,k be specified. The estimated model is again assumed to be the true one, $p_{io} = \frac{\hat{N}(\underline{x}_i)}{\sum_{j=1}^{k} \hat{N}(\underline{x}_j)}$; p_i is similarly defined with the number of observations in cell $\underline{x}_i$ ($N(\underline{x}_i)$) replacing the number of expected events in the previous expression. The observations in each cell have a multinomial distribution with parameters $n,p_1,...,p_k$, and the CDF of the random variable

$$C = \sum_{i=1}^{k} \frac{(N(\underline{x}_i)-np_{io})^2}{np_{io}} \tag{3.5}$$

converges to the CDF of the χ^2 distribution with (k-1) degrees of freedom.

Small values of C lead us to conclude that the distribution with probability masses p_{io} is the true distribution, for example, if $C < \chi^2_{(k-1),\alpha}$ in which α is the level of significance. For this application, it is conveninet to use the same grid as the one used for the estimation of the parameters $a(\underline{x})$ and $b(\underline{x})$. If the number of spatial cells is large, k is reduced by aggregating neighboring cells, so that np_i is not less than 5 in each resulting cell (Larsen, 1981) (the χ^2 test is exact only asymptotically, for $np_i \rightarrow \infty$).

3.2.3 Log-likelihood

Another statistic which can be used is the log-likelihood. The log-likelihood $l(\underline{x})$ for cell $\underline{x}$ may be written in different ways, depending on whether and how earthquakes are classified according to size and time of occurence. For example, if the events are classified only according to geographical location $\underline{x}$, then

$$L = \sum_{\underline{x}} L(\underline{x}) = \sum_{\underline{x}} N(\underline{x}) \bullet ln[\lambda(\underline{x})T(\underline{x})] - \lambda(\underline{x})T(\underline{x})) + constant \tag{3.6}$$

where $\lambda(\underline{x})$ is the rate of events defined in equation 2.1, $N(\underline{x})$ is the number of observations and L is the log-likelihood. The expected value and variance associated with the total log-likelihood can be derived assuming that the estimated model is the true one and that each $L(\underline{x})$ is Poisson distributed with parameter $\lambda(\underline{x})$.

The expression for the variance is approximate because the log-likelihoods for each cell are not independent due to the smoothing of $\hat{a}(\underline{x})$ and $\hat{b}(\underline{x})$. Figure 3.2 illustrates how the (log-)likelihood, its expected value and variance vary as a function of the expected and observed number of events in a single cell. The (log-)likelihood is maximum when the number of observations is equal to the expected number of events and the range for which the (log-)likelihood remains larger than its expected value corresponds to models which, by our definition, overfit the data. However, these are within the range of acceptable models if we consider the uncertainty on the (log-)likelihood and accept all models within one standard deviation of the expected (log-)likelihood.

3.2.4 Flagging of significant overpredictions and underpredictions

Another procedure to judge the goodness-of-fit of the model, is to compare the number of expected and observed events in various partitions of the catalog under an assumption of Poisson occurences. Flags are assigned to significant overpredictions (">" for $P[N \leq \hat{N}] < 2\%$, "+" for $P[N \leq \hat{N}] < 10\%$) and significant underpredictions ("<" for $P[N \geq \hat{N}] < 2\%$, "-" for $P[N \geq \hat{N}] < 10\%$). The first significance level corresponds to mild deviations from the model assumptions, while the second significance level corresponds to more severe deviations. In this application, the tests are routinely performed on the total number of events at each location for given partitions of the catalog in time and size measure. Tests are also performed with respect to the distribution of the total number of events in each size interval, as well as each

discretization interval for the probability of detection. The tests can be used to select optimal penalties, by comparing the number of flags to the number of expected flags [20% of the number of tests performed for (+,-), and 4% of the number of tests performed for (<,>)]. The sequential or spatial distribution of the flags is informative with respect to the goodness-of-fit of the model. If flags of a given sign are clustered in space, it indicates that the model fails to capture the trend of spatial variation of seismicity and systematically underpredicts or overpredicts the number of events depending on their location.

3.2.5 Distance measures

Other goodness-of-fit statistics for point processes are the distribution of nearest neighbor distances and the distribution of the shortest distance from a random point to an event from the process (Diggle, 1983). The distribution of these statistics is usually obtained through Monte Carlo simulations. Using the estimated model, several new catalogs are simulated, and the distribution of some specified statistic determined. The estimated model is rejected if the observed statistic is outside a specified range of the ordered simulated values. A possible selection rule for P_a based on this procedure is to select the smoothest model which is not rejected at a specified significance level.

3.2.6 Statistics for the selection of the grid size

Statistics other than the previous ones have been developed in the context of spatial point processes for the selection of the grid size for analysis. The litterature on point processes refers to this issue as a problem of scales of patterns (Diggle, 1983), (Ripley, 1981), (Pielou, 1969). The most common of these procedures is that proposed by (Greig-Smith, 1952): The data is partitioned accordinng to a grid of contiguous cells (or quadrats). The sample variance-to-mean ratio ($\frac{\sigma^2}{m}$), or index of dispersion of the events is calculated for this basic grid and for coarser grids obtained by successive

combinations of adjacent cells into 2x2, 4x4, etc... blocks. The index of dispersion is then plotted against block size and peaks ($\frac{\sigma^2}{m} > 1$) or throughs ($\frac{\sigma^2}{m} < 1$) in the graph are interpreted as evidence of scales of patterns (aggregation or repulsion respectively at the proper scale). Note that $\frac{\sigma^2}{m} = 1$ is the value obtained under complete spatial randomness (Poisson process).

However, these procedures are not applicable in the context of the present application because they assume that the characteristics of clustering or regularity are unifrom for the whole region. The selection of the proper discretization is , however, an important issue and will be addressed in section 3.3.

3.2.7 Combining several statistics

Each of the previous statistics can be used separately to select a particular parameter of the model. Alternatively, one may combine several of them for the selection of a single parameter. Several equivalent procedures are available for this purpose (Gibbons, 1985), (Bradley, 1968), and (Krishnaiah, 1984). The one which has been used in this application is the following, proposed by Good and Gaskins (1980), which assumes that the target is the median value of the statistics.

If several statistics, $S_1,S_2,...,S_n$ are used for the selection of a parameter, the corresponding tail-area probabilities (P_i) can be combined through an harmonic-mean (Good, 1958)($\frac{1}{h_P} = \frac{1}{n}\ [\frac{1}{P_1} + ... + \frac{1}{P_n}]$). However, our problem is special in that left and right tails of the distribution of the statistics correspond to conflicting phenomena, roughness and smoothness. The procedure which is proposed is symmetrical with respect to the two tail areas. Given m tail-area probabilities less than 0.5, $P_1,P_2,...,P_m$, and n that are at least 0.5, $Q_1,Q_2,...,Q_n$, the harmonic means h_p of the P's and k_q of the (1-Q)'s are computed, converted to odds ratios, and then weighted through a geometric mean,

$$O = [\ \frac{h_p}{(1-h_p)}\ ]^{\frac{m}{(m+n)}} [\ \frac{(1-k_q)}{k_q}\ ]^{\frac{n}{(m+n)}} \tag{3.7}$$

and finally converted to a resultant probability

$$R = 2\ [\ \frac{O}{(1+O)} - 0.5\ ] \tag{3.8}$$

where $0 \leq R \leq 1.0$. Good and Gaskin suggest that the minimum value of R be used for the selection of the optimal parameter.

3.2.8 Applications

The target-statistics method is illustrated here for the selection of the parameter P_a which controls the smoothness of the estimator $\hat{a}(\underline{x})$. The analogous parameter for $\hat{b}(\underline{x})$ is fixed to 1000, a value which produces high smoothing. The catalog used is the one compiled by Chiburis (1981) with space discretized to one degree cells. Modified Mercalli Intensity (I_o) is used as a measure of earthquake size. The discretizations in time and intensity are selected as a function of the accuracy and history of reporting. Intensity is already reported in a discrete scale, and for it, unit intervals are a natural choice. The discretization in time and space for the probability of detection is determined from an analysis of reported events and the mode of reporting (VanDyck, 1986) and was described in the previous chapter. In this and in following applications, a prior mean of 1.3 is assigned to b, which is the value obtained under complete smoothing of the b parameter. A prior variance of 10 is specified on the basis of work by VanDyck (1985, Chapter 4). This is a mild prior, which however stabilizes the estimate of $b(\underline{x})$ in areas of sparse data. The interpolators $\hat{a}(\underline{x})$ and $\hat{b}(\underline{x})$ are the averages of a and b over the eight cells that are closest to $\underline{x}$.

Figure 3.3a shows the variation with P_a of the Chi-square and the Kolmogorov-Smirnov statistics computed for one degree cells and all of the observations used for estimation. The dashed line corresponds to the diagnostic quantity proposed by (Good

and Gaskins, 1980)(Equation 3.8). If one selects medians as the target values S^*, one finds optimal penalties between 20 and 35 and rather smooth associated estimates of $a(\underline{x})$ when comparing the expected and observed number of events. The grid of aggregated cells used to compute the Chi-square statistic is shown in Figure 3.4a. Note that when we are comparing the observations and predictions for the whole catalog (the estimation and validation data sets are the same), the exceedence probabilities associated with each statistic are monotonically increasing. Notice that the target-statistics procedure is originally intended to be used with the full data set.

The optimal choice is less clear when the criterion is applied to different subsets of the catalog. In this case, the statistics are computed for only part of the data set which was used for the estimation of the model. The aggregated cells that are used to compute the Chi-square statistic are shown in Figures 3.4b,c,d for each subset. Notice that due to the smaller number of events, the total number of observations decreases and so does the power of the tests. For events with intensity greater than 4.5, the χ^2 and Kolmogorov-Smirnov statistics vary differently with P_a. The tail probabilities associated with the χ^2 statistic increases monotonically with the penalty and its median corresponds to P_a=5 (Figure 3.3b). The tail probabilities associated with the Komogorov-Smirnov statistic are not as well behaved because the test is based on sparse observations within one degree cells. In addition, many of the most active cells are located at the periphery of the region and end up in the tails of the one-dimensional histogram (obtained by joining successive rows of cells) where the test lacks power. The criterion proposed by Good and Gaskins is controlled by the variation of the χ^2 statistic and identifies an optimal penalty P_a=3. Such a result is consistent with the observation that events of larger intensity are more likely in regions of high activity.

The data sets corresponding to the most recent events (since 1915) with $I \geq 3.5$ or $I \geq 4.5$, contain fewer events than the previous case. For the recent events with $I \geq 3.5$, the

χ^2 test favors no smoothing while the K-S test fails to identify an optimal penalty (Figure 3.3b). For the recent events of larger intensity both tests fail to identify an optimal penalty (Figure 3.3d). A comparison of the data sets for the periods 1627-1915 and 1915-1981 shows that the lack-of-fit is a consequence of a change in the pattern of seismicity between the two periods (Figure 3.5b). The partition of the region corresponds to homogeneous seismogenic provinces proposed by (WGC, 1980) and is displayed only to facilitate the description of the data and results (Figure 3.5a). In this case, the Adirondack Uplift, Piedmont Atlantic Gravity, and Merrimack Synclinorium show an increase in activity while there is a sharp decrease of activity in the Massachusetts Thrust Fault Complex. This raises the issue of possible lack-of-fit of the model which will be addressed more in detail in later sections. In consequence, this procedure should be used only with respect to the full data set for the selection of optimal penalties.

A simple test which illustrates the lack-of-fit is the flagging of significant deviations between model predictions and observations for different partitions of the catalog. For observations within on degree cells, Figure 3.6 shows how the total number of flags varies as a function of P_a and Figure 3.7 shows how these are distributed spatially for a selection of penalties and partitions of the catalog. For identifying the optimal penalty, the criterion is formulated such that the number of flagged cells ("+" or "-") corresponds to 20% of the number of cells (N_{test}) for which the test is powerful. For the events with $I \geq 3.5$ and $t \geq 1627$, N_{test} is approximately 40 and the the optimal penalty is P_a=7 (Figure 3.7a). For other partitions of the catalog, the test is not as powerful because of the smaller number of evnets (for $I \geq 4.5$ and $t \geq 1627$, N_{test}=25; for $I \geq 3.5$ and $t \geq 1915$, N_{test}=35; and for $I \geq 4.5$ and $t \geq 1915$, N_{test}=16). For the events of larger intensity, the flags indicate that the number of events are underpredicted in the Piedmont Atlantic Gravity province indicating that the estimates

of b($\underline{x}$) are locally too large (Figure 3.7b,d). Finally, the procedure fails to identify an optimal penalty for $I \geq 3.5$ or $I \geq 4.5$ and $t \geq 1915$, because the number of flags is too large whatever the penalty due to lack-of-fit of the model. Again this procedure is intended to be used with the full data set for the selection of the optimal penalty and tests on subsets of the catalog are only useful for analysing the goodness-of-fit of the selected models.

The log-likelihood based on the total number of observations in each cell, L=ln$l$ with l in Eq. 3.6, is plotted in Figure 3.8 as a function of P_a. Also shown in that figure are the expected value and the one-standard-deviation bounds on L, under the assumption that the estimated model is the true one. For $I \geq 3.5$ and $t \geq 1627$, the log-likelihood is equal to its expected value for a penalty P_a=20 and the one standard deviation bounds on L correspond to a range of P_a between about 12 and 30. Figure 3.9 shows L-E[L] decomposed in space for differnet penalties P_a and for L computed using the entire catalog. For very small penalties, L is greater than E[L] across all of the region because there is a perfect match betwewn the predictions and observations. At the optimal penalty, L is smaller than E[L] only in zones of high activity. For higher penalties, L is dominated by contributions from the most active cells (near Boston and south eastern Quebec). These results show that the likelihood remains close to its expected value in regions of low activity. For $I \geq 3.5$ and $t \geq 1915$, the optimal penalty is small ($0 \leq P_a \leq 3$), which is consistent with the previous results. For the other two cases ($I \geq 4.5$ for $t \geq 1627$ or $t \geq 1915$) the procedure fails to identify a particular model given the uncertainty on the log-likelihood.

3.2.9 Conclusion

The Chi-square and Kolmogorov-Smirnov tests require some subjectivity respectively to aggregate cells and to construct a one-dimensional distribution function. These

statistics perform adequately given a sufficient number of observations, and the combined statistic proposed by Good and Gaskins is then useful in identifying penalties for which the observed statistics are close to their median value. However, the procedure fails to identify a model when applied to smaller subsets of the catalog and provides little insight on the possible sources of lack-of-fit when present. The flagging procedure has the advantage of visually displaying where lack-of-fit occurs and is a useful tool for the subjective evaluation of the goodness-of-fit, however, the procedure is not powerful when there are few observations in each cell.

The target-statistics method has the advantage of being intuitive and easy to implement. However, the method lacks predictive interpretation and cannot be used to rank alternative estimators. On the latter scores, cross-validation procedures, which are described next, should be preferred.

3.3 Cross-validation

Cross-validation aims at maximizing the predictive ability of a model: Suppose that, besides the original earthquake catalog (estimation data set E), additional observations (validation data set V) are available from the earthquake generation process. Also let S be a statistic that compares the validation data with predictions when the model is fitted to the estimation data. It would be natural then to rank alternative estimators of $a(\underline{x})$ and $b(\underline{x})$ based on the values of S.

Using different statistics S will usually lead to the selection of different optimal penalties. For the purpose of seismic hazard analysis, where one is interested in the probabiliy of occurence of events, the maximization of a likelihood-based criterion appears to be a natural choice. The cross-validated log-likelihood is given by

$$L_{cv} = \sum_{t}\sum_{I}\sum_{\underline{x}} L[N(\underline{x},I,t)|(\hat{a}(\underline{x}),\hat{b}(\underline{x}))_t] \qquad (3.9)$$

where $(\hat{a}(\underline{x}),\hat{b}(\underline{x}))_t$ are estimators from observations prior to the t^{th} time interval. The method of cross-validated likelihood is a natural development of the idea of using the likelihood to judge the adequacy of fit of a statistical model.

Another popular statistic is the squared error,

$$SE_{cv} = \sum_t \sum_I \sum_{\underline{x}} (N(\underline{x},I,t)-\hat{N}(\underline{x},I,t))^2 I(\hat{a}(\underline{x}),\hat{b}(\underline{x}))_t] \qquad (3.10)$$

Figure 3.10 shows how the log-likelihood and squared error vary as a function of the number of observations (N) and recurrence rate (λ). The sqaured error is symmetrical with respect to N and λ and only depends on the absolute deviations $|N-\lambda|$, while the log-likelihood penalizes the same deviations, but in a way that depends on λ (higher penalty for lower expected counts). The only combination for which the log-likelihood is more sensitive than the squared error is when there is a large number of observations and a small recurrence rate which does not occur in the models fitted in the following sections.

In practice, validation data sets are not available and the method is applied by partitioning the actual sample in various ways into an estimation subset E_i and validation subset V_i. The cross-validated estimator is the one which optimizes the total score, say $\sum S_i$ *or* $\prod S_i$ (Silverman, 1985, Titterington, 1985, Hand, 1982). The estimation data set E_i associated with the validation data set V_i can be defined from any subset of the data remaining after removing V_i. Two basic methods of defining the validation data sets are extrapolation or interpolation. In extrapolation, only the data preceeding a validation interval is used for the associated estimation data set while in interpolation all of the remaining data is used. The extent of the estimation data set preceeding a validation interval can be defined so that it includes increasing larger portions of the recent seismicity. In the presence of non-stationarities, this may be useful in determining the extent of the memory of the process.

It is recommended that the associated estimation subsets E_i contain only data prior to V_i. The reason why data following V_i should not be included in E_i is that, if the assumption of stationarity and Poisson independence do not hold exactly, the use for estimation of events on both sides of the validation subset artificially increases the prediction ability of the fitted model.

3.3.1 Applications - Chiburis catalog

The estimation and validation subsets should be defined so as to replicate as closely as possible the features of the actual data and of the events to be predicted. In the analysis of Northeastern U.S. seismicity, the Chiburis catalog was divided into ten intervals with nearly equal numbers of recorded main events and the last five intervals were used as validation subsets (V_i, i = 6,...,10) (Table 3.I). For the more recent time intervals, this corresponds to validation periods of approximately 15 years. Other partitions were also investigated, with 2,3,5,20 intervals. The optimal penalties obtained for the last 10 of 20 intervals were the same as for the last 5 of 10 intervals. The optimal penalties obtained from the longer validation intervals (catalog partitioned into 2,3 or 5 intervals) resulted in larger optimal penalties partly because of the decrease in the sample size of the estimation data sets E_i and partly because of migration of seismicity (see below). Note that for the purpose of computing the cross-validated scores, the probability of detection has been fixed to estimates previously obtained by using the entire catalog.

The first step in the specification of the model is the selection of the grid size for the estimation of the seismicity parameters, which directly affects the degree of smoothness attainable by the estimates. The effect of the grid-size was investigated for the subregion identified in Figure 3.11a for $(1.0)^2$, $(0.5)^2$ and $(0.25)^2$ square degree cells. All scores are computed with respect to quarter degree cells. For example, the

seismicity parameters which are estimated for a one square degree cell are assigned to each of its 16 quarter degree subcells for the computation of the scores. It has been found that, for the region under study, there is a significant gain in prediction accuracy when going from one-degree to half-degree cells, but that no additional gain results from using quarter-degree cells (Figure 3.11b). This may be a consequence of the fact that many events in the catalog are located with an accuracy not higher than one quarter degree.

Starting with a fixed neighborhood, a given value of P_a ($P_a = 7$, which turns out to be the optimum value), and a discretization into half-degree cells, optimization was first performed with respect to P_b (Figure 3.12). The large optimal penalty (P_b=1000) is a consequence of the inaccurate estimation of b using data from only very few cells. Figure 3.12 indicates that it is best to use a high penalty P_b and introduce bias into $\hat{b}(\underline{x})$, in order to reduce the large estimation variance. With P_b fixed to 1000, the optimal penalty for $a(\underline{x})$ has been determined and found to be low (P_a=7), meaning that this parameter is best estimated locally; see Figure 3.13.

A higher optimal penalty, around 15, is found when using a cross-validated squared-error criterion; see Figure 3.13b. In this case, one penalizes quadratically the deviations of the actual counts $N(\underline{x},I,t)$. Notice that the log-likelihood penalizes the same deviations, but in a way that depends on $\lambda(\underline{x},I,t)$ (higher penalty for lower expected counts). The reason for the increased optimal penalty for the squared error is that this quantity is more sensitive than the log-likelihood to large deviations of the actual counts from the expected counts. These deviations are reduced by using higher smoothing. The only combinations to which the log-likelihood is very sensitive (nearly zero expected counts and large actual counts) do not occur in the data.

The scores of Equation 3.9 can be decomposed in space, time, and earthquake size to investigate the features of the process in more detail. Figure 3.14 shows the

decomposition of the cross-validated log-likelihood for different intensity intervals. From these we can observe that the main contribution to the total scores is from the smaller intensity interval which contains the largest number of events. The optimal penalty is the same for the first two intensity intervals (P_a=7) indicating that events in these two intervals have similar spatial distributions. For the third interval ($5.5 \leq I < 6.5$), there is a local maximum at the optimal penalty of the first two intervals and an overall maximum at larger penalties. The presence of the two maxima is indicative of two trends in the data set. The first one indicates that part of the observations has a distribution similar to the first two intervals. The other indicates that some of the observations occur in unexpected areas, which in this case is the cluster of events in the south west corner of the region. Similar remarks can be made with respect to the last two intervals. However, in the latter cases, the scores are based on very few events and their uncertainty is large.

The clustering of events in the validation data sets raises the issue of lack-of-fit which was also raised in the previous section with respect to the spatial distribution of events for the first five (1627-1815) and the last five (1915-1981) time intervals of the catalog. Figure 3.15a shows the distribution of the events for the last five time intervals . The optimal penalties for the individual time intervals vary between 3 and 10 which is a range that can be expected given the variability due to the small sample sizes (Figure 3.15b). In that case, the cross-validated likelihood is obtained for each interval by summing only with respect to location and intensity in Eq. 3.9. For the last time interval, there are two maxima, the first is within the range of the optimal penalties for the other time intrevals while the second is dûe to an unexpected increase of activity in southeastern Quebec.

Lack-of-fit of the model in space, time and magnitude can be investigated using various procedures. A simple one is to flag significant residuals by comparing the

number of observations to the number of predicted observations for the different partitions of the catalog. For the purpose of comparing $N(\underline{x},t)$ and $\hat{N}(\underline{x},t)$, counts in each time interval are aggregated over regions (provinces) of homogeneous seismicity proposed by the Weston Geophysical Company (Figure 3.5). Confidence intervals at the 2, 5, and 10% significance levels are indicated and the validation subsets are numbered from 1 to 5 (Figure 3.16). Significant underpredictions of seismicity are identified in the Valley and Ridge, Piedmont Atlantic Coastal Gravity, Adirondack Uplift, and Merrimack Synclinorium provinces. Note that for a Poisson process, underpredictions of a given magnitude are more significant than overpredictions of similar magnitude (Fig. 3.10).

All previous results are for interpolation neighborhoods of fixed geometry and size (section 2.2). An undesirable feature of the estimates is that the boundaries between highly active and less active areas, which should appear as sharp discontinuities of $\hat{a}(\underline{x})$, are blurred. A simple corrective procedure could be to vary the penalty (P_a) as a function of location ($P_a(\underline{x})$) or as a function of the total number of observations in each cell $P_a(N(\underline{x}))$. Less smoothing should be required where there are many observations and more smoothing where observations are sparse. However, Figure 3.17 shows that there is no clear pattern in the optimal penalties as a function of location or as a function of the total munber of observations in a cell. Note that the optimal smoothing in these figures is for each individual cell and that it does not take into consideration how these penalties affect the estimates at the neighboring locations.

A better procedure to allow differential smoothing across the region is to use local-neighborhoods as described in section 2.3. In order to compare fixed-neighborhood with local-neighborhood estimators on the basis of L_{cv} in Equation 3.9, one should cross-validate the local neighborhoods. One can do so in two different ways: For each validation interval t, one can estimate the local neighborhoods 1. using only the data set

E_t, or 2. using all the data with V_t removed. If P_a is kept to 7, the cross-validated likelihoods for the two options are respectively -774 and -745. The value L_{cv} for the case with fixed neighborhoods is -760, as shown in Figure 3.13. These results indicate that accurate estimation of the local neighborhoods requires large amounts of data, hence the option 1 may not be representative of the accuracy achievable at the present time. Option 2 gives a more realistic evaluation and shows improvement over the analysis with fixed neighborhoods.

As one would expect, a decomposition of the cross-validated likelihood in space indicates that, in regions of pronounced seismicity gradients, the likelihood increases with increasing the significance level α for the identification of the local neighborhoods. The opposite is true in areas where the long-term seismicity appears homogeneous, although the earthquake pattern has changed, sometimes significantly, over shorter intervals of time. One way to further improve the local-neighborhood solution is to allow α to vary as a function of location. Analyses of this type were made, limiting the choice of $\alpha(\underline{x})$ to just two values: the value 0, which corresponds to a neighborhood of fixed geometry, and the value 0.15, which produces neighborhoods of homogeneous cells. The cross-validated likelihood of each cell was calculated for both α=0 and α=0.15 and the value of $\alpha(\underline{x})$ was fixed to 0 or to 0.15 if the local likelihood in one solution was larger than the same likelihood in the other solution by more than a given factor (10% in this case); see unshaded and heavily shaded cells in Figure 3.18a. For other cells, two cases have been considered, one favoring the fixed neighborhoods (α=0), the other favoring the local neighborhoods (α=0.15).

The estimates $\hat{a}(\underline{x})$ that result from the two analyses are displayed in Figure 3.18b. Except in the Southeastern corner (New Jersey, Eastern Pennsylvania, and Northeastern Maryland), the contour lines of $\hat{a}$ are almost the same in the two cases. The reason is that, for most of the cells that are indifferent to setting α equal to 0 or

0.15, the local and fixed neighborhoods coincide or are very similar. Because keeping α fixed is a special case of letting α vary with $\underline{x}$, one cannot compare the estimators of Figure 3.18 with those of Figure 2.7(b) in terms of their cross-validated likelihood. It is however clear that the estimates $\hat{a}(\underline{x})$ are not much different in the two cases and hence that, for the region under study, there is little incentive to use the more complicated estimator with variable α.

Another modification of the estimators of Figure 2.7 that is considered consists of finding $\hat{a}(\underline{x})$ and $\hat{b}(\underline{x})$ from only the more recent part of the catalog. Doing so should produce better predictions if the earthquake process has memory or is nonstationary, so that seismicity in the near future should resemble more the recent past than the average seismicity during long periods of time. This idea was implemented by including in the estimation subsets E_i only the two time intervals that precede V_i. The estimate of $\hat{a}(\underline{x})=ln\lambda(4)$ from the last two periods (1957-1981) is shown in Figure 3.19b and is quite different from estimates that use the entire ctalog, e.g. the estimate of Figure 2.7, which is reproduced as Figure 3.19a.

The optimum penalty P_a when using only the more recent data is around 10, and this is the value used in Figure 3.19a. The penalty P_b and the prior on $b(\underline{x})$ are the same as for Figure 3.19b. Because of the reduced amount of data, the estimates of $a(\underline{x})$ and $b(\underline{x})$ based only on recent seismicity are smoother (b is almost flat over the entire region, with values between 1.23 and 1.29). Other differences between the estimates of $a(\underline{x})$ in Figures 3.19a and 3.19b are that, in the former, earthquake activity is higher in New Jersey and lower in Eastern Massachusetts and Southern New Hampshire.

The cross-validated likelihood is nearly the same for the two analyses. This is probably the net effect, in the case when only recent data are used, of an increase in prediction accuracy due to the higher similarity of seismicity and a decrease in prediction accuracy from the smaller estimation samples. In order to evaluate how the

differences in $\hat{a}(\underline{x})$ and $\hat{b}(\underline{x})$ affect the recurrence of large events, Figure 3.19c and 3.19d show contour plots of $ln\lambda(8)= \hat{a}-4\hat{b}$. It is interesting, but probably fortuitous, that the differences in a and b in the two analyses have compensating effects, so that the estimates of $ln\lambda(8)$ are more similar than the estimates of $ln\lambda(4)$. The main differences for earthquakes of MM intensity 8 are that, when only the recent data are used, the estimated rate is smoother over the entire region and is higher (by a factor of about 2) in the New Jersey area. Because $\hat{a}(\underline{x})$ is sensitive to the portion of the catalog used for the estimation and the compensation of $\hat{a}$ and $\hat{b}$ for high intensities is of suspect generality, it is concluded that one should consider seismicity estimators that are local in time, especially when their cross-validated likelihood is high.

3.3.2 EPRI catalog

In this section, the cross-validation procedure is applied to a much larger region which allows a better analysis of the spatial distribution of seismicity and phenomenas such as burst and migrations of activity. This gives rise to some difficulties, mainly in the definition of validation samples because of the large differences in the incompleteness as a function of location.

The region which is analyzed covers latitudes 25° N to 52° N and longitudes 60° W to 90° W. Nova Scotia is purposely left out because of its short historical record and the difficulty in defining a validation interval for it. Events that are used for estimation or validation have magnitude greater than 3.3 and are within the time-magnitude envelopes of Table 2.II which limits the data set to the most complete portions of the historical record for each incompleteness region (see Figure 2.8).

In this and the following applications, the interpolators $\tilde{a}(\underline{x})$ and $\tilde{b}(\underline{x})$ used in Eq. 2.15 are the averages of a and b over the eight cells that are closest to $\underline{x}$ unless otherwise specified, and the probability of detection is set to the estimates in EPRI (1985) (Table 2.II).

The discretization of space is into one degree cells. The variation of the scores as a function of P_b is investigated for 4 levels of smoothing (P_b=10,100,1000,10000) where the largest of these penalties results in nearly uniform estimates of b($\underline{x}$), and the range of variation of P_a is from 1 to 100. Only one validation data set is defined because of the short history of reporting across most of the region (Table 2.II). Three different partitions of the catalog into an estimation and a validation set are considered : In the first two partitions, the prediction set includes all the events in the last 15 or 30 years of the catalog (Figure 3.20a,b). These periods correspond to fairly complete portions of the catalog and to typical prediction horizons in seismic hazard analysis. The number of events in the resulting partitions are shown in Figure 3.21 (see also Table 3.III). Note that these events correspond to equivalent periods of observation T($\underline{x}$,m) which vary both in space and in magnitude because of incompleteness (Table 2.II(b)).

Another possibility is to partition the catalog according to the equivalent period of observation for each location and magnitude interval. The estimation data sets is then defined such that it covers the first α% of the total equivalent period of observation T($\underline{x}$,m) and the validation data set, the rest. The advantage of such a partition is that the proportion of events of different magnitude and at different locations between the validation and estimation subsets is preserved. In particular, this choice eliminates instances where the estimation subset may be smaller than the validation subset (locations with short histories of reporting, e.g. incompleteness regions 8,9,11 in Table 2.II(b)). This partition produces unequal periods of observation as a function of space and magnitude which must be kept in mind when later interpreting the cross-validation results.

Figure 3.20c shows the partition obtained when the percentage α is set to 67% of the equivalent period of observation T($\underline{x}$,m) (in the following, this will be referred to as the (2/3,1/3) partition of the catalog). The unmarked regions correspond to incomplete

portions of the catalog which were not used in the analysis. The events corresponding to this partition are shown in Figure 3.21c.

3.3.2.1 The Catalog

The seismicity of the area has been the subject of several investigations, some of which do not support the assumptions of stationarity in time and space and exponentiality in size distribution. The following is a summary of the comments from previous studies.

(Mitronovas, 1981) suggests that the activity in the northeastern U.S. during the past 300 years shows secular variations lasting up to 100 years. In particular, he reports periods of greater activity between 1720-1790, 1830-1880 and 1910 through the present in the state of New York, with the local activity alternating between subregions within the state. (Armbruster and Seeber, 1987) note that the pattern of seismicity derived from recent short term instrumental data ressembles in general the pattern of seismicity derived from long-term samples of historic data, but acknowledge that some changes in the temporal pattern of seismicity can result after large events such as the 1886 Charleston S.C. earthquake. (Ebel, 1987) observes that the mean rate of earthquake occurences in the northeast U.S. as a whole has been approximately stable with time, however, his observations are based only on the similarity of the a-values for the whole region during the periods 1938-1986 and 1975-1986. Variations in the rate of occurence of small and large magnitude events in the eastern U.S. have also been noted by other authors (Chiburis, 1981) (Shakal and Toksoz, 1977) (Veneziano and VanDyck, 1987).

The Chinese earthquake catalog (3000 years of observations) provides support for similar non-stationarities in intraplate regions : (McGuire, 1979) observes temporal periodicities in the order of hundreds of years in which rates of activity changed by as much as a factor of 10. However, despite these non-stationarities, predictions using a Poisson model were adequate when based on observations immediately preceeding the interval.

The presence of possible non-stationarities in the EPRI catalog can be illustrated by comparing the spatial density of events between the (2/3) and (1/3) partitions of the catalog (Figure 3.21c). Under the assumption that seismicity is stationarity and that the probability of detection is correct, the density of events should be similar spatially and in a proportion of 2 to 1 between the two subsets. However, this is not the case in the Ottawa River Valley, Eastern Tennessee and Lower St-Lawrence Valley areas where the proportion of events in the recent past is larger than expected. Conversely, in the recent past there are fewer events than expected in the Boston and Western Tennessee areas. The types of non-stationarity and their effect on the selection of optimal penalties are examined in more detail in the following section.

3.3.2.2 Discussion of the results

The issues addressed in this Chapter are divided into 3 main groups. These are : the selection of the optimal models, the influence of the statistic used in the cross-validation, and the goodness-of-fit of the predictions.

• Selection of Optimal Models

The cross-validated scores are computed for three different magnitude discretizations of the validation data set. The optimal penalties are calculated for each discretization and separately for each magnitude range. The first discretization is into 7 intervals of width 0.6 from 3.3 to 7.5. This is also the discretization used for the estimation of the parameters a and b. The second discretization is into three intervals (m = 3.3-3.9, 3.9-4.5, 4.5-7.5), to increase the sample size for the larger events. The third case is to use only one interval (the total number of events in each cell) and disregard the distribution in magnitude of the events. The discretizations of magnitude into 7, 3 and 1 intervals result in similar optimal penalties for all three validation subsets (Figure 3.22).

The optimal penalties are respectively small and large for P_a and P_b. The small penalty on $a(\underline{x})$ indicates that most predicted events are located in historically active areas. As in the previous application, the optimal penalty on $b(\underline{x})$ is large and corresponds to almost constant values. These results confirm the uniformity of b-value obtained by (Chinnery, 1979) in his study of frequency-MMI intensity data from the southeastern U.S., central Mississipi Valley and Southern New England. However, the effect of the penalty P_b is small compared to the effect due to P_a, specially when considering the uncertainty on the cross-validated likelihood statistics (see below). Notice that the effect of P_b is almost nil when the cross-validated statistics are computed for the total number of observations (the recurrence rate is dominated by $a(\underline{x})$.

The optimal penalty on $a(\underline{x})$ for predicting the events during the last 15 years of the catalog (Figure 3.22a) are slightly smaller than the optimal penalties for predicting events during the last 30 years (Figure 3.22b) or the last 1/3 portion of the catalog (Figure 3.22c) because of the larger estimation sample size for the 15 years partition. In all cases, the optimal penalty P_a is slightly larger for SE_{cv} than for L_{cv} for reasons explained in the previous section. The scores for the low magnitude events ($3.3 \leq m < 3.9$) dominate the total scores and are, not surprisingly, small for $a(\underline{x})$ and large for $b(\underline{x})$ (Figure 3.23). For the events of intermediate magnitude ($3.9 \leq m < 4.5$), either a small penalty on $b(\underline{x})$ and a moderate penalty on $a(\underline{x})$ or a large penalty on $b(\underline{x})$ with a small penalty on $a(\underline{x})$ are optimal (Figure 3.24). The two cases result in similar recurrence rates for this magnitude interval, because estimates of $b(\underline{x})$ are negatively correlated with the level of activity (section 2.2). The number of large magnitude events ($4.5 \leq m < 7.5$) in the validation subsets is very small and consists mainly of events with magnitudes 4.5 to 5.1 (Table 3.III). For the last 15 years of data and the (2/3,1/3) partition, the optimal penalties are similar as those identified for the previous 2 magnitude intervals (Fig. 3.25). The penalty P_b, however, has more influence on the

cross-validated statistics. For these two cases, the optimal penalties are very large for P_b. For the last 30 years of observations (Table 3.III), the optimal penalties are respectively large for P_a and small for P_b due to a cluster of events in western Tennessee. Notice that there are fewer events than expected in the Boston and Charlevoix areas while there are more events than expected in the Lower St-Lawrence Valley. The latter underprediction results from the short period of reporting in the St-Lawrence Valley where most of the reported events occured during the last 30 years of the catalog (Figure 3.21b). The validation subset for the (2/3,1/3) partition of the catalog corrects for this imbalance by increasing the period of observation for the validation subset in regions with long historical records (such as the Boston and Charlevoix regions) and decreasing it where the historical record is short (such as the Lower St-Lawrence Valley). In addition, the (2/3,1/3) partition increases the sample size for the larger magnitudes.

In conclusion, the optimal penalties appear to be insensitive to the discretization in magnitude, because of the dominance by the low magnitude events. They are respectively low for P_a and high for P_b. These penalties appear to be optimal for all ranges of magnitude. The issues of non-stationarity and non-exponentiality and their effect on the estimation of an optimal model are further addressed in the following section.

All previous results are for interpolation neighborhoods of fixed geometry and size. Various plots of $\hat{a}(\underline{x})$ are presented in Figure 3.29 to show the effect of changing the significance level α which defines the degree of homogeneity of the local neighborhoods. and the penalty P_a. For α=10%, the function $\hat{a}(\underline{x})$ displays plateaus of nearly constant activity, in some cases separated by sharp discontinuities, in other cases connected by gradual ramps. The local neighborhoods have no influence where the seismicity is uniform or where the data is too sparse to identify contrasts of seismicity.

The estimates are less sensitive with respect to P_a and preserve a high level of contrast for a large penalty P_a. Significant boundaries of activity are identified along the Atlantic coast, around Charleston, Western Tennessee, Boston and Charlevoix.

Estimates are obtained with α=10% for predicting the events during the last 30 years and the last 1/3 partition of the catalog. In this application, the local neighborhoods do not perform as well as anticipated because the patterns of events in the estimation and validation data sets are significantly different. For example, the introduction of the local neighborhoods in regions which have been historically more active than presently, such as the Charlevoix and Massachusetts regions, typically increases the magnitude of the predictions. Similarly, predictions in areas which have been quiescent in the past and more active recently improve with increased smoothing.

• L_{cv} versus SE_{cv}

L_{cv} and SE_{cv} identify similar optimal penalties, SE_{cv} tending to select penalties slightly larger than L_{cv} (Figure 3.22). As in the application to the Chiburis catalog, the reason for the increased optimal penalty for the squared error is that this quantity is more sensitive than L_{cv} to large deviations of the actual counts. These large deviations are reduced by using higher smoothing. The only combinations to which L_{cv} is very sensitive (nearly zero expected counts and large actual counts) do not occur in the data. The decomposition of the squared error in space shows that large contributions correspond to significant overpredictions or underpredictions (<,> in terms of flags), which is limited in general to a few cells, while contributions to the log-likelihood tend to be more uniform over the region (note that overpredictions are not as heavily weighted as underpredictions for the log-likelihood). Examining the sequence of Figure 3.26, shows that the variation of the statistics as a function of P_a is dominated by the most active cells of the validation subset and that many deviations between observations and predictions remain large also at the optimal penalty. This brings up the issue of goodnees-of-fit which is addressed next.

• Goodness-of-fit

For goodness-of-fit, two basic procedures are used. The first is based on the flagging of significant residuals and the other is based on the probability distribution function of the cross-validated likelihood. These can be used, when applied to given discretizations in space and magnitude, to test the assumptions of a stationary Poisson process and exponentiality in size distribution. The spatial goodness-of-fit of the model in magnitude is checked through tests that compare the observed and expected mean magnitude of events in each cell. Under the null hypothesis, the events in a cell of the complete catalog are exponentially distributed in magnitude with parameter b($\underline{x}$). For the incomplete catalog the events are similarly distributed with a correction to account for the probability of detection. For the generic event recorded at location $\underline{x}$, the magnitude has probability mass function

$$p(m) = \frac{P_D(\underline{x},m)e^{-b(\underline{x})m}}{\sum_{m_1}^{m_n} P_D(\underline{x},m)e^{-b(\underline{x})m}} \qquad m = m_1,...m_n \tag{3.11}$$

where

$$P_D(\underline{x},m) = \sum_{1}^{N_t} \frac{P_D(\underline{x},m,t)\bullet T(\underline{x},m)_t}{\sum_{1}^{N_t} T(\underline{x},m)_t}$$

where $T(\underline{x},m)_t$ is the lenght of the t^{th} period of observation for magnitude m at location $\underline{x}$ and N_t is the number of observation periods.

For two or more events, the distribution of the average (incomplete) magnitude can be obtained through successive convolutions of the previous probability mass function and is therefore tedious to calculate. However, a test based on a normal approximation using the mean and variance for N($\underline{x}$) independent observations from the distribution in Equation 3.11 can be used. In this application, the power of the test is weak because of the small number of observations in each cell. Eliminating empty cells of the validation subset leaves numerous cells with single events in historically active areas which reject the hypothesis of a constamt b($\underline{x}$). (Figure 3.27b).

Next, the goodness-of-fit with respect to the number of events predicted in various partitions of the catalog is checked using the flagging of significant residuals (section 3.2.4). We first examine predictions of the total number of events regardless of magnitude and location, and proceed with partitions of the data set in magnitude and in space. The first test concerns global non-stationarities with respect to the whole region, the second is a test of non-exponentiality for all the events in each magnitude interval and the last test is a test of the spatial distribution of events. It is found that the total number of events is overpredicted in all the validation subsets, indicating that the rate of activity over the EUS has decreased significantly in recent times (Table 3.III). The most significant deviations are for the last 15 years of data and for the validation period covered by the (2/3,1/3) split. The total number of events for the last 30 years is not as significantly overpredicted implying that the number of events in the first period of 15 years is underpredicted by an amount similar to the overpredictions of the second period of 15 years. With respect to the magnitude distribution irrespective of location, the largest deviations from the assumption of exponentiality are detected for events with size $3.9 \leq m \leq 4.5$ (Table 3.III). The fit of the model with respect to $a(\underline{x})$ is also checked spatially by performing the test with respect to the predictions of the number of events in each spatial cell. The number of flags as well as their spatial distribution can be used to judge the goodness-of-fit of the predictions. Relative to the historical record, there is a decrease in the rate of events in the Boston and Western Tennessee areas and an increase in the rate of events in the Lower St-Lawrence Valley, Ottawa River Valley, and Eastern Tennessee areas in recent times.

A comparison of the spatial distribution of the flags (Figure 3.27a) with $SE_{cv}(\underline{x})$ (Figure 3.26) shows that the maximum scores correspond almost exclusively with flags at the 2% significance level (< or >) which are extremes of overprediction or underprediction. On the other hand, the maximum scores of $L_{cv}(\underline{x})$ (low likelihoods,

poor predictions) correspond to flags at either the 2% or 10% level with more emphasis on underpredictions (+ or >) than on overpredictions (- or <).

The tests on the spatial distribution are also performed after correcting the total counts for imbalances between predicted and observed counts. The correction is implemented by adding a positive or negative constant Δa to each estimate $\hat{a}(\underline{x})$ such that the total number of observed and predicted events are equal for the whole region. The constant is calculated under the assumption that the imbalance is the result of a unifrom change in the level of activity across the region. The correction is meant to reduce nonstationarity flagging due to lack-of-fit of the model, i.e. it allows one to better separate between anomalies in the time-average activity and anomalies in the spatial distribution of seismicity. With the correction, the number of flags due to overpedictions were reduced but the main features of the pattern of flags remained the same.

The nature and severity of the lack-of-fit of the predictions can also be investigated through the probability distribution function of the cross-validated log-likelihood (L_{cv}). The expectation and standard deviation of L_{cv} are computed through parametric simulations (the number of simulations is 20 in this application) of the number of observations in each spatial cell and magnitude interval of the estimation and validation subsets. An approximate and computationally less tedious expression for estimating the expected value is obtained by taking expectation with respect to the number of observations in each cell of the validation data set and assuming that the estimated model is the true one.

$$E[L_{CV}]= \sum_{\underline{x}}\sum_{m}\sum_{N(\underline{x},m)_V=0}^{\infty} P[N(\underline{x},m)_V \mid \hat{\lambda}'(\underline{x},m)_E] \bullet lnP[N(\underline{x},m)_V \mid \hat{\lambda}'(\underline{x},m)_E] \qquad (3.12)$$

where $P[N(\underline{x},m)_V \mid \hat{\lambda}'(\underline{x},m)_E]$ is the probability of observing $N(\underline{x},m)_V$ events in a cell for a Poisson process with recurrence rate $\hat{\lambda}'(\underline{x},m)_E$, E and V refer respectively to the

estimation and validation sets. A similar expression is obtained for the variance of L_{cv} in each cell ($\underline{x}$,m). The variance on the total score is then simply the summation of the variances for each individual cell assuming independence. The approximation is accurate when the penalty P_a is large but overestimates $E[L_{cv}]$ for small penalties because the expectation does not take into account the finite sample size of the catalog and expectation is not taken with respect to the number of events in the estimation data set.

The procedure finds the model to be acceptable if the average of L_{cv} over many cells is close to $E[L_{cv}]$ (within 1.5 standard deviations for example). It is assumed that there is undersmoothing when L_{cv} is greater than $E[L_{cv}]$ (the fit is too good to be true). Local lack-of-fit caused by regional overpredictions or undepredictions, can be assessed through a spatial decomposition of L_{cv} and $E[L_{cv}]$ which can also be decomposed in magnitude to judge the goodness-of-fit of events over different magnitude ranges. Values of $L_{cv}(m)$ and $E[L_{cv}(m)]$ tend to decrease in absolute value with an increase in magnitude because of the decrease in the number of observations and recurence rate, $\sigma[L_{cv}(m)]$ also decreass but at a slower rate. In consequence, the coefficient of variation on L_{cv} increases with magnitude, smaller data sets and degree of discretization in space, time and magnitude.

Goodness-of-fit tests through simulation are performed only for the (2/3,1/3) partition of the catalog and P_b=10000. Figure 3.28 shows L_{cv}, $E[L_{cv}]$ and the one standard deviation envelopes obtained both through simulation and through the previous approximate expression. When L_{cv} is not discretized in magnitude, the lack-of-fit with respect to the spatial distribution of the total counts is emphasized. When the magnitude is discretized, the fit with respect to the distribution in magnitude is stressed, and the fit with respect to the total counts in not as important. In this application, the model is not rejected when the scores are computed on the basis of the

total number of events despite large shifts in the spatial distribution of events between the estimation and validation subsets (Figure 3.28b). When magnitude is discretized, the lack-of-fit is slightly more severe, however, L_{cv} is within acceptable limits at the optimal penalties (Figure 3.28a,c). For the low magnitude events, lack-of-fit is largest near the optimal penalty (Figure 3.28d). For the intermediate magnitude events, the models are accepted for the full range of penalties P_a. Results for large magnitude events indicate that the statistics for any penalty are within their expected range (Figure 3.28f).

A decomposition of the results in space indicates that the most serious lack-of-fit occurs for cells with underpredictions. A close examination of the results shows that many of the deviations are the result of local shifts of seismicity in time (e.g. Charleston, the Ottawa River Valley and Eastern Tennessee), which have a minor effect on the seismic risk when seismicity is integrated over the whole region. More serious deviations which cannot be explained through a local accounting of events occur in Western Tennessee, the lower St-Lawrence Valley, and the Ottawa River Valley areas. Note the large local residuals may be partly a consequence of large changes in the probability of detection as a function of location.

3.4 Conclusion

In this section, it was shown that cross-validation is a good procedure for the simultaneous selection of model parameters such as, the penalties on $a(\underline{x})$ and $b(\underline{x})$, the grid size, the local neighborhood characteristics and the time-magnitude region for estimation. The method is appealing for seismic hazard applications because it emphasizss the predictive ability of the model.

For the Northeastern U.S., we find that half-degree cells give an appropriate geographical discretization. The optimal degree of smoothing for $b(\underline{x})$ is high

reflecting the low accuracy with which this parameter is estimated from small samples. By contrast, the optimal estimator of $a(\underline{x})$ is highly variable and closely follows the pattern of historical seismicity. With respect to the latter, it is recommended to use the local neighborhood estimator in the abscence of strong physical evidence on the location of major discontinuities. No advantage was found from allowing the parameter α that controls the homogeneity of the local neighborhoods to vary on the geographical plane. However, it is recommended that alternative estimators be considered, which use different portions of the historical record. Doing so is especially important in regions where nonstationarities of the earthquake process have been observed or are suspected to exist.

Methods to investigate the goodness-of-fit of seismicity models have been developed. It is found that the spatial distribution of the total number of predicted events may significantly differ from the historical distributions of seismicity for short time intervals, but has almost no influence on the selection of the optimal smoothing parameters.

Number of intervals	interval	from	to
2	1	1627	1915
	2	1915	1981
3	1	1627	1883
	2	1883	1939
	3	1939	1981
5	1	1627	1847
	2	1847	1897
	3	1897	1930
	4	1930	1957
	5	1957	1981
10	1	1627	1790
	2	1790	1847
	3	1847	1876
	4	1876	1897
	5	1897	1915
	6	1915	1930
	7	1930	1944
	8	1944	1957
	9	1957	1969
	10	1969	1981
20	1	1627	1720
	2	1720	1790
	3	1790	1819
	4	1819	1847
	5	1847	1866
	6	1866	1876
	7	1876	1887
	8	1887	1897
	9	1897	1908
	10	1908	1915
	11	1915	1923
	12	1923	1930
	13	1930	1937
	14	1937	1944
	15	1944	1951
	16	1951	1957
	17	1957	1963
	18	1963	1969
	19	1969	1975
	20	1975	1981

Table 3-I: Decomposition of the Chiburis catalog into intervals containing approximately identical numbers of events.

	1915-1930		1930-1944		1944-1957		1957-1969		1969-1981	
I	obs	exp	obs	exp	obs	exp	obs	exp	obs	exp
3.5-4.5	32	28.2	28	28.5	35	29.2	21	31.1	25	31.1
4.5-5.5	12	8.	10	8.5	15	8.2	20	8.2	11	8.2
5.5-6.5	3	2.0	1	2.2	2	2.0	3	2.1	6	2.1
6.5-7.5	0	0.6	2	0.5	1	0.5	0	0.4	0	0.4
7.5-8.5	0	0.1	0	0.1	1	0.1	0	0.1	0	0.1
	47	39.3	41	39.7	54	39.9	44	41.8	42	41.8

I	obs	exp
3.5-4.5	141	148.1
4.5-5.5	68	41.6
5.5-6.5	12	10.4
6.5-7.5	3	2.4
7.5-8.5	1	0

Table 3-II: Number of observed and expected events as a function of the validation interval and intensity for the Chiburis catalog.

15 years interval -121-

Estimation data set				Validation data set			
mag	obs	exp		mag	obs	exp	
1	630	667.98	-	1	129	179.6	-
2	356	299.64	+	2	28	50.88	-
3	115	115.18		3	7	13.2	-
4	25	41.04	-	4	2	3.43	
5	10	12.79		5	0	0.89	
6	4	3.46		6	0	0.23	
7	1	0.90		7	0	0.06	
	1141	1141			166	248.29	-

30 years interval

Estimation data set				Validation data set			
mag	obs	exp		mag	obs	exp	
1	475	495.92		1	284	310.1	-
2	269	239.18	+	2	115	94.21	+
3	102	97.79		3	20	24.53	
4	23	35.89	-	4	4	6.38	
5	10	11.33		5	0	1.66	
6	4	3.09		6	0	0.43	
7	1	0.80		7	0	0.11	
	884	884			423	437.42	

(2/3,1/3) split

Estimation data set				Validation data set			
mag	obs	exp		mag	obs	exp	
1	529	550.93		1	230	264.40	-
2	256	228.26	+	2	128	109.60	+
3	92	83.23		3	30	40.50	-
4	16	28.71	-	4	12	14.03	
5	7	8.85		5	3	4.34	
6	2	2.40		6	2	1.18	
7	1	0.63		7	0	0.31	
	903	903			404	434.36	-

Table 3-III: Expected and observed number of events in the estimation and validation subsets for the 3 partitions of the catalog.

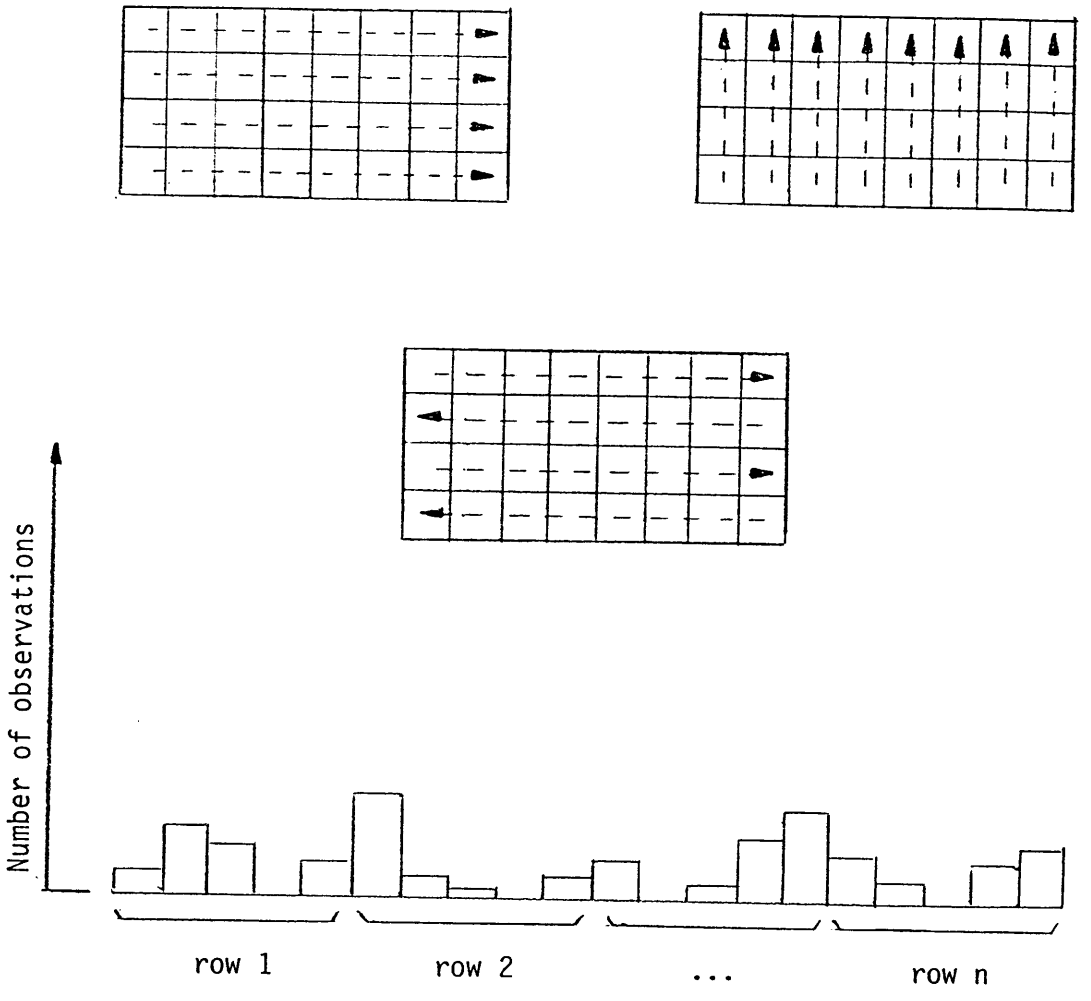


Figure 3-1: Construction of a one-dimensional histogram from a spatial point process.

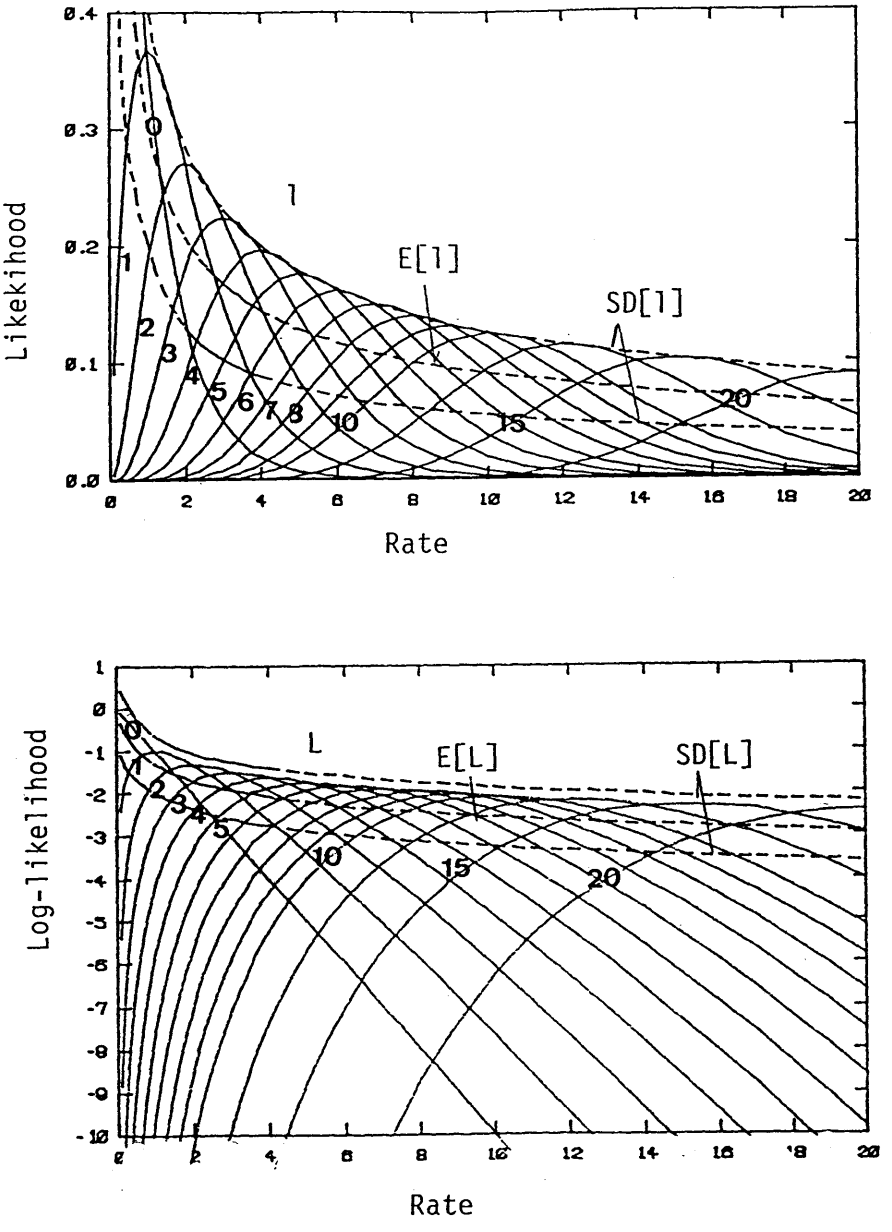


Figure 3-2: (Log)-likelihood and expected (Log)-likelihood, as a function of the recurrence rate and the number of observations for a Poisson process.

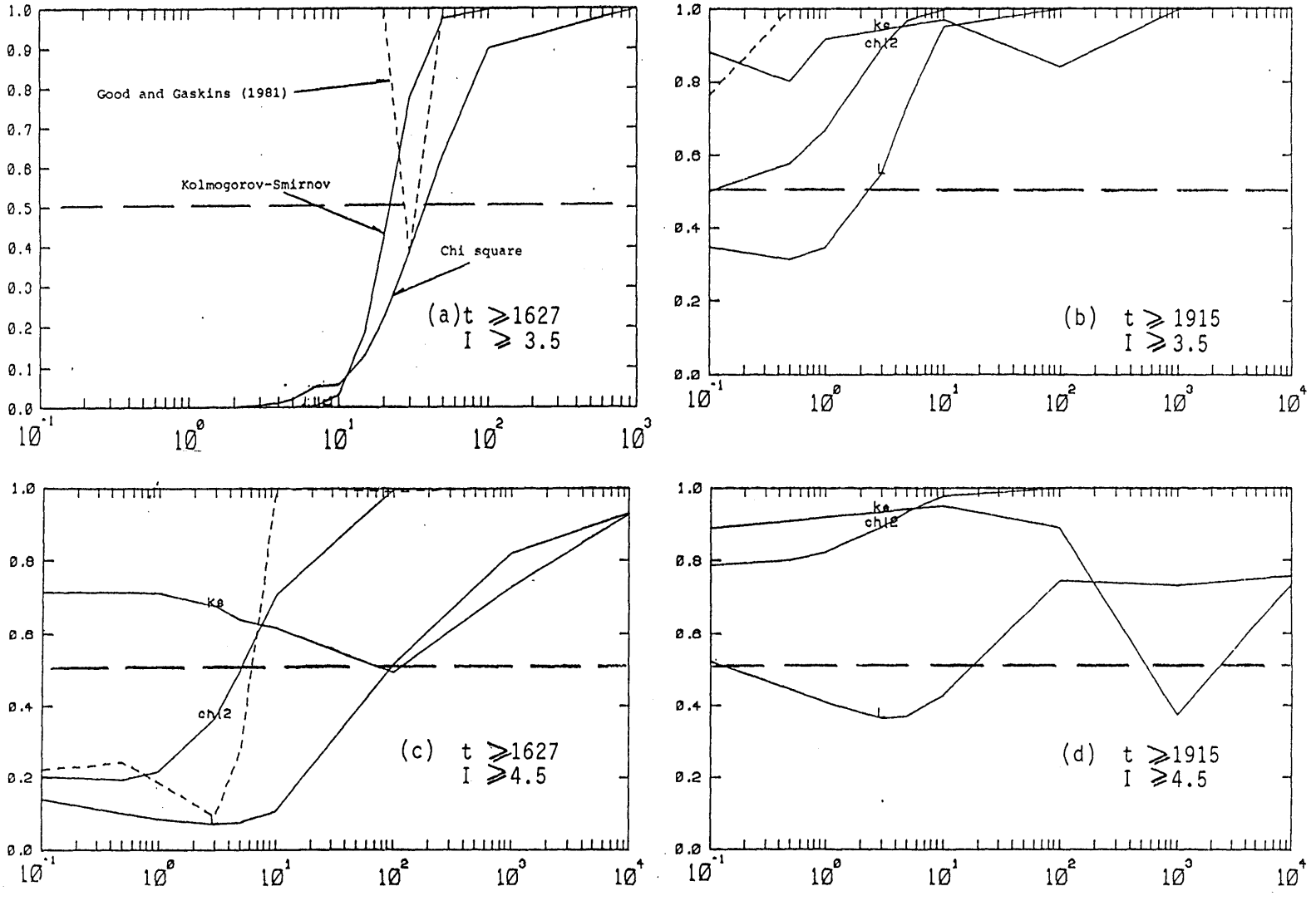
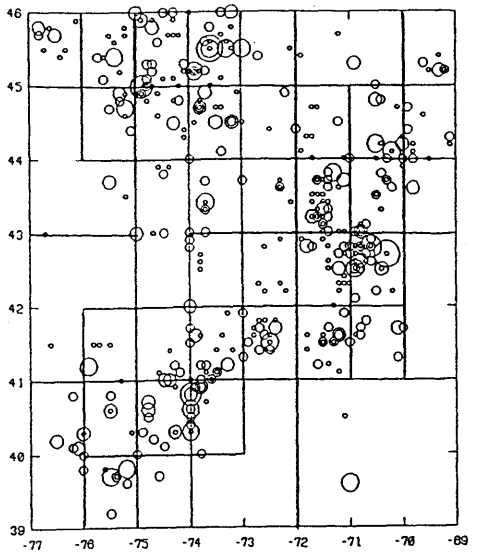
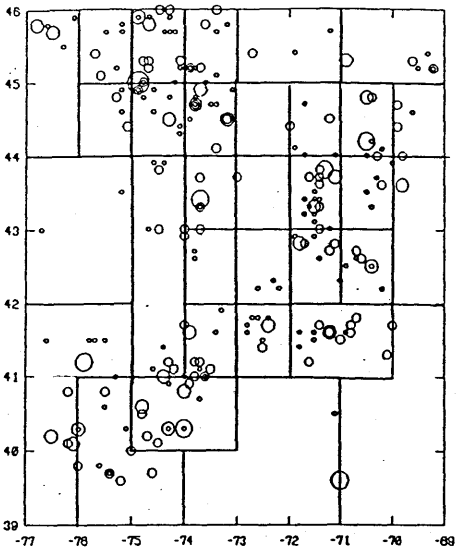


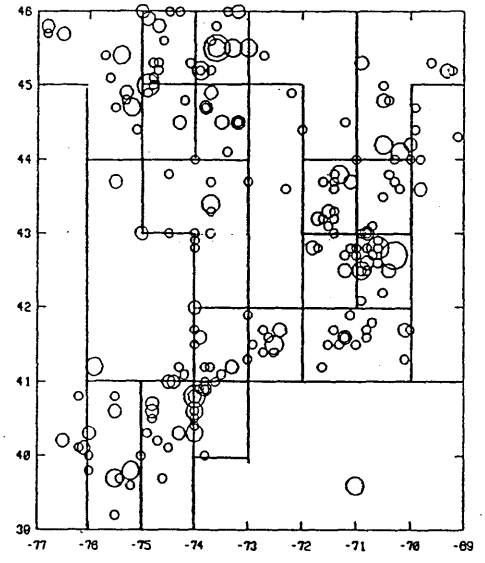
Figure 3-3: Target-statistics procedure for the selection of the optimal P_a over different subsets of the Chiburis catalog.



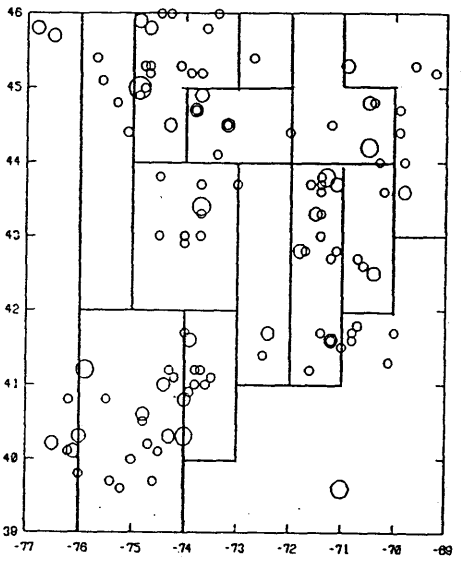
(a) I$\geqslant$3.5, since 1627



(b) I $\geqslant$ 3.5, since 1915

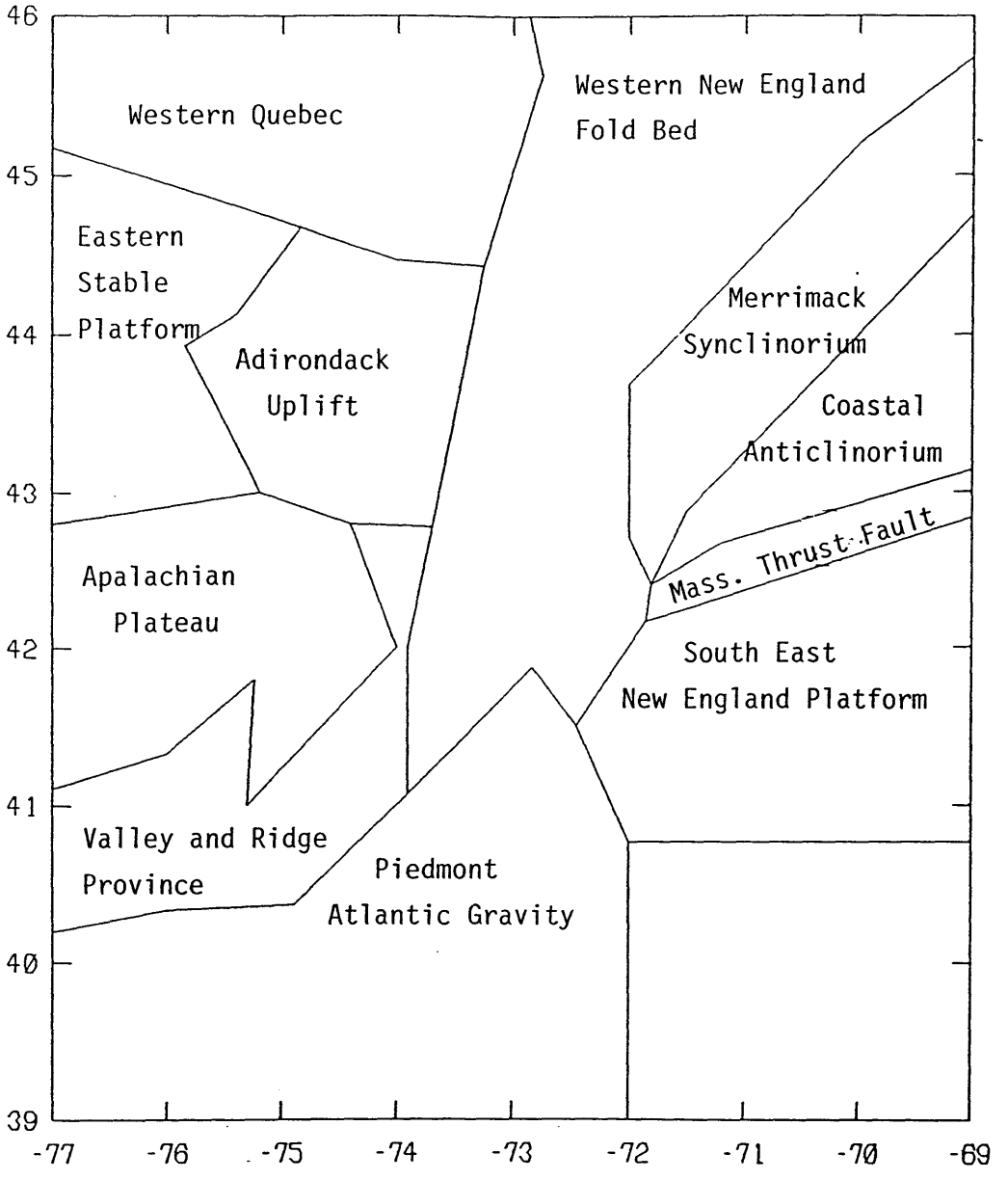


(c) I $\geqslant$4.5, since 1627



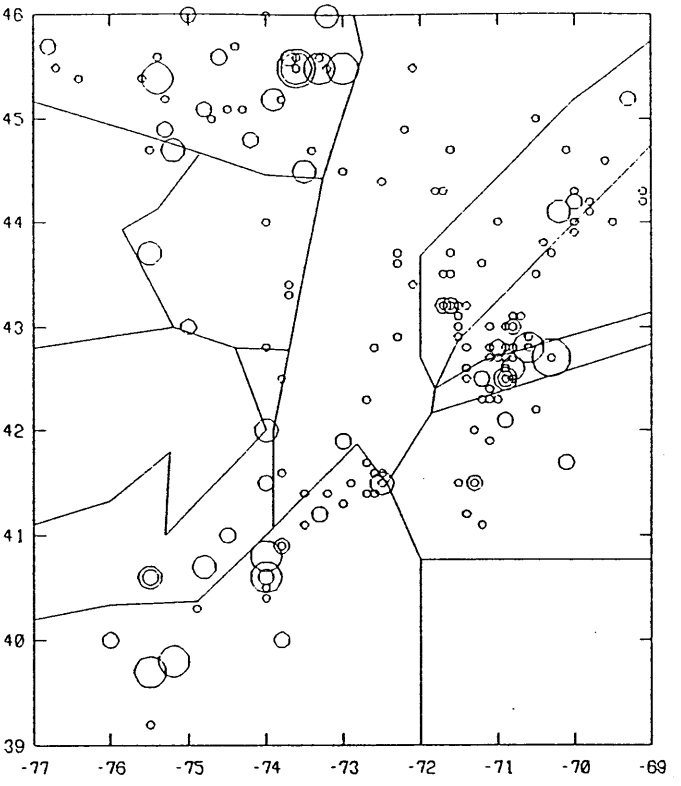
(d) I $\geqslant$4.5, since 1915

Figure 3-4: Aggregated cells for the computation of the Chi-square statistic for different subsets of the Chiburis catalog.



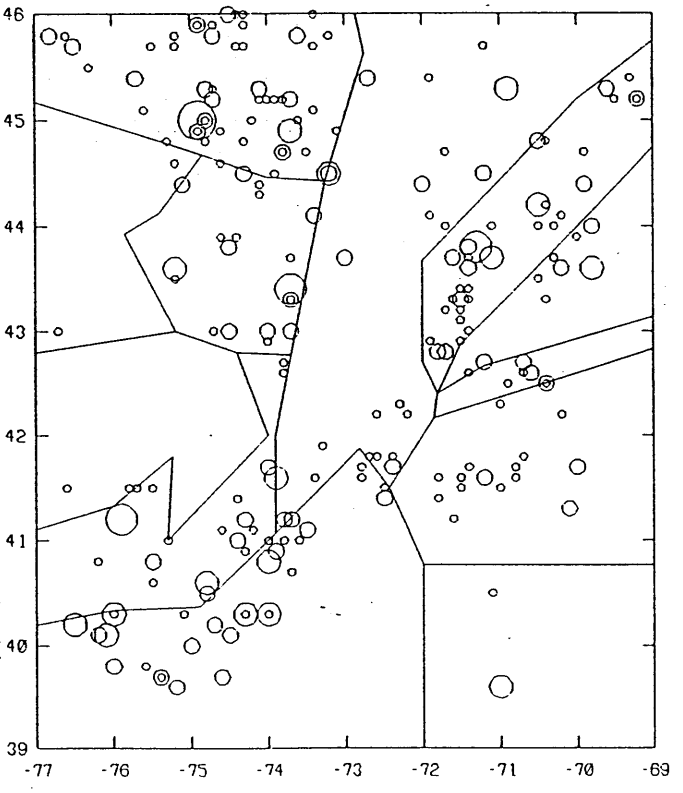
(a)

Figure 3-5: (a) Seismogenic provinces proposed by Weston Geophysical Co. and (b) partition of the Chiburis catalog into two subsets.

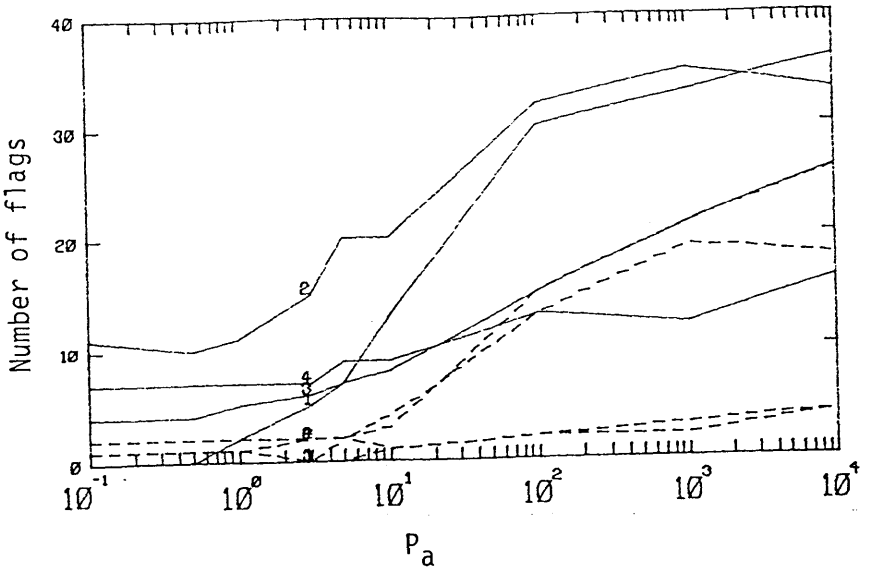


(1627 - 1915)

(b)



(1915 - 1981)



1 : $I \geq 3.5, t \geq 1627$

2 : $I \geq 3.5, t \geq 1915$

3 : $I \geq 4.5, t \geq 1627$

4 : $I \geq 4.5, t \geq 1915$

(_____) : number of flags at the 10% significance level
(_ _ _) : number of flags at the 2% significance level

Figure 3-6: Number of flags for significant residuals as a function of the penalty P_a

P_a=1

Observed cumulative count (mb*>3.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	7.0	9.0	21.0	27.0	3.0	2.0	1.0	6.0
44.0	0.0	7.0	18.0	15.0	3.0	7.0	8.0	9.0
43.0	0.0	2.0	3.0	8.0	5.0	30.0	18.0	7.0
42.0	1.0	0.0	3.0	7.0	8.0	20.0	41.0	0.0
41.0	1.0	4.0	4.0	17.0	22.0	17.0	6.0	1.0
40.0	4.0	9.0	11.0	16.0	0.0	1.0	0.0	0.0
39.0	0.0	9.0	3.0	1.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>3.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	6.4	8.9	21.0	26.0	3.7	2.3	2.0	5.8
44.0	1.1	6.5	17.6	14.5	3.8	6.3	7.7	8.5
43.0	0.6	1.7	3.4	7.8	5.6	29.5	17.9	7.1
42.0	0.9	0.7	3.1	6.9	8.5	20.3	39.6	1.5
41.0	1.4	3.5	4.2	16.6	21.4	16.2	5.4	1.1
40.0	3.3	8.1	10.8	14.7	1.0	1.0	0.4	0.4
39.0	1.7	8.1	3.9	1.3	0.6	0.3	0.4	0.2

Significance test (mb*>3.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0								
44.0								
43.0								
42.0							-	
41.0								
40.0								
39.0	-							

P_a=7

Observed cumulative count (mb*>3.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	7.0	9.0	21.0	27.0	3.0	2.0	1.0	6.0
44.0	0.0	7.0	18.0	15.0	3.0	7.0	8.0	9.0
43.0	0.0	2.0	3.0	8.0	5.0	30.0	18.0	7.0
42.0	1.0	0.0	3.0	7.0	8.0	20.0	41.0	0.0
41.0	1.0	4.0	4.0	17.0	22.0	17.0	6.0	1.0
40.0	4.0	9.0	11.0	16.0	0.0	1.0	0.0	0.0
39.0	0.0	9.0	3.0	1.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>3.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	5.9	9.0	19.7	22.5	5.8	3.6	3.7	5.7
44.0	2.7	6.2	15.7	13.6	6.0	5.9	6.9	7.4
43.0	1.6	2.5	4.6	7.8	7.1	27.5	18.4	7.7
42.0	1.5	1.6	3.6	6.9	9.1	20.8	33.8	3.7
41.0	1.8	2.8	4.3	15.4	19.1	14.3	5.9	2.0
40.0	2.6	6.7	9.7	12.1	2.3	1.6	0.9	0.8
39.0	3.9	6.9	5.7	2.3	1.5	0.9	0.7	0.6

Significance test (mb*>3.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0					-		-	
44.0	-				-			
43.0	-							
42.0		-					+	<
41.0								
40.0					-			
39.0	<				-			

Observed cumulative count (mb*>3.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	7.0	9.0	21.0	27.0	3.0	2.0	1.0	6.0
44.0	0.0	7.0	18.0	15.0	3.0	7.0	8.0	9.0
43.0	0.0	2.0	3.0	8.0	5.0	30.0	18.0	7.0
42.0	1.0	0.0	3.0	7.0	8.0	20.0	41.0	0.0
41.0	1.0	4.0	4.0	17.0	22.0	17.0	6.0	1.0
40.0	4.0	9.0	11.0	16.0	0.0	1.0	0.0	0.0
39.0	0.0	9.0	3.0	1.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>3.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	6.7	8.8	14.0	15.2	8.8	6.4	5.8	6.2
44.0	5.1	7.1	11.7	12.1	8.8	7.2	6.9	6.8
43.0	3.6	4.4	5.3	8.3	8.4	24.3	19.6	7.2
42.0	2.8	3.1	4.5	5.4	7.6	19.6	22.3	5.2
41.0	2.5	2.8	3.8	13.3	15.2	13.2	9.4	3.4
40.0	2.4	6.6	8.5	10.1	3.2	2.7	2.1	2.0
39.0	5.7	6.6	7.1	2.9	2.6	2.1	1.8	1.8

Significance test (mb*>3.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0			+	>	<	-	<	
44.0	<		+		<			
43.0	<	-	-		-			
42.0		<					>	<
41.0					+			-
40.0				+	<		-	-
39.0	<		-	-	-	-		-

(a) P_a=50

Figure 3-7: Number of flags for significant residuals in the spatial cells as a function of the penalty P_a.

P_a=1

Observed cumulative count (mb*>3.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	4.0	5.0	15.0	11.0	1.0	2.0	1.0	5.0
44.0	0.0	4.0	16.0	14.0	0.0	4.0	5.0	2.0
43.0	0.0	1.0	3.0	5.0	1.0	16.0	7.0	3.0
42.0	1.0	0.0	2.0	5.0	4.0	8.0	8.0	0.0
41.0	1.0	4.0	4.0	7.0	9.0	8.0	5.0	1.0
40.0	4.0	7.0	8.0	9.0	0.0	1.0	0.0	0.0
39.0	0.0	5.0	3.0	0.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>3.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	4.2	5.9	13.9	17.2	2.5	1.5	1.4	3.8
44.0	0.7	4.3	11.7	9.6	2.5	4.2	5.1	5.7
43.0	0.4	1.1	2.3	5.2	3.7	12.7	7.7	4.7
42.0	0.6	0.4	2.0	4.6	5.6	8.7	17.0	1.0
41.0	0.9	2.3	2.8	7.1	9.1	6.9	2.3	0.8
40.0	2.2	3.4	4.6	6.2	0.6	0.7	0.3	0.2
39.0	0.7	3.4	1.7	0.9	0.4	0.2	0.3	0.2

Significance test (mb*>3.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0				-				
44.0			+	+	-			-
43.0					-			
42.0							<	
41.0							+	
40.0		+	+					
39.0								

P_a=7

Observed cumulative count (mb*>3.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	4.0	5.0	15.0	11.0	1.0	2.0	1.0	5.0
44.0	0.0	4.0	16.0	14.0	0.0	4.0	5.0	2.0
43.0	0.0	1.0	3.0	5.0	1.0	16.0	7.0	3.0
42.0	1.0	0.0	2.0	5.0	4.0	8.0	8.0	0.0
41.0	1.0	4.0	4.0	7.0	9.0	8.0	5.0	1.0
40.0	4.0	7.0	8.0	9.0	0.0	1.0	0.0	0.0
39.0	0.0	5.0	3.0	0.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>3.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	4.0	6.1	13.3	15.2	3.9	2.4	2.5	3.9
44.0	1.8	4.2	10.6	9.2	4.1	4.0	4.7	5.0
43.0	1.1	1.7	3.1	5.3	4.8	10.9	7.3	5.2
42.0	1.0	1.1	2.4	4.6	6.2	8.2	13.4	2.5
41.0	1.2	1.9	2.9	6.1	7.6	5.7	2.4	1.4
40.0	1.8	2.6	3.8	4.8	1.6	1.1	0.6	0.5
39.0	1.5	2.7	2.2	1.6	1.0	0.6	0.5	0.4

Significance test (mb*>3.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0					-			
44.0	-		+	+	<			-
43.0					-	+		
42.0							-	-
41.0		+					+	
40.0	+	>	+	+	-			
39.0	-	+		-				

P_a=50

Observed cumulative count (mb*>3.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	4.0	5.0	15.0	11.0	1.0	2.0	1.0	5.0
44.0	0.0	4.0	16.0	14.0	0.0	4.0	5.0	2.0
43.0	0.0	1.0	3.0	5.0	1.0	16.0	7.0	3.0
42.0	1.0	0.0	2.0	5.0	4.0	8.0	8.0	0.0
41.0	1.0	4.0	4.0	7.0	9.0	8.0	5.0	1.0
40.0	4.0	7.0	8.0	9.0	0.0	1.0	0.0	0.0
39.0	0.0	5.0	3.0	0.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>3.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	4.7	6.1	9.7	10.5	6.1	4.4	4.0	4.3
44.0	3.5	4.9	8.1	8.4	6.1	5.1	4.8	4.8
43.0	2.5	3.0	4.4	5.8	5.8	8.4	6.8	5.0
42.0	1.9	2.1	3.1	4.4	5.3	6.8	7.7	3.6
41.0	1.7	1.9	2.6	4.6	5.2	4.6	3.3	2.4
40.0	1.6	2.3	2.9	3.4	2.2	1.8	1.5	1.4
39.0	1.9	2.2	2.4	2.0	1.8	1.5	1.3	1.2

Significance test (mb*>3.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0			+		<	-	-	
44.0	<		>	+	<			-
43.0	-	-			<	>		
42.0		-						<
41.0		+			+	+		
40.0	+	>	>	>	-		-	-
39.0	-	+		-	-	-		

(b)

P_a=1

Observed cumulative count (mb*>4.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	3.0	2.0	9.0	9.0	2.0	0.0	1.0	3.0
44.0	0.0	3.0	5.0	7.0	0.0	2.0	3.0	2.0
43.0	0.0	1.0	1.0	2.0	1.0	8.0	1.0	2.0
42.0	0.0	0.0	2.0	2.0	0.0	4.0	12.0	0.0
41.0	0.0	1.0	1.0	8.0	4.0	3.0	2.0	1.0
40.0	3.0	4.0	8.0	7.0	0.0	0.0	0.0	0.0
39.0	0.0	6.0	3.0	1.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>4.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	2.2	3.1	7.3	9.0	1.3	0.8	0.7	1.9
44.0	0.4	2.3	6.2	5.0	1.3	2.1	2.5	2.8
43.0	0.2	0.6	1.2	2.7	1.9	9.0	5.4	2.3
42.0	0.3	0.2	1.1	2.4	2.9	6.2	12.0	0.5
41.0	0.5	1.3	1.5	5.5	6.8	5.0	1.7	0.4
40.0	1.3	2.9	3.8	5.0	0.3	0.3	0.1	0.1
39.0	0.6	2.9	1.4	0.5	0.2	0.1	0.1	0.1

Significance test (mb*>4.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0								
44.0								
43.0							<	
42.0					-			
41.0								
40.0	+		+					
39.0		+	+					

P_a=7

Observed cumulative count (mb*>4.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	3.0	2.0	9.0	9.0	2.0	0.0	1.0	3.0
44.0	0.0	3.0	5.0	7.0	0.0	2.0	3.0	2.0
43.0	0.0	1.0	1.0	2.0	1.0	8.0	1.0	2.0
42.0	0.0	0.0	2.0	2.0	0.0	4.0	12.0	0.0
41.0	0.0	1.0	1.0	8.0	4.0	3.0	2.0	1.0
40.0	3.0	4.0	8.0	7.0	0.0	0.0	0.0	0.0
39.0	0.0	6.0	3.0	1.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>4.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	2.1	3.2	7.0	7.9	2.0	1.2	1.2	1.9
44.0	1.0	2.2	5.6	4.7	2.0	2.0	2.3	2.4
43.0	0.6	0.9	1.6	2.7	2.4	8.2	5.4	2.5
42.0	0.6	0.6	1.3	2.4	3.1	6.2	10.0	1.2
41.0	0.7	1.1	1.6	5.0	5.9	4.3	1.8	0.7
40.0	1.0	2.4	3.3	4.0	0.8	0.5	0.3	0.3
39.0	1.4	2.4	2.0	0.9	0.5	0.3	0.2	0.2

Significance test (mb*>4.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0								
44.0					-			
43.0							<	
42.0					<			
41.0				+				
40.0	+			>	+			
39.0	-	+						

Observed cumulative count (mb*>4.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	3.0	2.0	9.0	9.0	2.0	0.0	1.0	3.0
44.0	0.0	3.0	5.0	7.0	0.0	2.0	3.0	2.0
43.0	0.0	1.0	1.0	2.0	1.0	8.0	1.0	2.0
42.0	0.0	0.0	2.0	2.0	0.0	4.0	12.0	0.0
41.0	0.0	1.0	1.0	8.0	4.0	3.0	2.0	1.0
40.0	3.0	4.0	8.0	7.0	0.0	0.0	0.0	0.0
39.0	0.0	6.0	3.0	1.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>4.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	2.4	3.1	5.0	5.4	3.0	2.1	1.9	2.0
44.0	1.8	2.5	4.2	4.2	3.0	2.4	2.3	2.2
43.0	1.3	1.6	2.2	2.9	2.8	7.2	5.7	2.3
42.0	1.0	1.1	1.6	2.3	2.6	5.8	6.6	1.7
41.0	0.9	1.1	1.4	4.3	4.7	4.0	2.8	1.1
40.0	0.9	2.3	2.9	3.3	1.2	0.9	0.7	0.7
39.0	2.0	2.3	2.4	1.1	0.9	0.7	0.6	0.6

Significance test (mb*>4.5; since 1627)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0			+	+		-		
44.0	-			+	<			
43.0					-		<	
42.0					-		+	-
41.0				+				
40.0	+			>	+			
39.0	-	>						

P_a=50

(c)

P_a=1

Observed cumulative count (mb*>4.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	2.0	1.0	6.0	2.0	1.0	0.0	1.0	2.0
44.0	0.0	1.0	4.0	6.0	0.0	2.0	2.0	1.0
43.0	0.0	0.0	1.0	2.0	1.0	6.0	1.0	2.0
42.0	0.0	0.0	1.0	2.0	0.0	3.0	3.0	0.0
41.0	0.0	1.0	1.0	5.0	2.0	2.0	1.0	1.0
40.0	3.0	2.0	6.0	3.0	0.0	0.0	0.0	0.0
39.0	0.0	3.0	3.0	0.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>4.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	1.7	2.4	5.5	6.8	1.0	0.6	0.5	1.4
44.0	0.3	1.7	4.6	3.8	1.0	1.6	1.9	2.1
43.0	0.2	0.5	0.9	2.0	1.4	4.3	2.6	1.8
42.0	0.3	0.2	0.8	1.8	2.2	3.0	5.8	0.4
41.0	0.4	1.0	1.2	2.6	3.3	2.4	0.8	0.3
40.0	0.9	1.3	1.8	2.4	0.3	0.3	0.1	0.1
39.0	0.3	1.4	0.6	0.4	0.1	0.1	0.1	0.1

Significance test (mb*>4.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0				<				
44.0								
43.0								
42.0					-		-	
41.0				+				
40.0	+			>				
39.0		+		+				

P_a=7

Observed cumulative count (mb*>4.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	2.0	1.0	6.0	2.0	1.0	0.0	1.0	2.0
44.0	0.0	1.0	4.0	6.0	0.0	2.0	2.0	1.0
43.0	0.0	0.0	1.0	2.0	1.0	6.0	1.0	2.0
42.0	0.0	0.0	1.0	2.0	0.0	3.0	3.0	0.0
41.0	0.0	1.0	1.0	5.0	2.0	2.0	1.0	1.0
40.0	3.0	2.0	6.0	3.0	0.0	0.0	0.0	0.0
39.0	0.0	3.0	3.0	0.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>4.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	1.6	2.4	5.3	6.0	1.5	0.9	0.9	1.4
44.0	0.7	1.7	4.2	3.6	1.6	1.5	1.8	1.9
43.0	0.4	0.7	1.2	2.1	1.8	3.7	2.4	1.9
42.0	0.4	0.4	1.0	1.8	2.4	2.8	4.5	0.9
41.0	0.5	0.8	1.2	2.2	2.7	1.9	0.8	0.5
40.0	0.8	1.0	1.5	1.8	0.6	0.4	0.2	0.2
39.0	0.6	1.1	0.9	0.7	0.4	0.2	0.2	0.2

Significance test (mb*>4.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0				-				
44.0				+	-			
43.0								
42.0					-			
41.0				+				
40.0	+			>				
39.0		+		+				

Observed cumulative count (mb*>4.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	2.0	1.0	6.0	2.0	1.0	0.0	1.0	2.0
44.0	0.0	1.0	4.0	6.0	0.0	2.0	2.0	1.0
43.0	0.0	0.0	1.0	2.0	1.0	6.0	1.0	2.0
42.0	0.0	0.0	1.0	2.0	0.0	3.0	3.0	0.0
41.0	0.0	1.0	1.0	5.0	2.0	2.0	1.0	1.0
40.0	3.0	2.0	6.0	3.0	0.0	0.0	0.0	0.0
39.0	0.0	3.0	3.0	0.0	0.0	0.0	1.0	0.0

Expected cumulative count (mb*>4.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0	1.8	2.4	3.9	4.2	2.3	1.7	1.5	1.6
44.0	1.4	1.9	3.2	3.3	2.3	1.9	1.8	1.7
43.0	1.0	1.2	1.7	2.3	2.2	2.8	2.2	1.8
42.0	0.8	0.9	1.3	1.7	2.0	2.3	2.6	1.4
41.0	0.7	0.8	1.1	1.7	1.8	1.5	1.1	0.9
40.0	0.7	0.9	1.1	1.3	0.9	0.7	0.6	0.5
39.0	0.7	0.9	0.9	0.8	0.7	0.6	0.5	0.5

Significance test (mb*>4.5; since 1915)

	76.0	75.0	74.0	73.0	72.0	71.0	70.0	69.0
45.0						-		
44.0	-			+	-			
43.0					-	+		
42.0					-			-
41.0				>				
40.0	+			>	+			
39.0		+		+				

P_a=50

(d)

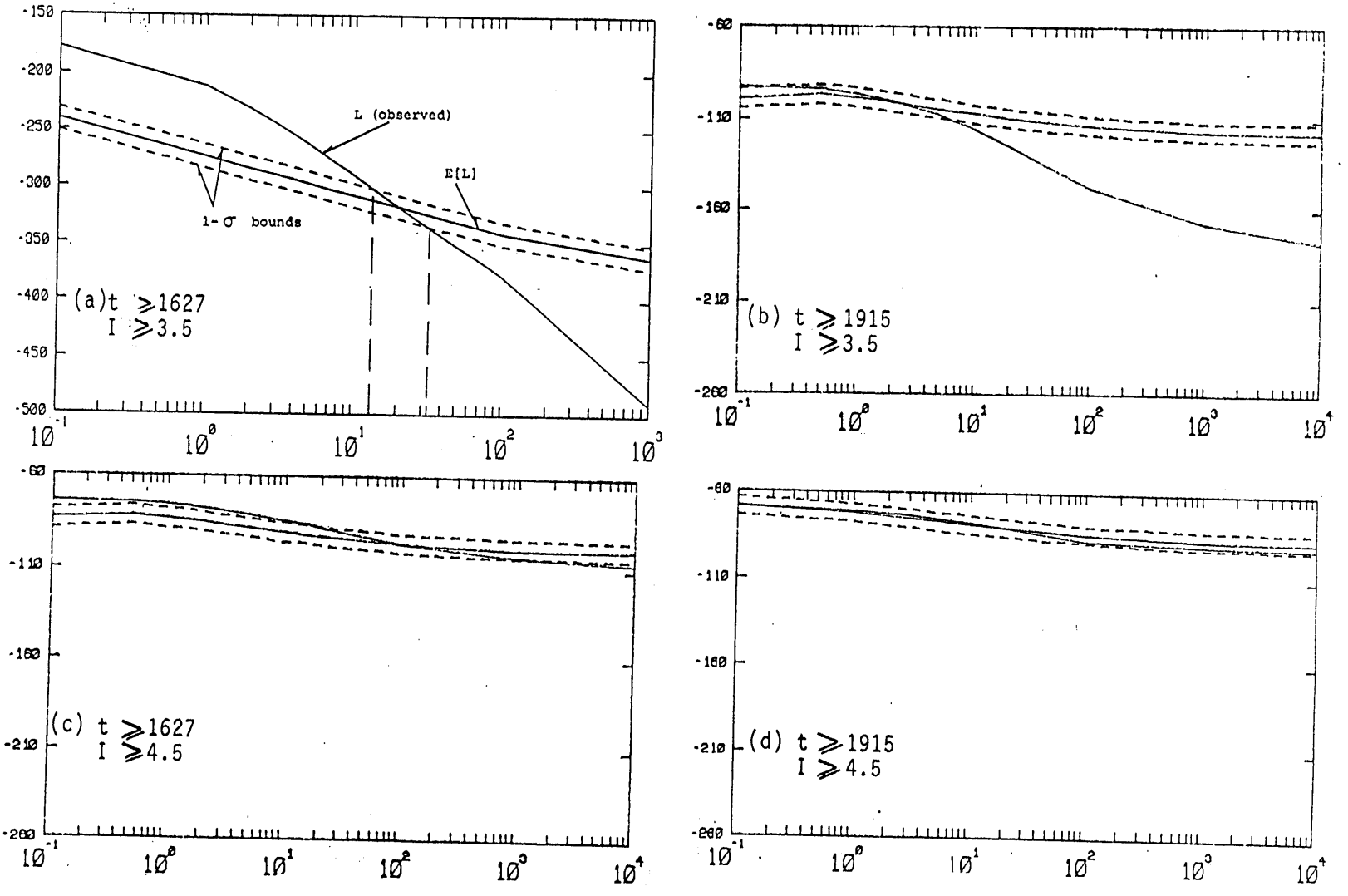


Figure 3-8: Selection of the penalty parameter P_a using the expected log-likelihood

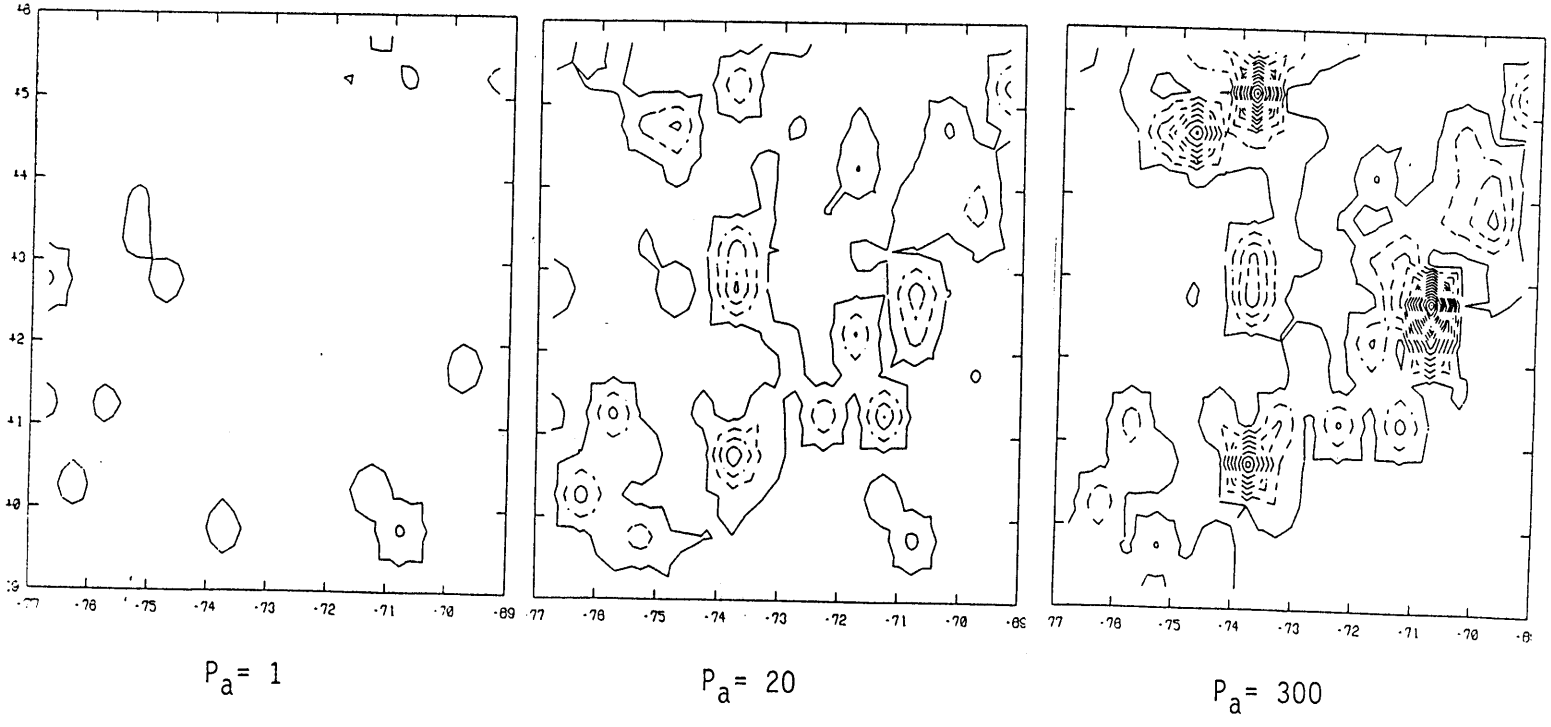


Figure 3-9: Spatial decomposition of L-E[L] for different penalties P_a.

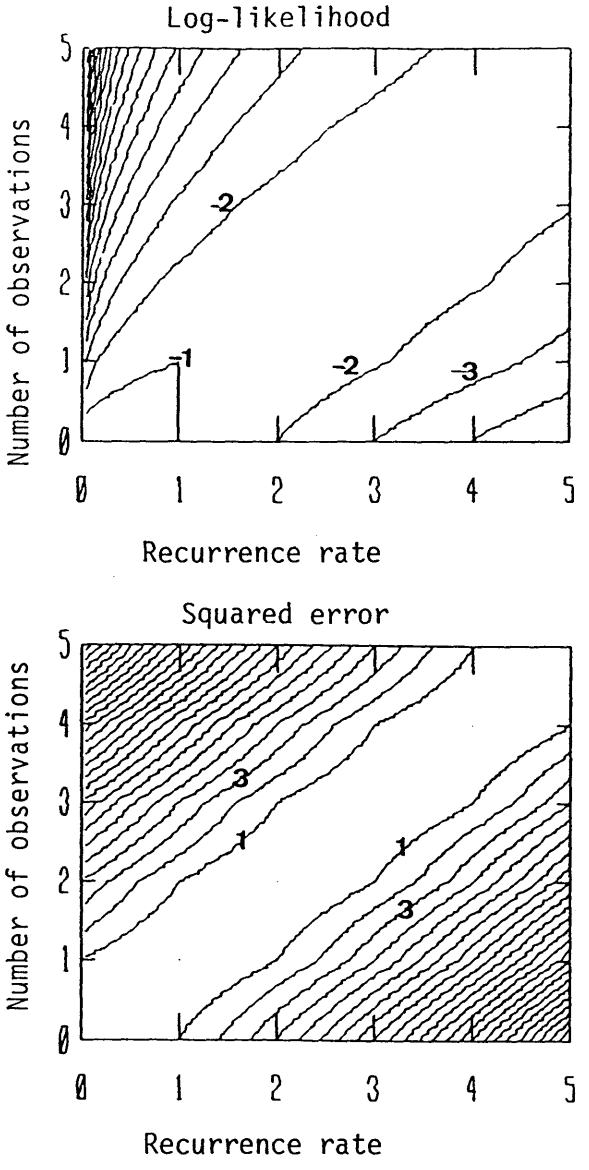


Figure 3-10: Log-likelihood and squared error as a function of the recurrence rate and the number of observations.

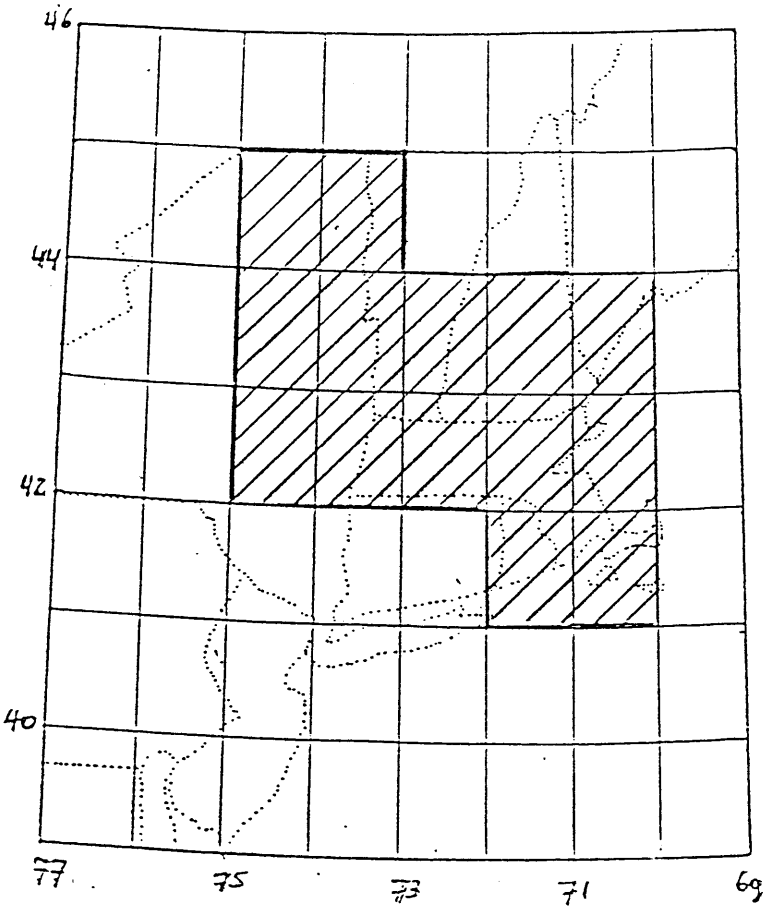
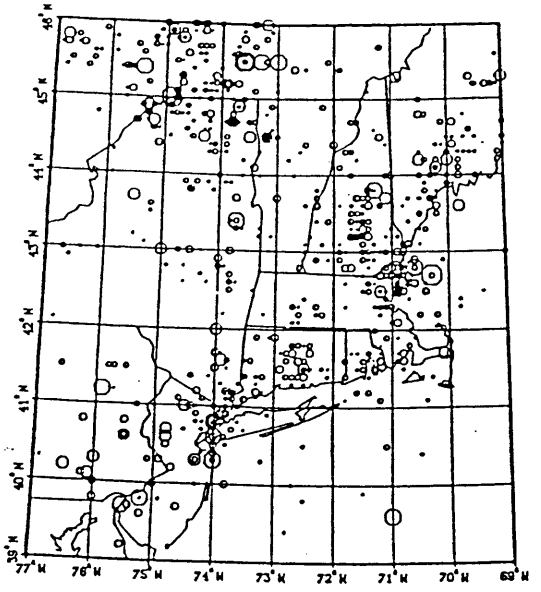
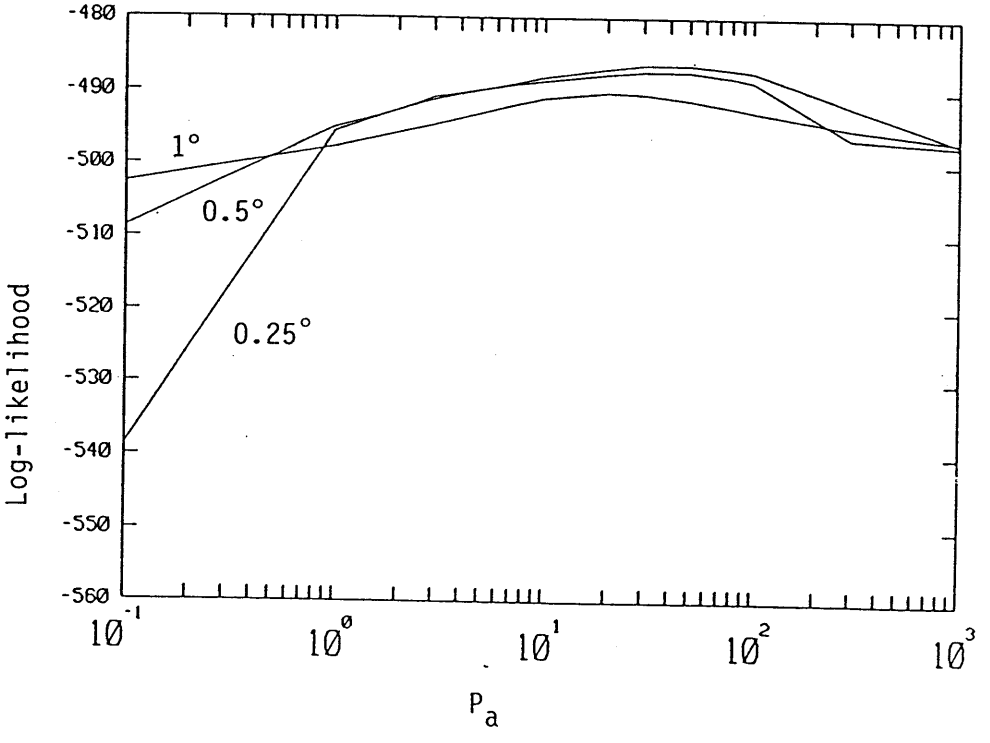


Figure 3-11: (a) Subregion for the selection of the optimal grid size and (b) selection of the optimal grid-size using the Log-likelihood.


46°N
45°N
44°N
43°N
42°N
41°N
40°N
39°N
77°W
76°W
75°W
74°W
73°W
72°W
71°W
70°W
69°W


-480
-490
-500
-510
-520
-530
-540
-550
-560
Log-likelihood
1°
0.5°
0.25°
10^-1
10^0
10^1
10^2
10^3
Pa

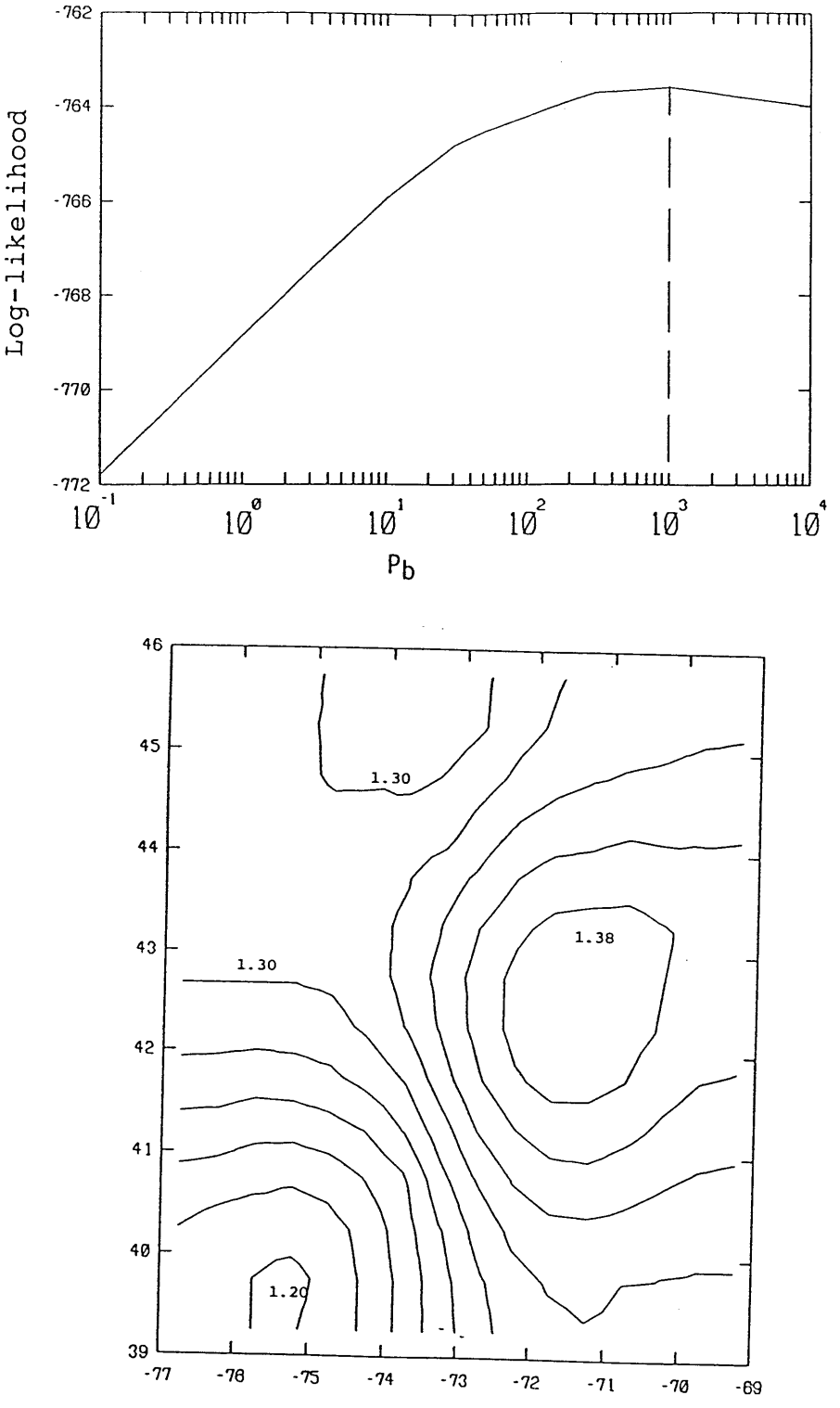


Figure 3-12: (a) Optimal penalty P_b and (b) associated estimate of $b(\underline{x})$. The penalty P_a is fixed to 7.

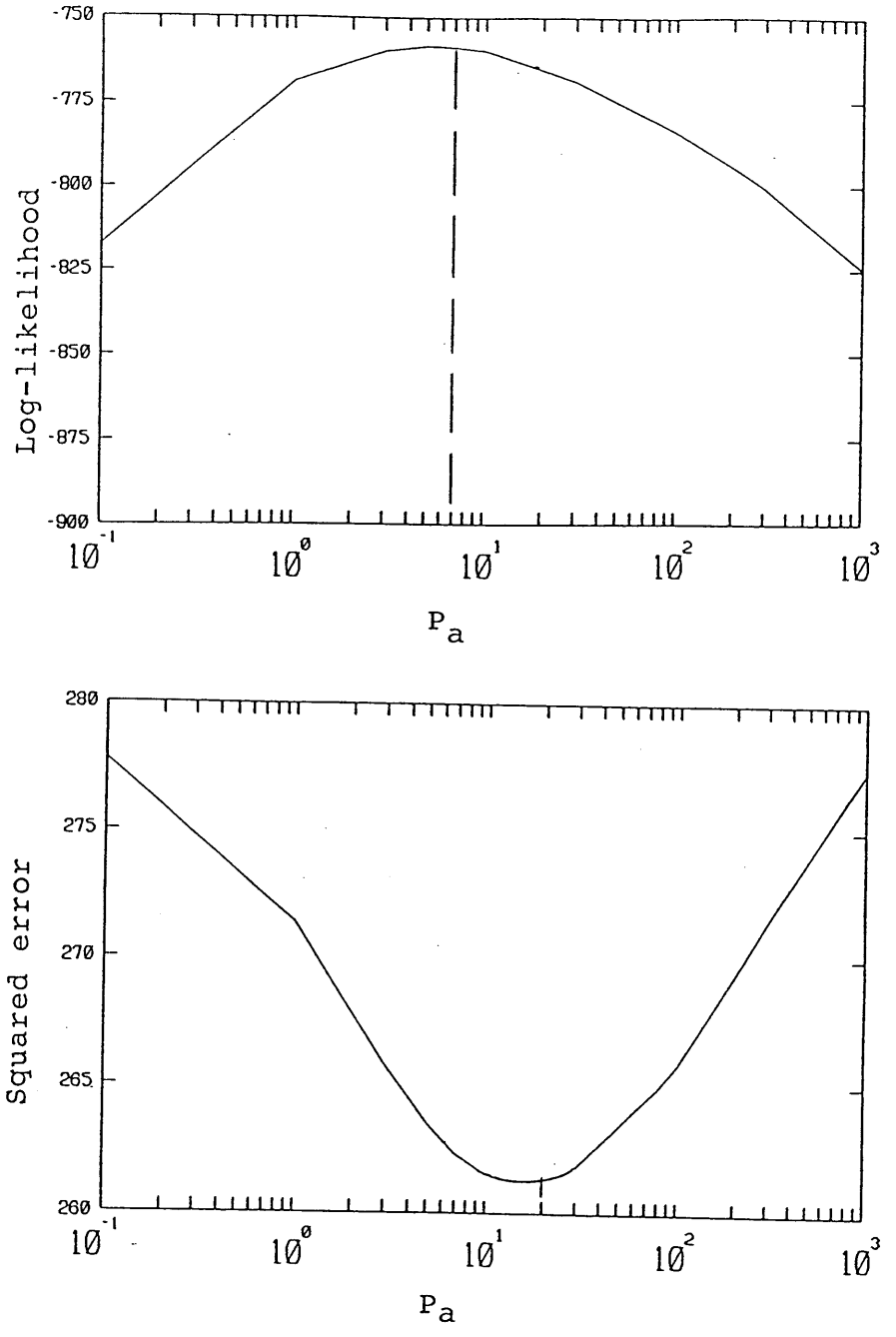


Figure 3-13: Optimal penalty P_a according to the cross-validated log-likelihood and cross-validated squared error criterion.

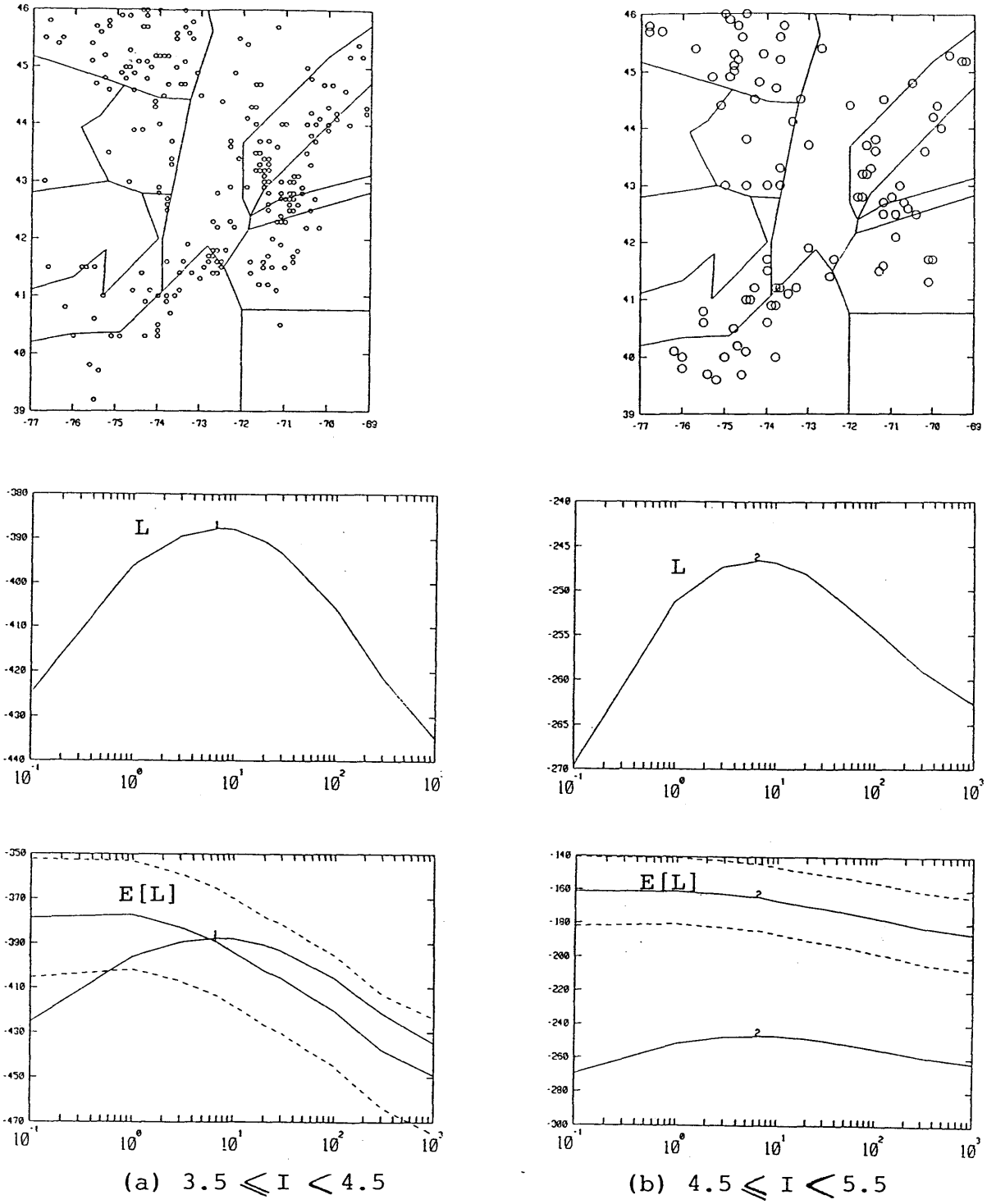
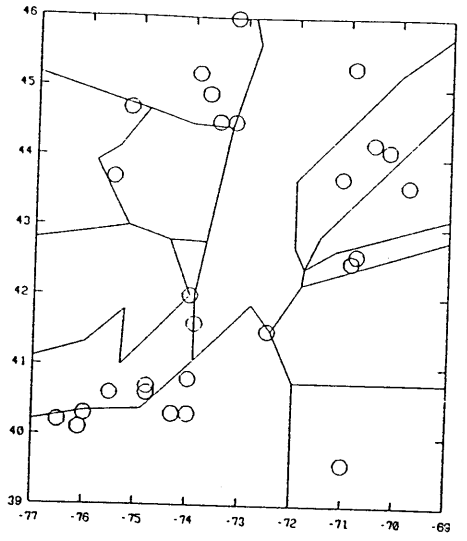
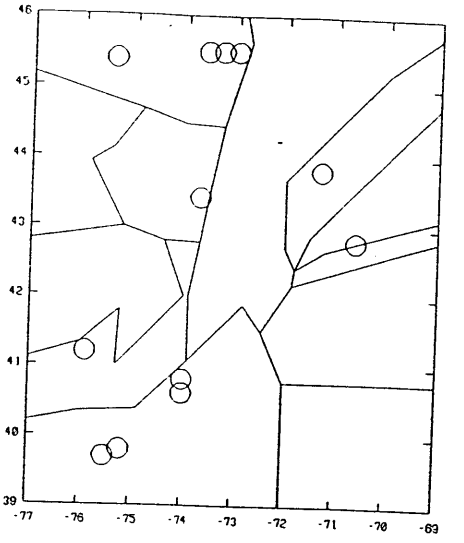
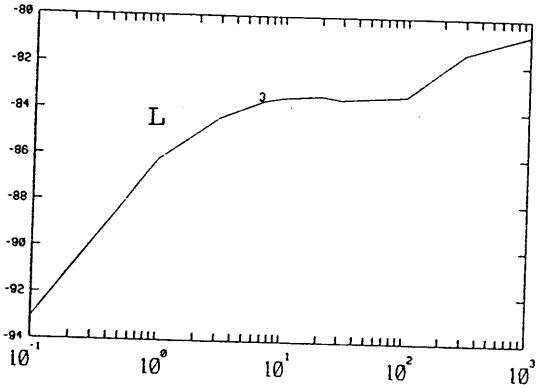
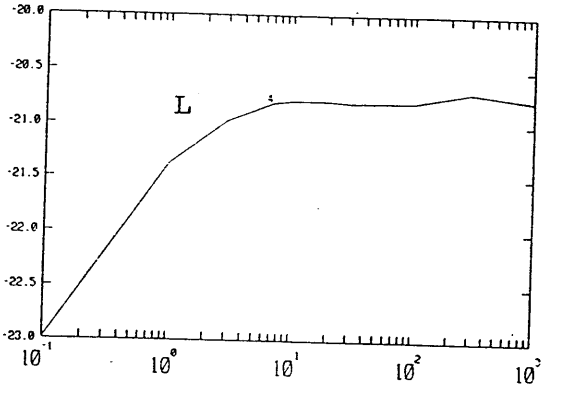


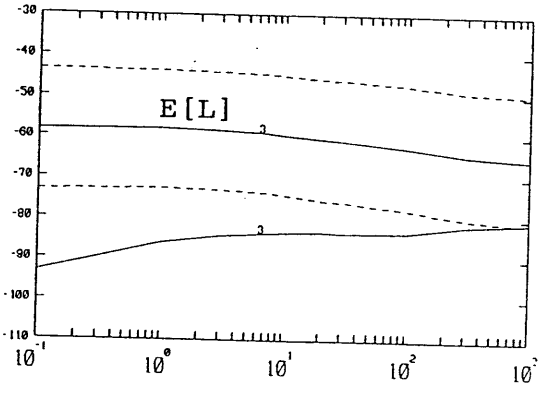
Figure 3-14: Cross-validated log-likelihood as a function of the penalty P_a for different intensity intervals.

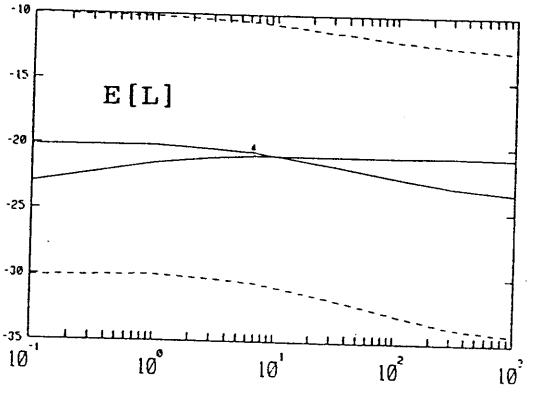






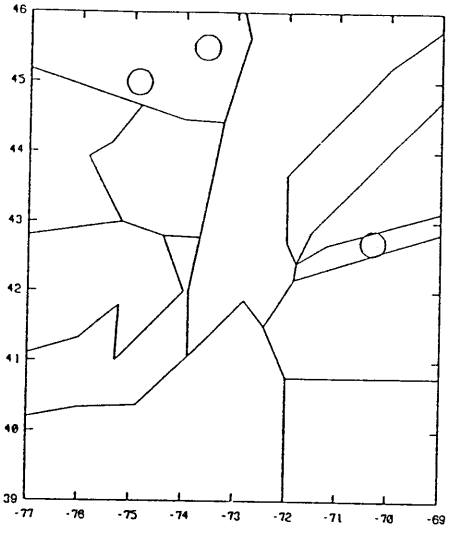


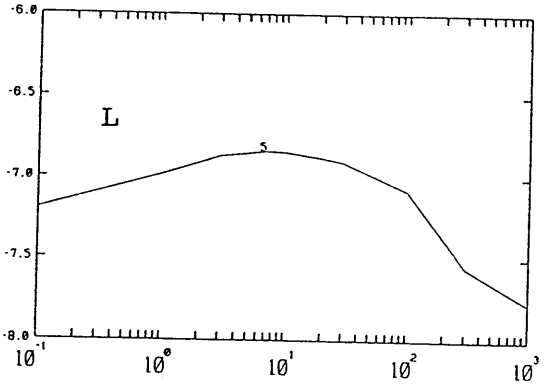


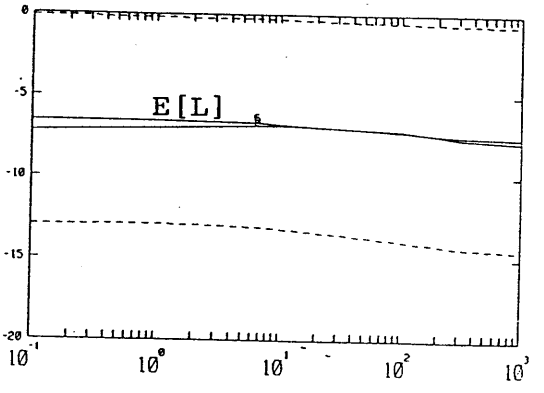


(c) $5.5 \leqslant I < 6.5$

(d) $6.5 \leqslant I < 7.5$







(e) $7.5 \leqslant I < 8.5$

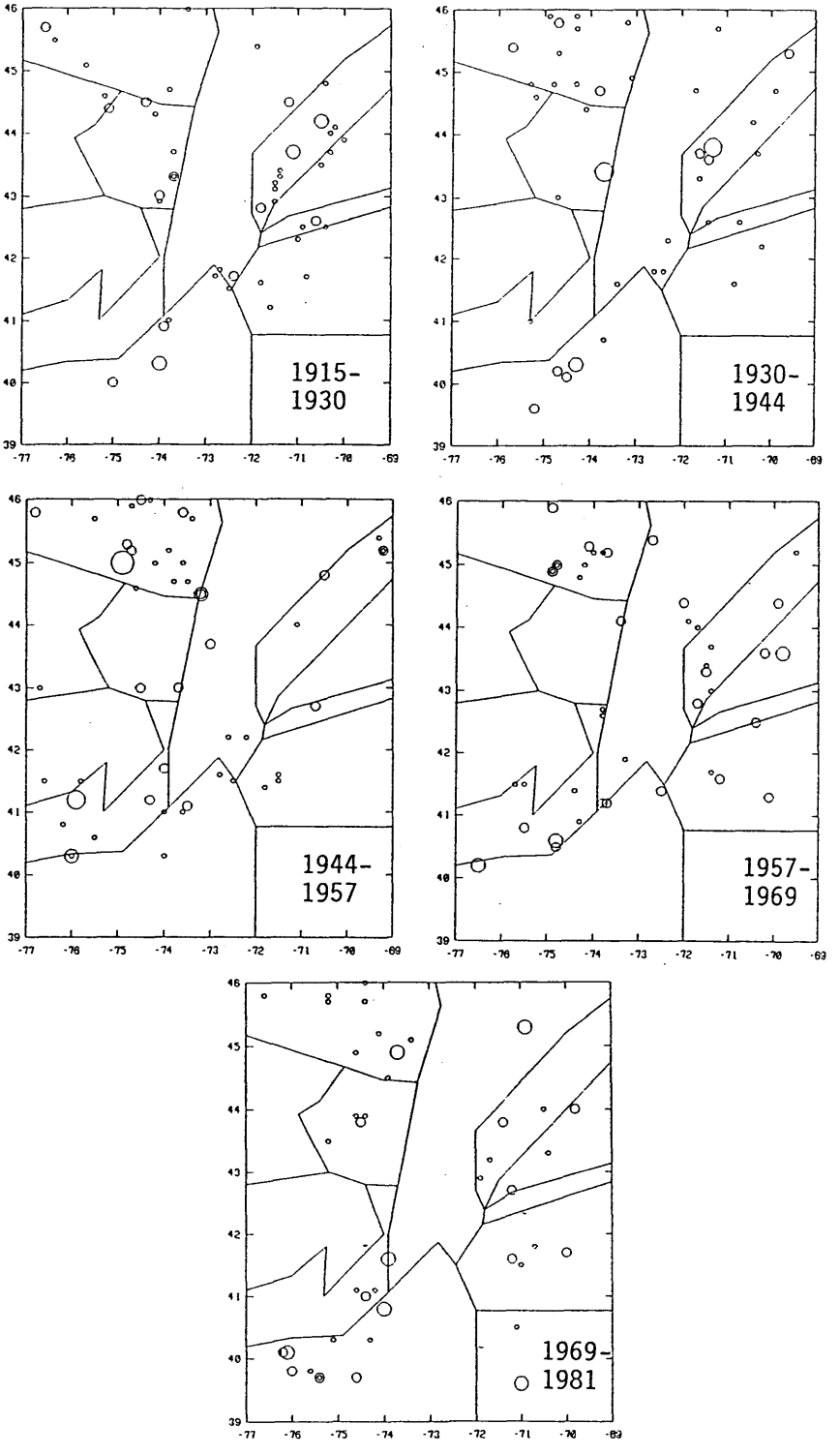
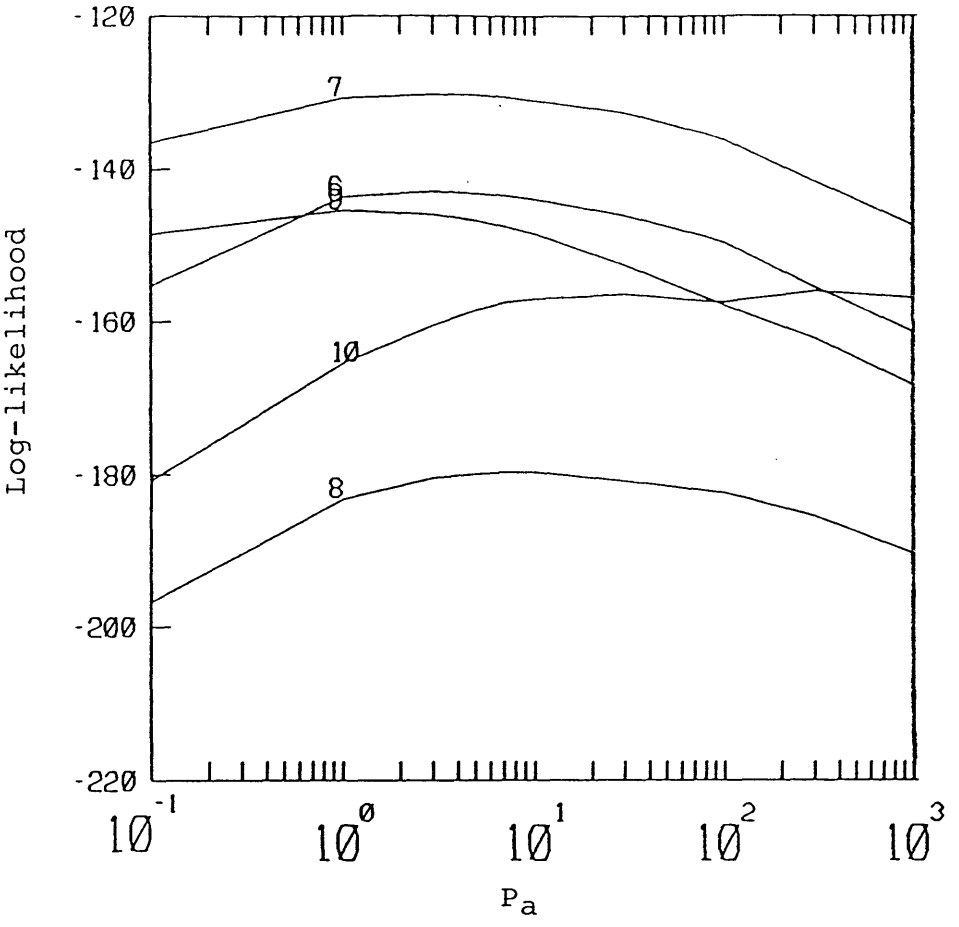


Figure 3-15: (a) Spatial distribution of the events for the 5 time intervals used for cross-validation, and (b) associated cross-validated log-likelihood as a function of P_a.



Period 6 : 1915-1930
Period 7 : 1930-1944
Period 8 : 1944-1957
Period 9 : 1957-1969
Period 10: 1969-1981

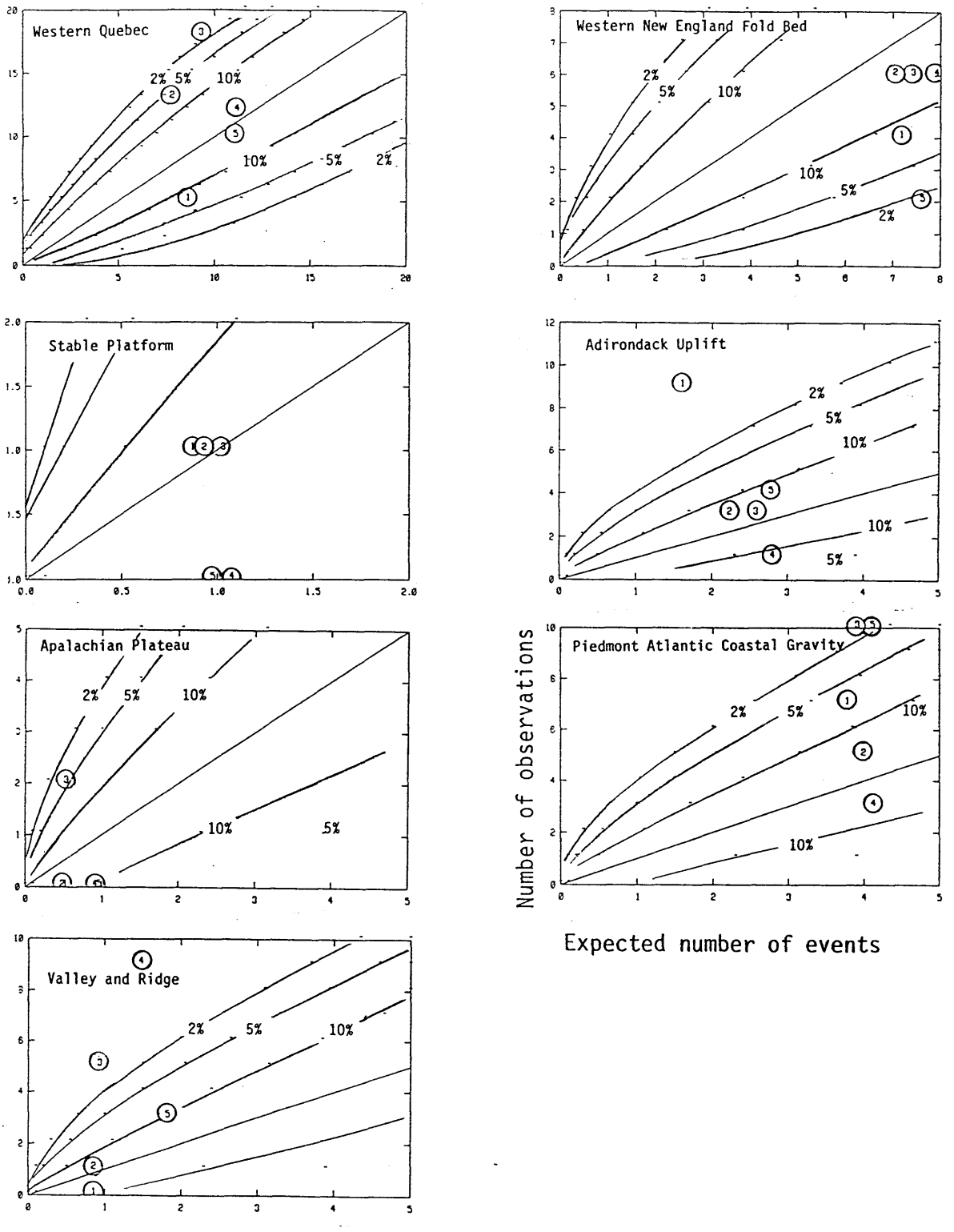
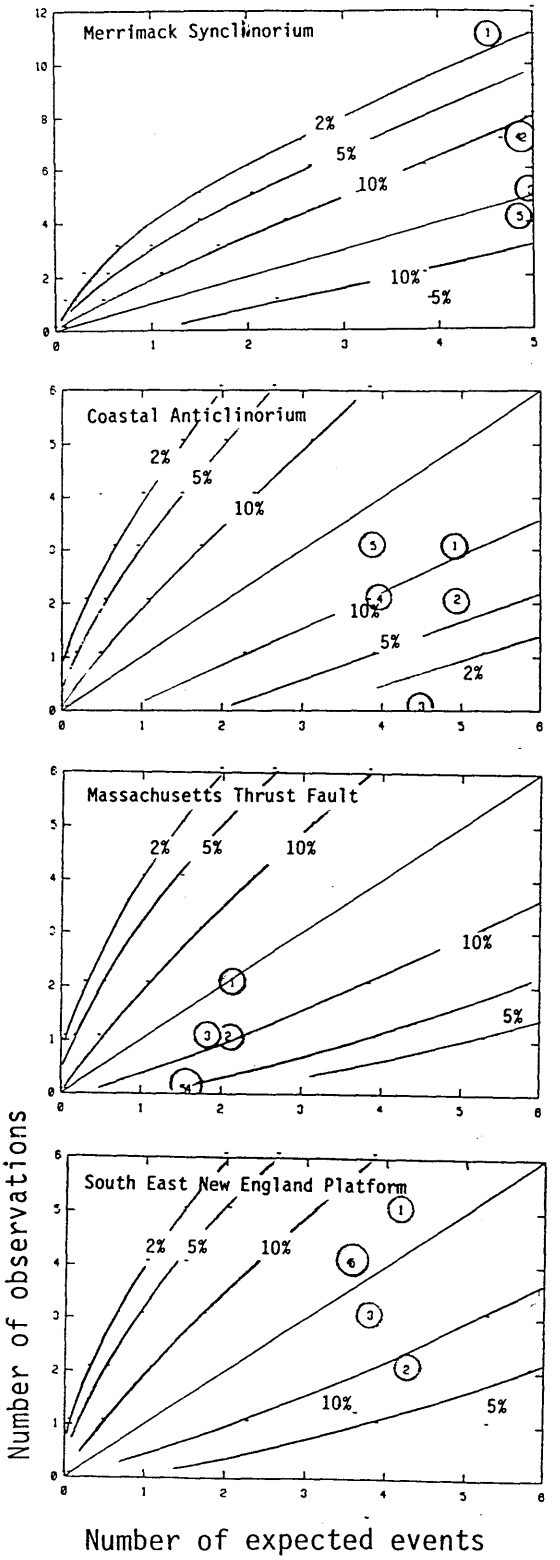
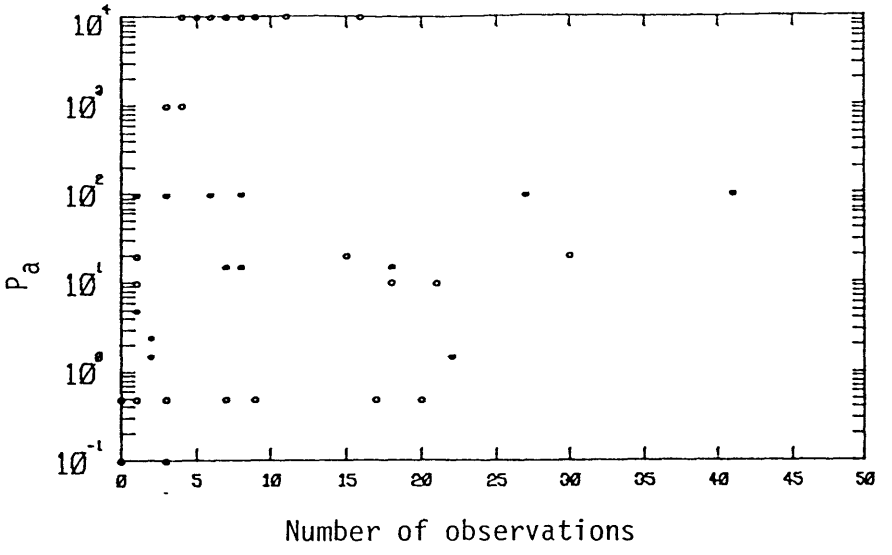


Figure 3-16: Comparison of the number of expected and observed events in each validation sample and for the seismogenic regions of Figure 5a.


Merrimack Synclinorium
2%
5%
10%
10%
5%
Coastal Anticlinorium
2%
5%
10%
10%
5%
2%
Massachusetts Thrust Fault
2%
5%
10%
10%
5%
South East New England Platform
2%
5%
10%
10%
5%
Number of observations
Number of expected events



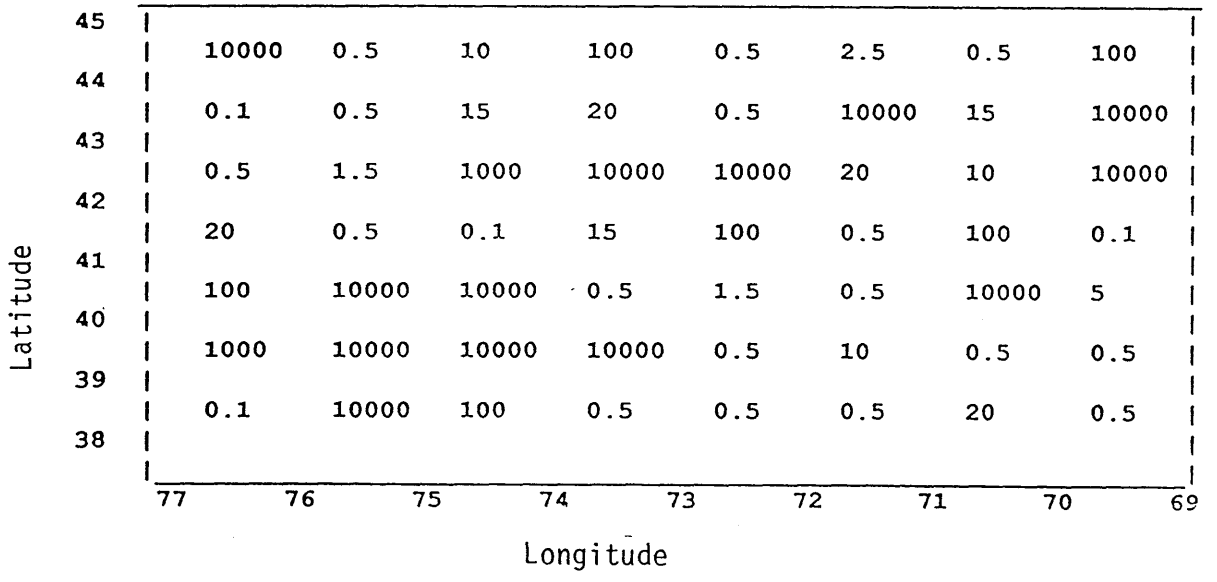


Figure 3-17: Optimal penalties P_a as a function of location and the total number of observations in each cell.

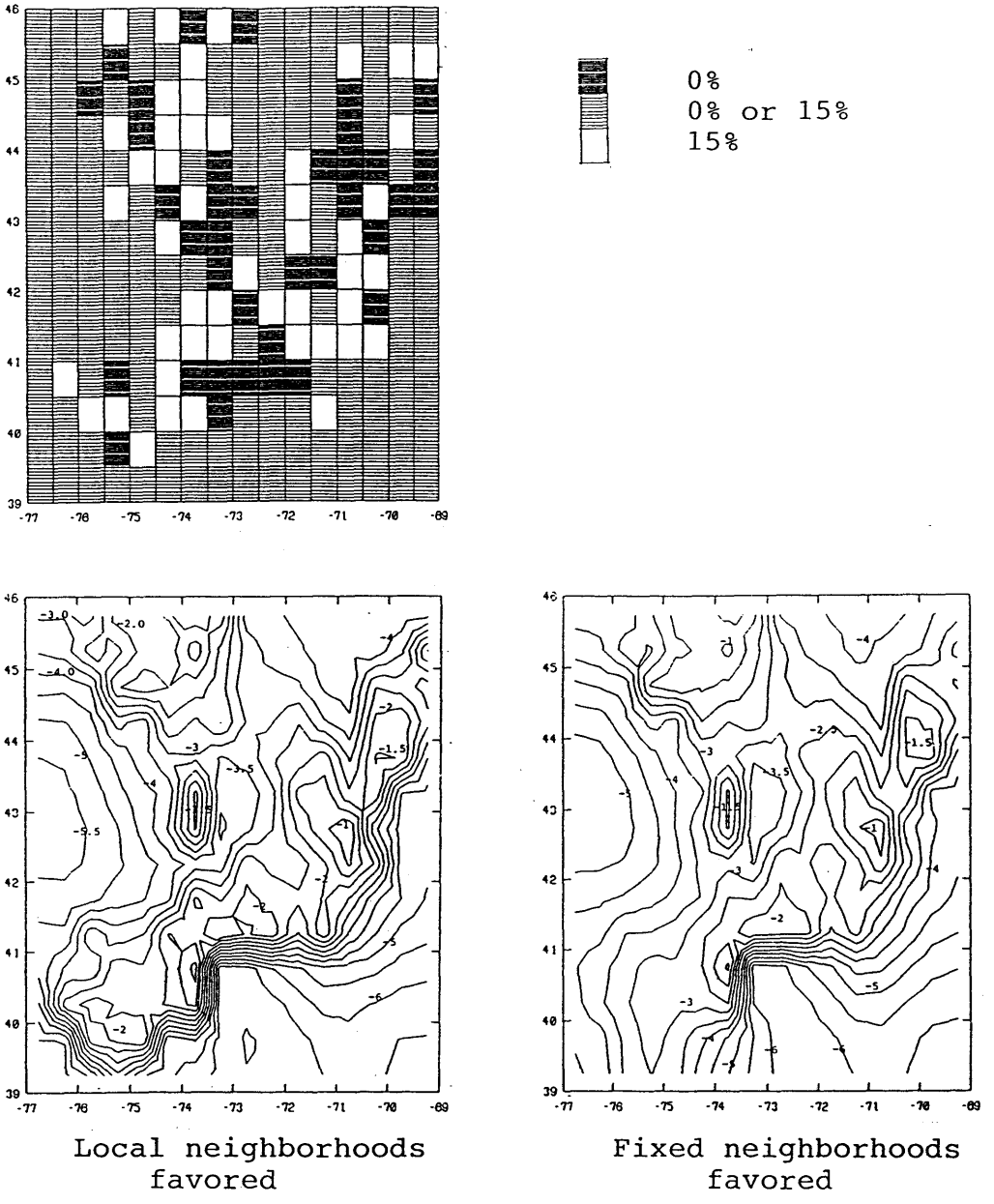


Figure 3-18: Two solutions with α that varies from cell to cell (a) shows the cells that are sensitive to α and those for which the optimal value of α are 0% and 15%.

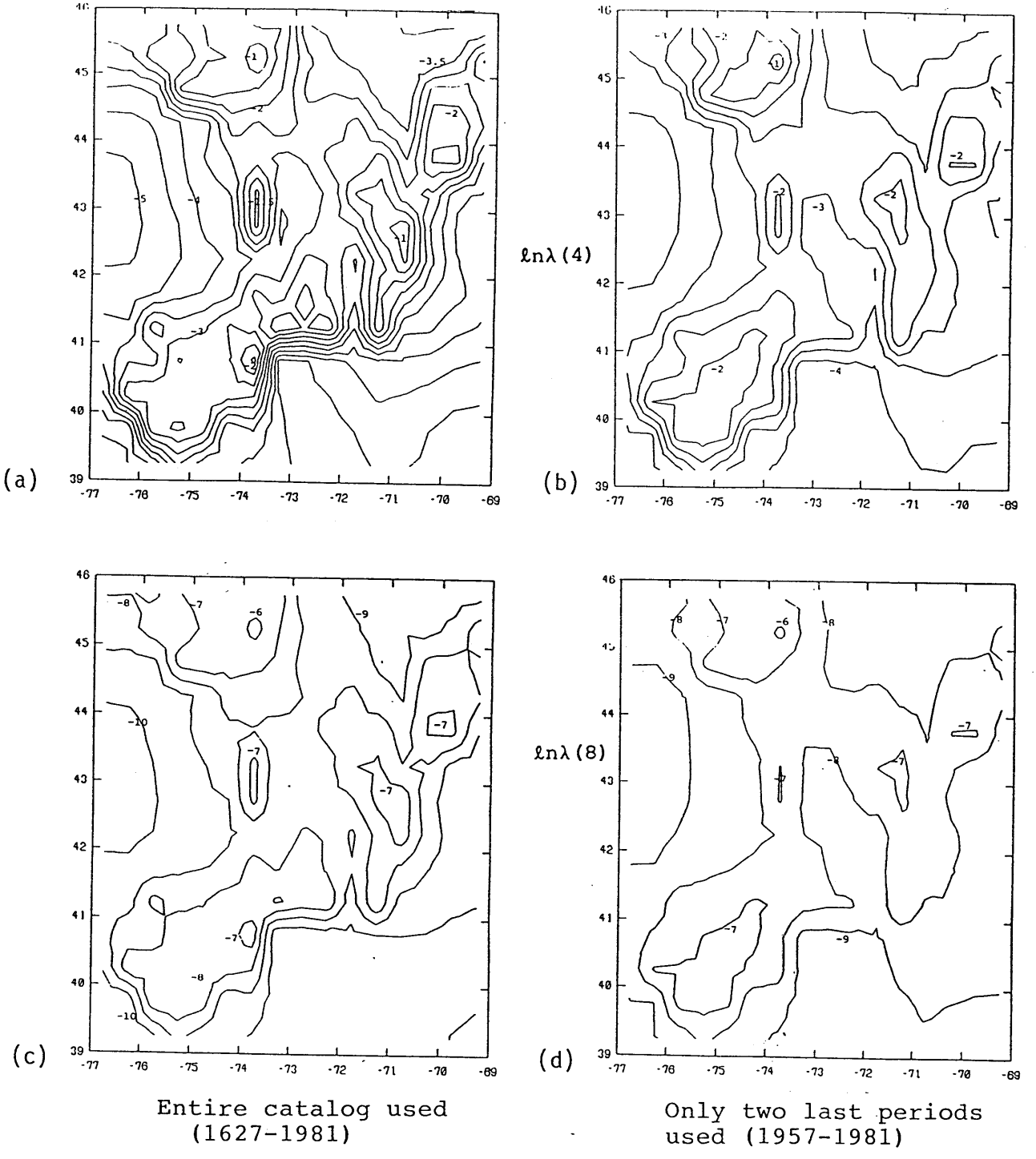


Figure 3-19: Comparison of the log-rates using the entire catalog and the last 2 time intervals.

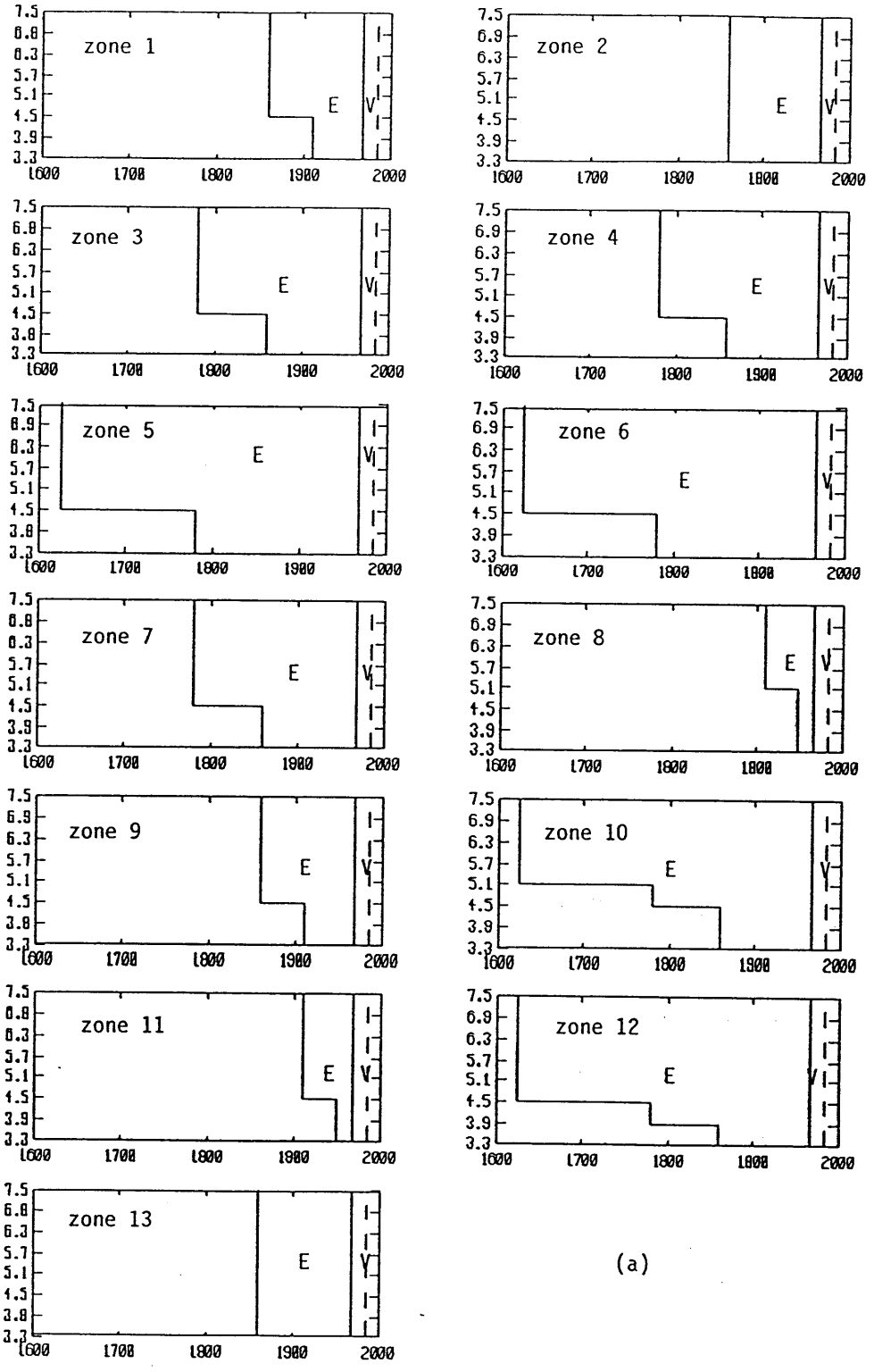
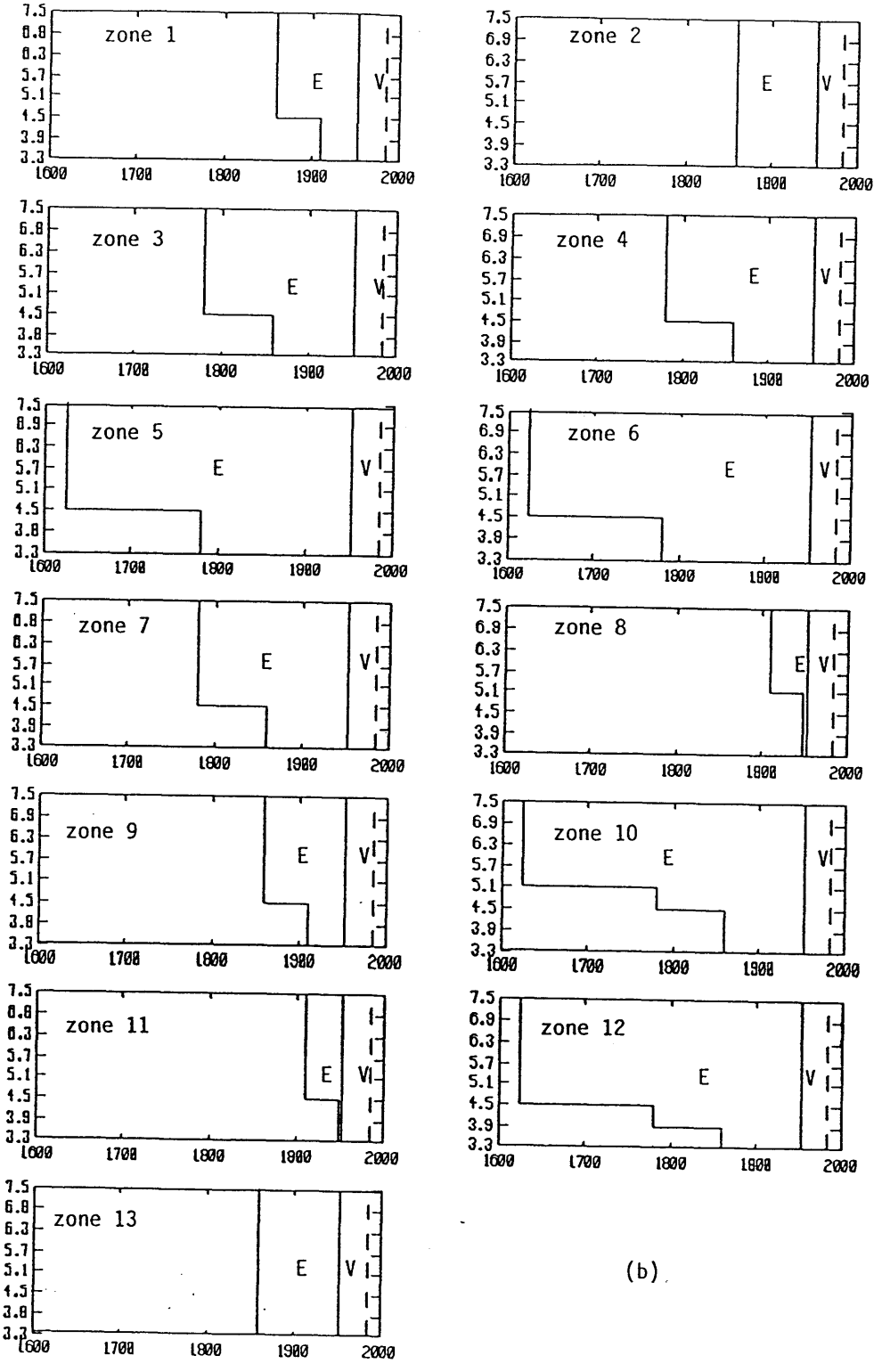
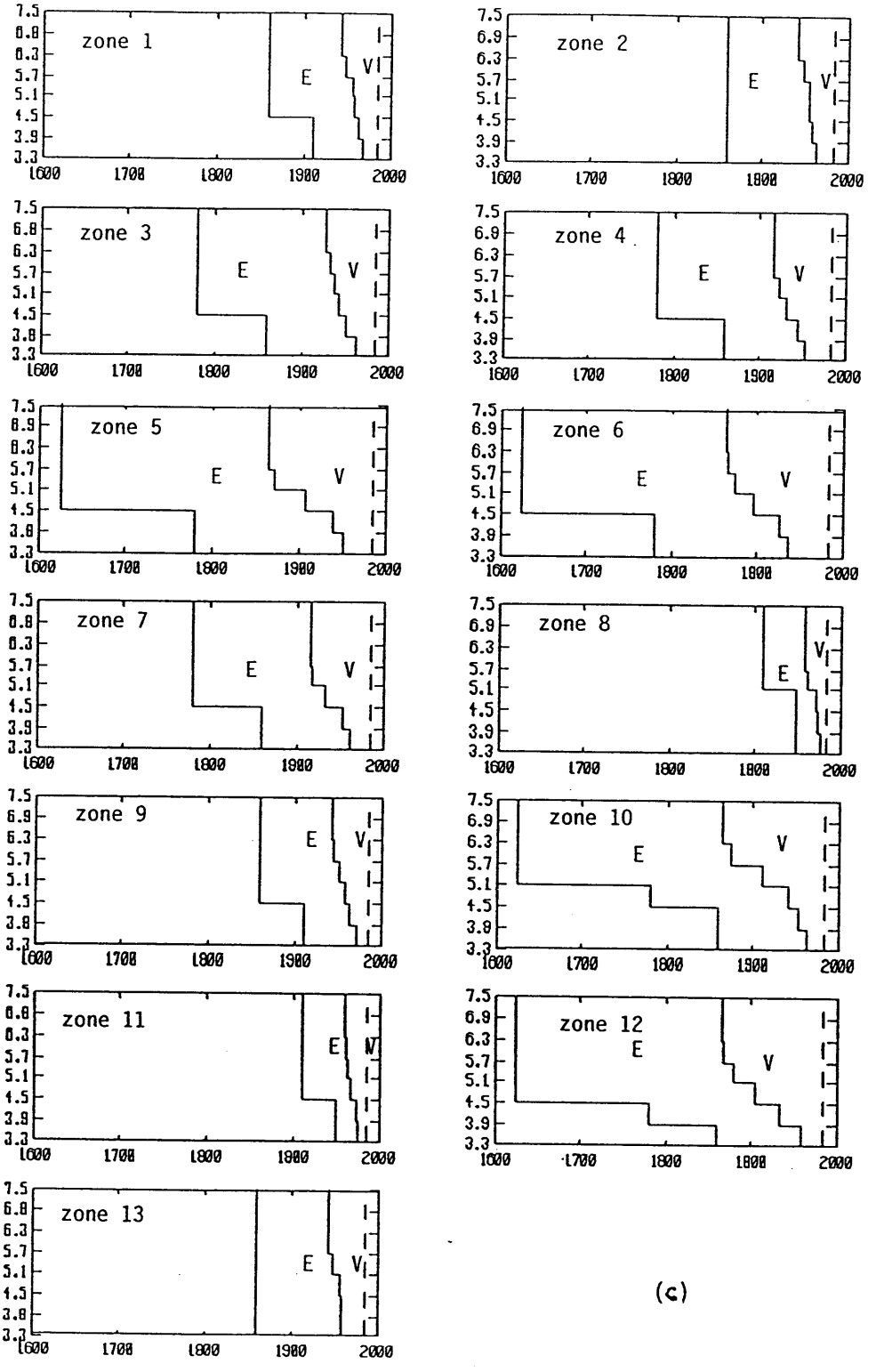


Figure 3-20: Partition of the catalog into a validation and estimation data set for each zone of incompleteness as a function of time and magnitude (a) last 15 years of observations, (b) last 30 years of observation, and (c) the last 1/3 of the complete catalog.


zone 1
zone 2
zone 3
zone 4
zone 5
zone 6
zone 7
zone 8
zone 9
zone 10
zone 11
zone 12
zone 13
E
V
7.5
6.9
6.3
5.7
5.1
4.5
3.9
3.3
1600
1700
1800
1900
2000

(b)


zone 1
zone 2
zone 3
zone 4
zone 5
zone 6
zone 7
zone 8
zone 9
zone 10
zone 11
zone 12
zone 13
E
V
7.5
6.8
6.3
5.7
5.1
4.5
3.8
3.3
1600
1700
1800
1900
2000

(c)

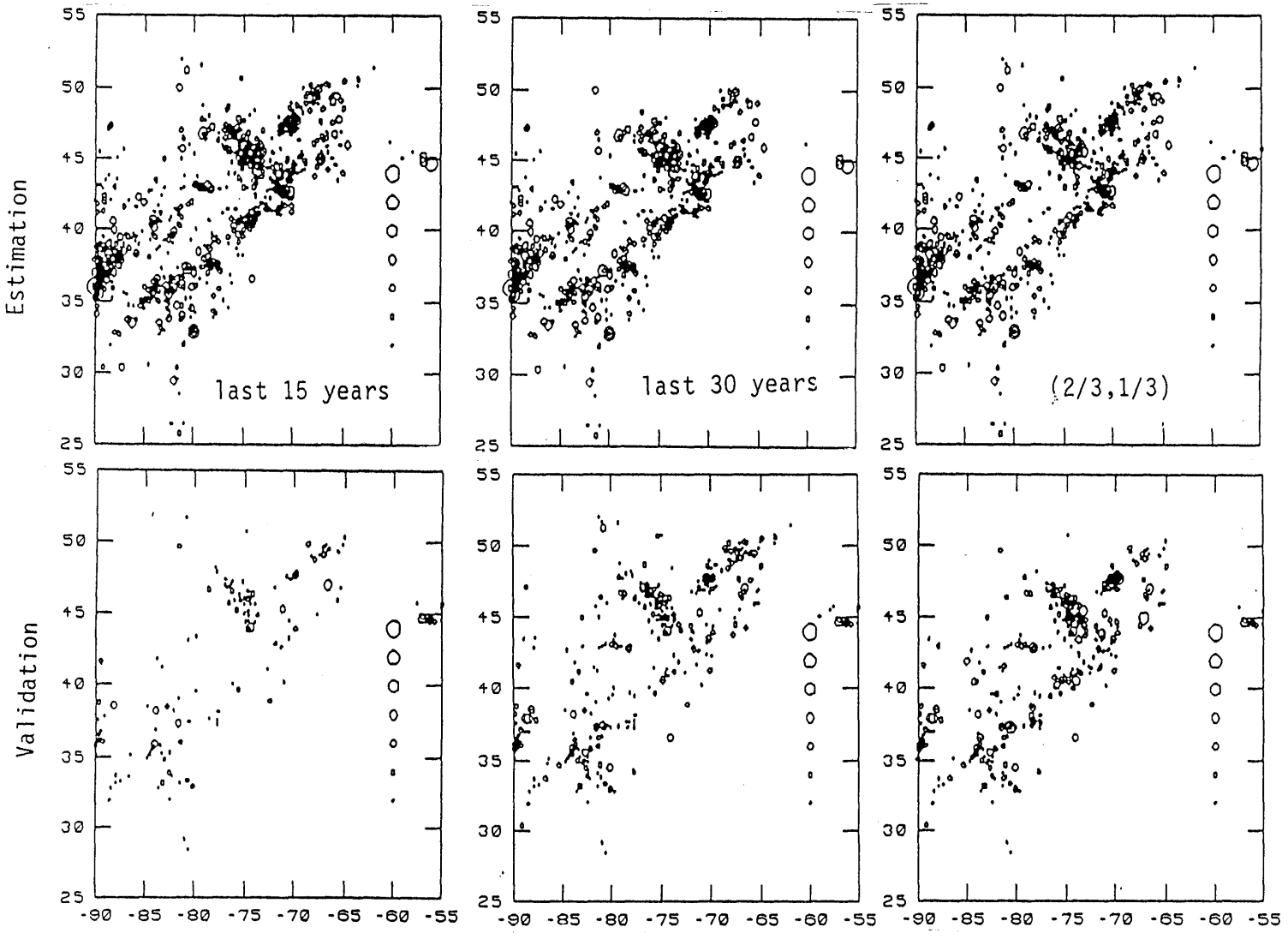
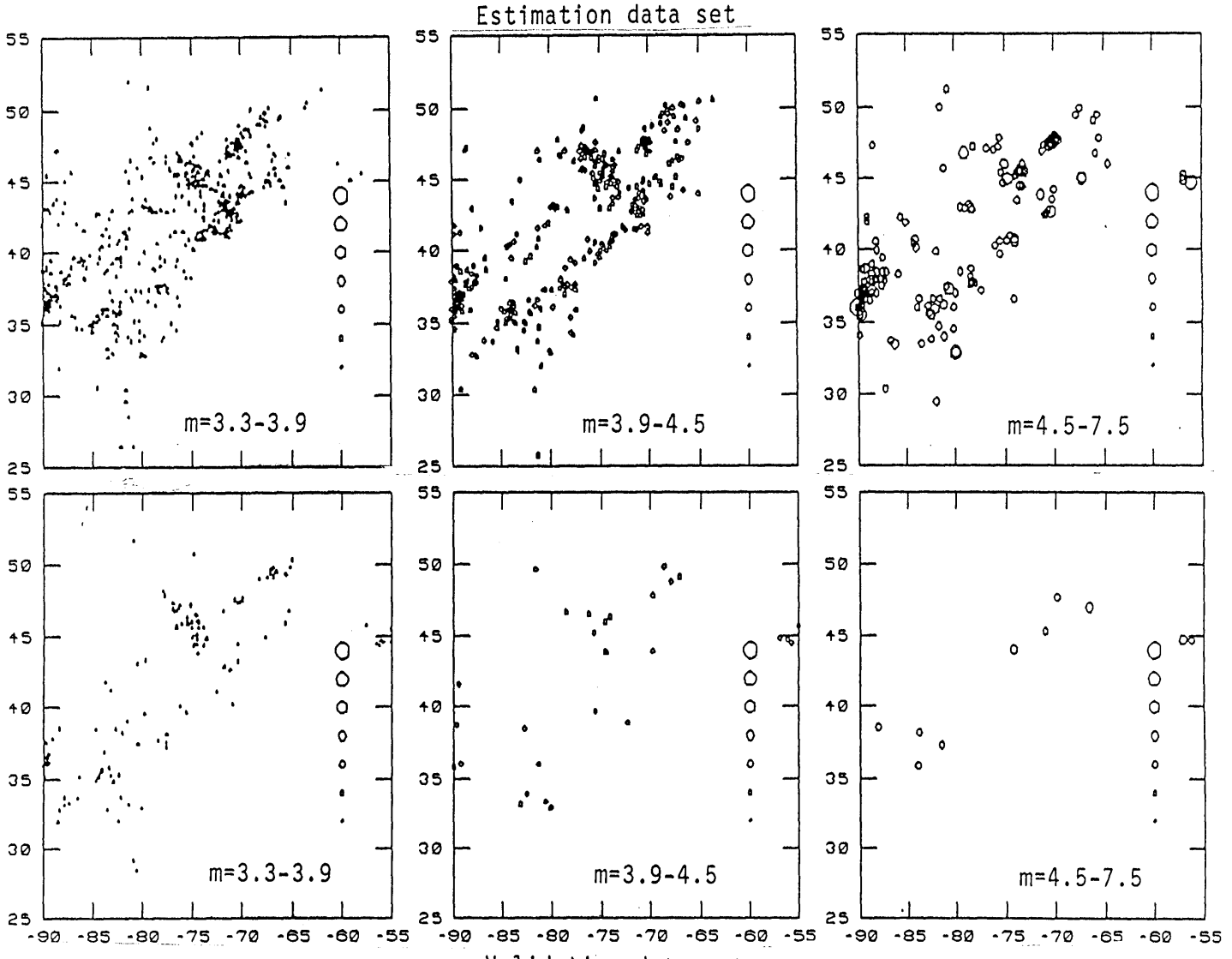
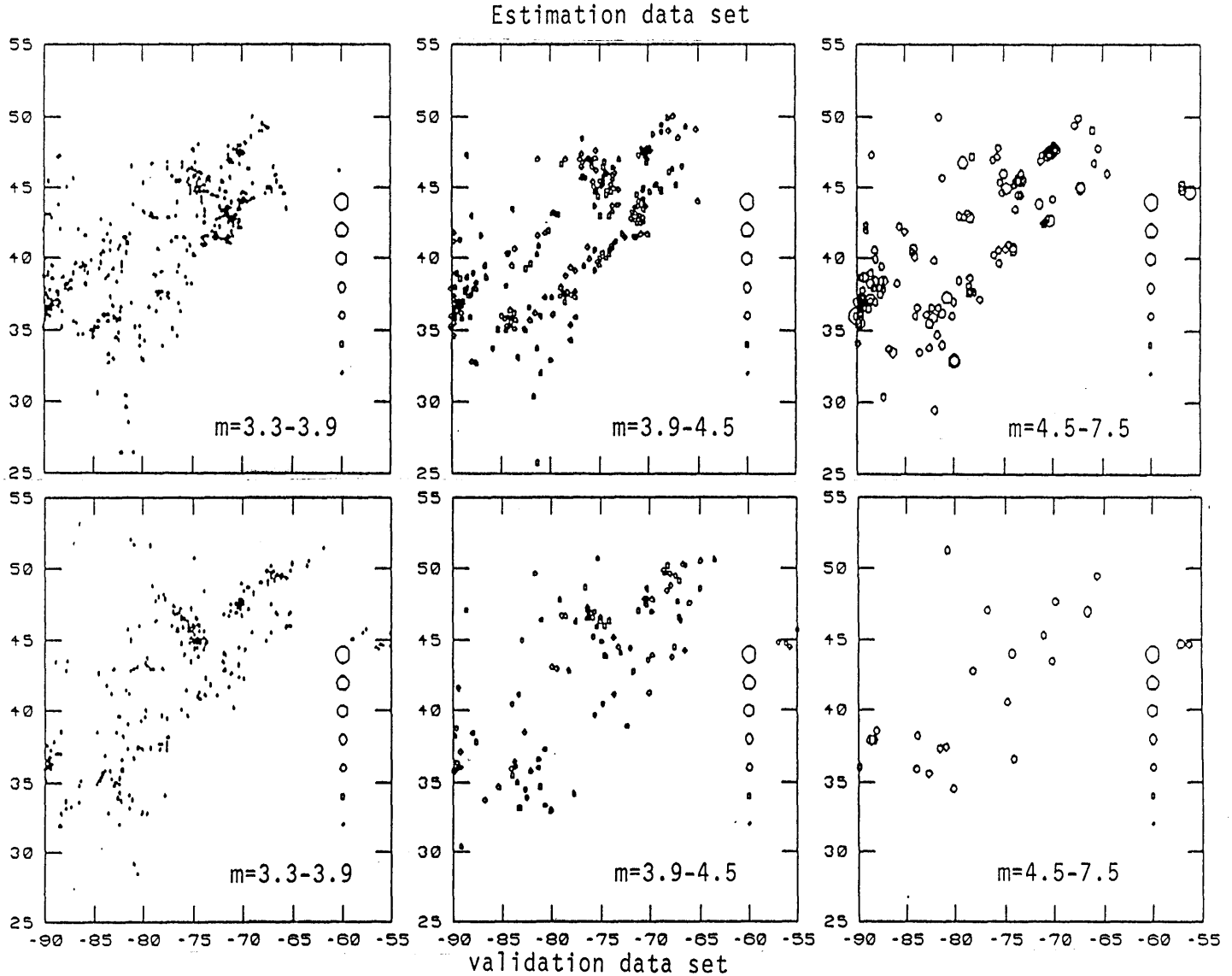
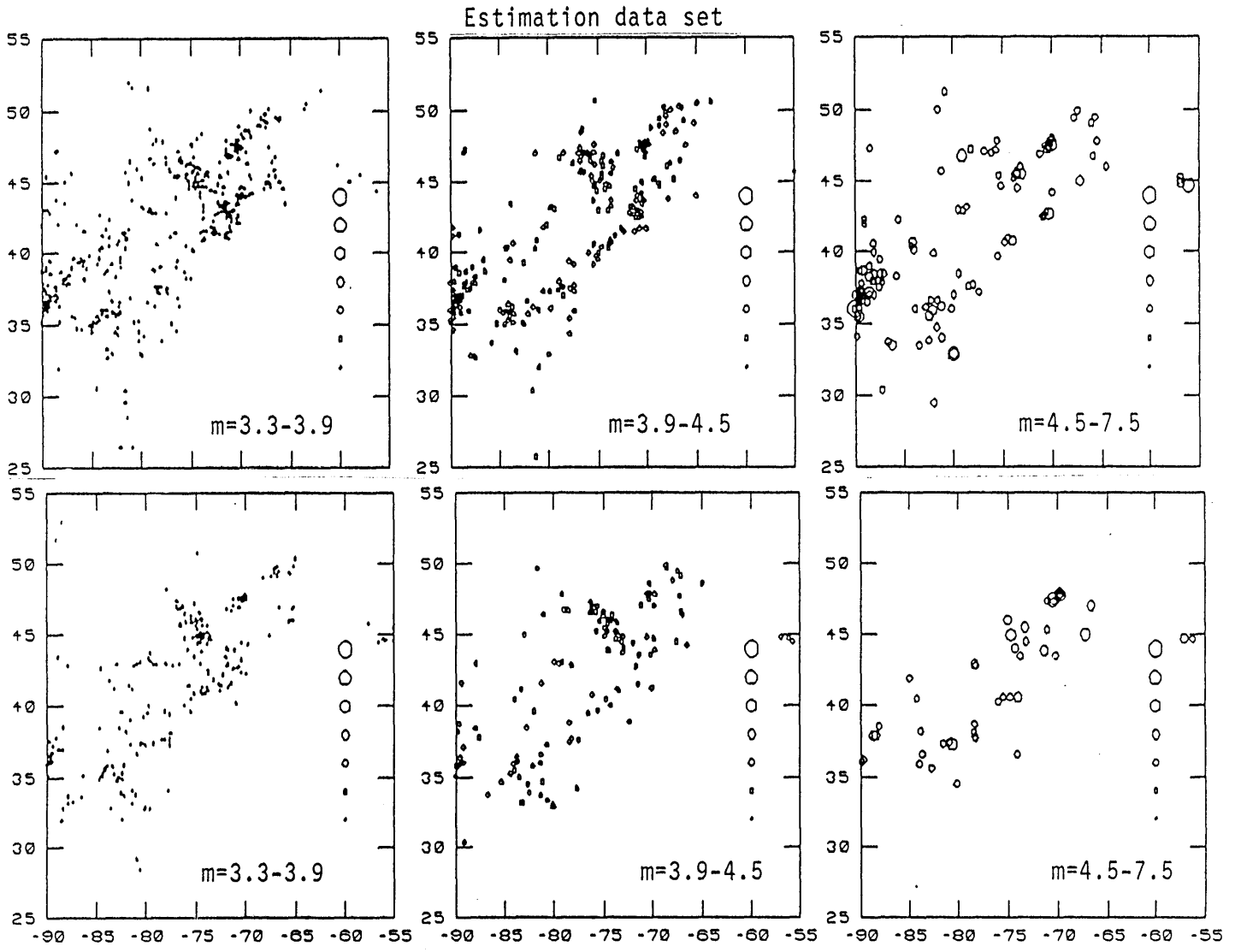


Figure 3-21: Spatial distribution of the total number of events in the estimation and validation subsets for the 3 partitions of the catalog.


Estimation data set
m=3.3-3.9
m=3.9-4.5
m=4.5-7.5
m=3.3-3.9
m=3.9-4.5
m=4.5-7.5
55
50
45
40
35
30
25
-90 -85 -80 -75 -70 -65 -60 -55
Validation data set
(b) last 15 years


Estimation data set
m=3.3-3.9
m=3.9-4.5
m=4.5-7.5
m=3.3-3.9
m=3.9-4.5
m=4.5-7.5
55
50
45
40
35
30
25
-90 -85 -80 -75 -70 -65 -60 -55
validation data set
(c) last 30 years



(d) (2/3,1/3) partition

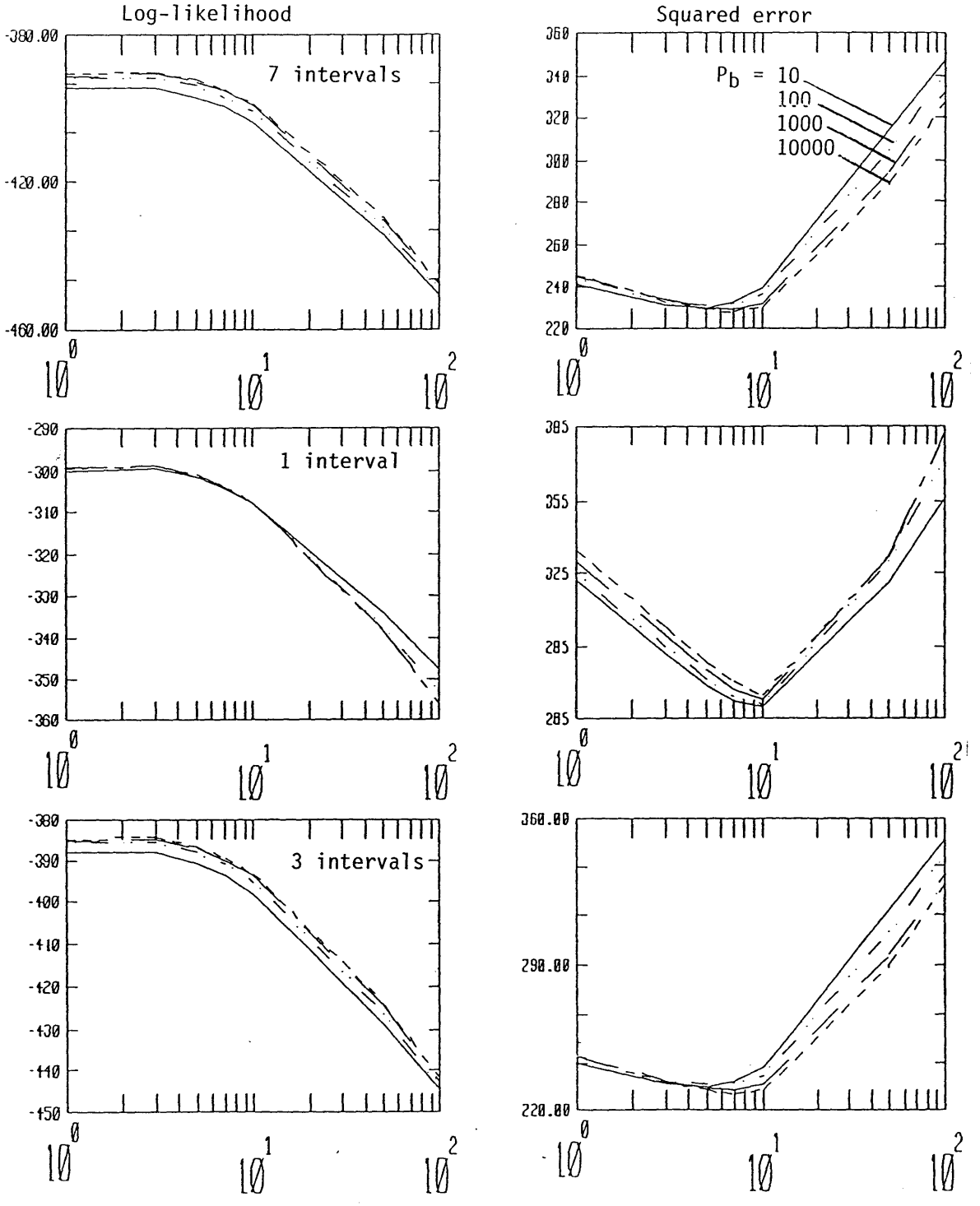
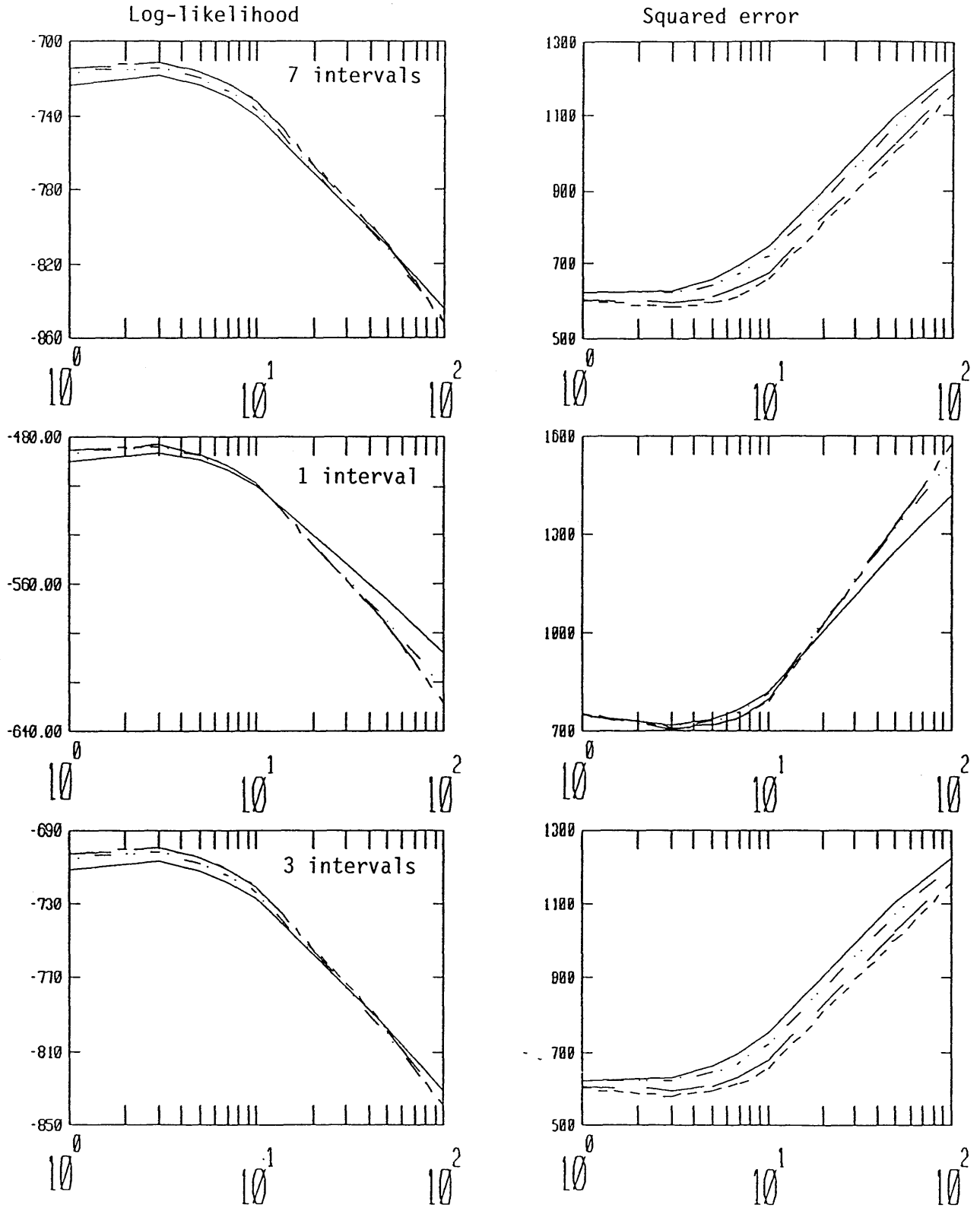
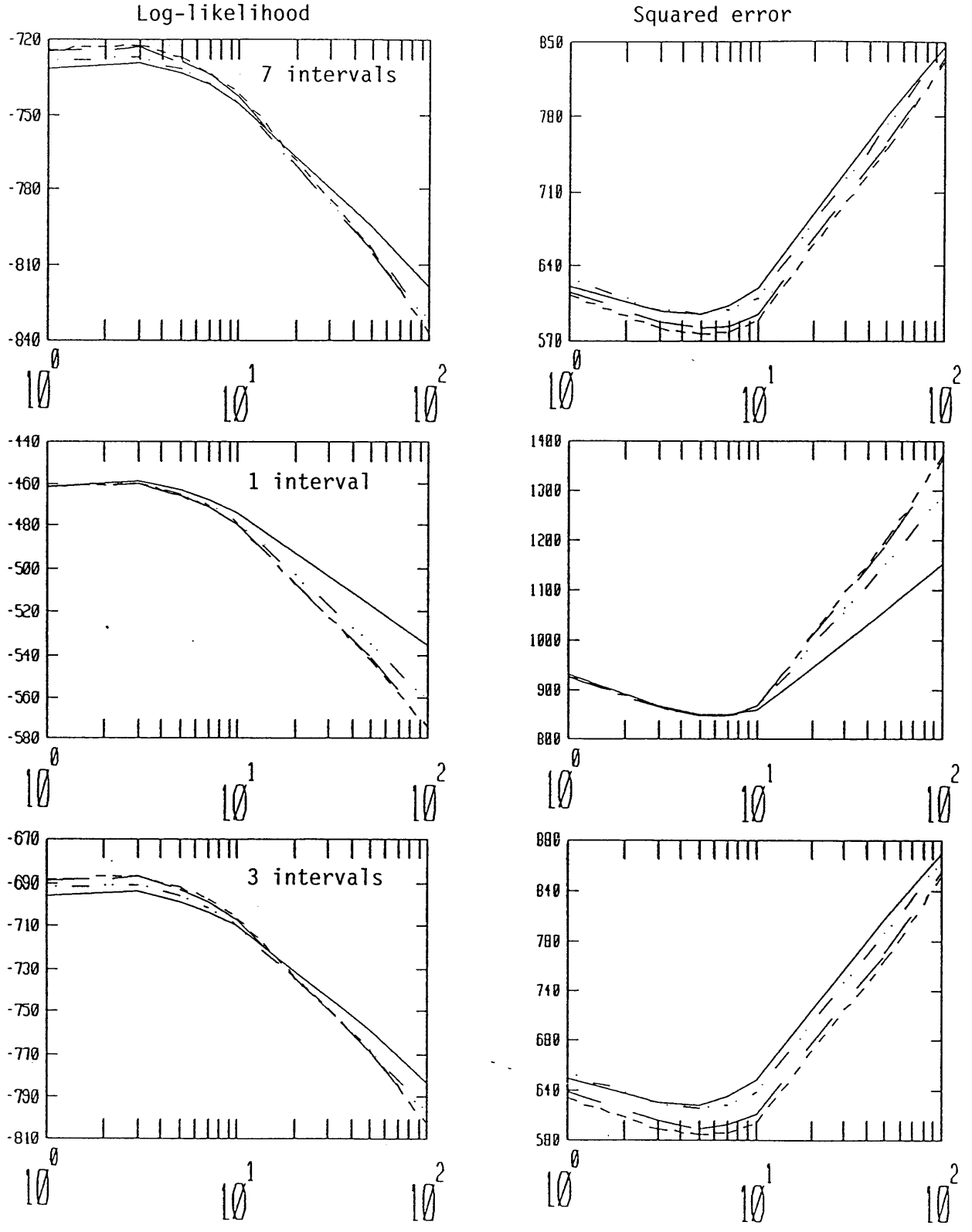


Figure 3-22: Selection of the optimal penalties P_a and P_b using cross-validation for different discretizations of magnitude and the 3 partitions of the catalog.



(b) last 30 years



(c) 2/3,1/3 partition

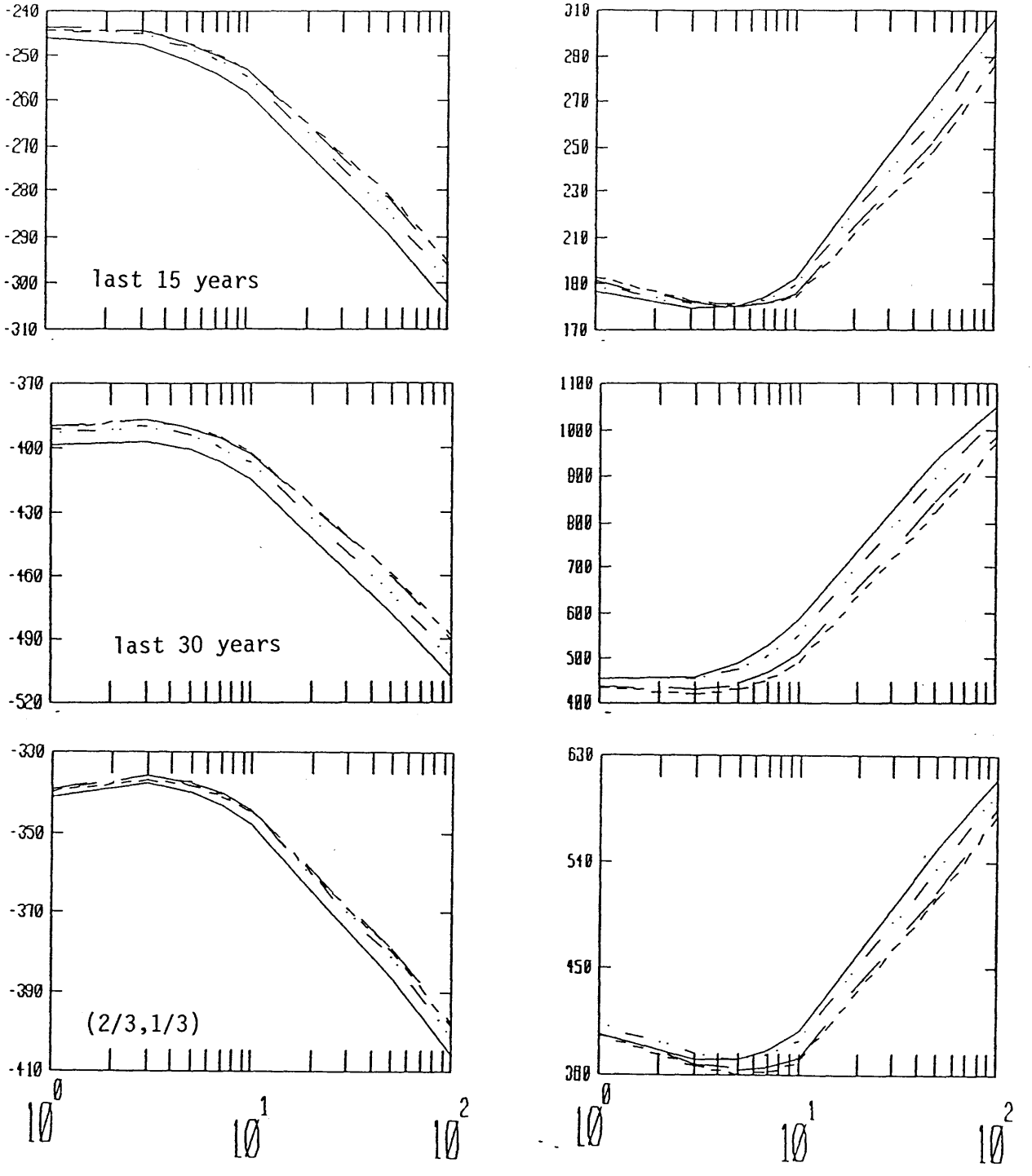


Figure 3-23: Selection of the optimal penalties P_a and P_b for low magnitude events ($3.3 \leq m \leq 3.9$) and the 3 partitions of the catalog.

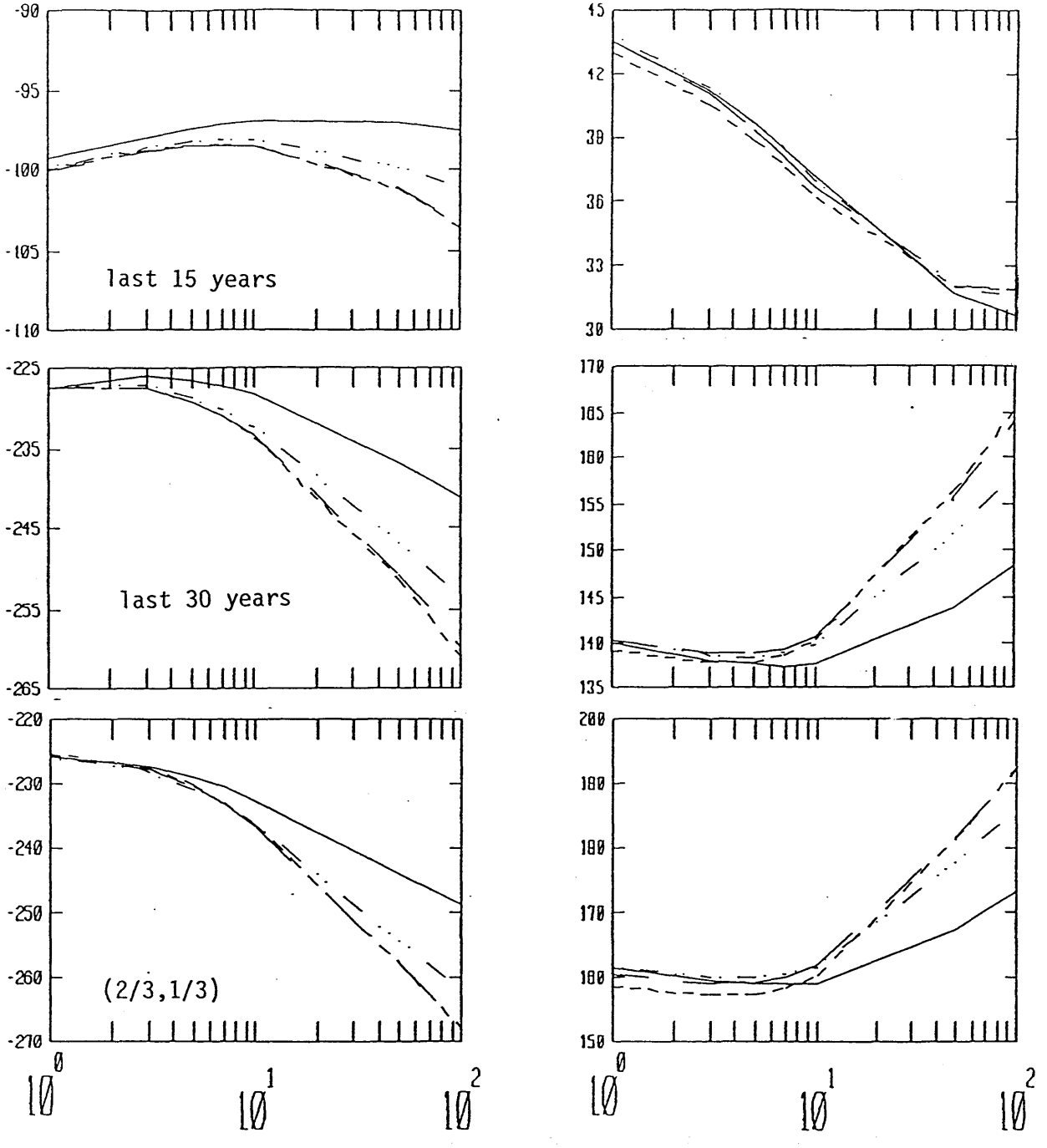


Figure 3-24: Selection of the optimal penalties P_a and P_b for intermediate magnitude events ($3.9 \leq m \leq 4.5$) and the 3 partitions of the catalog.

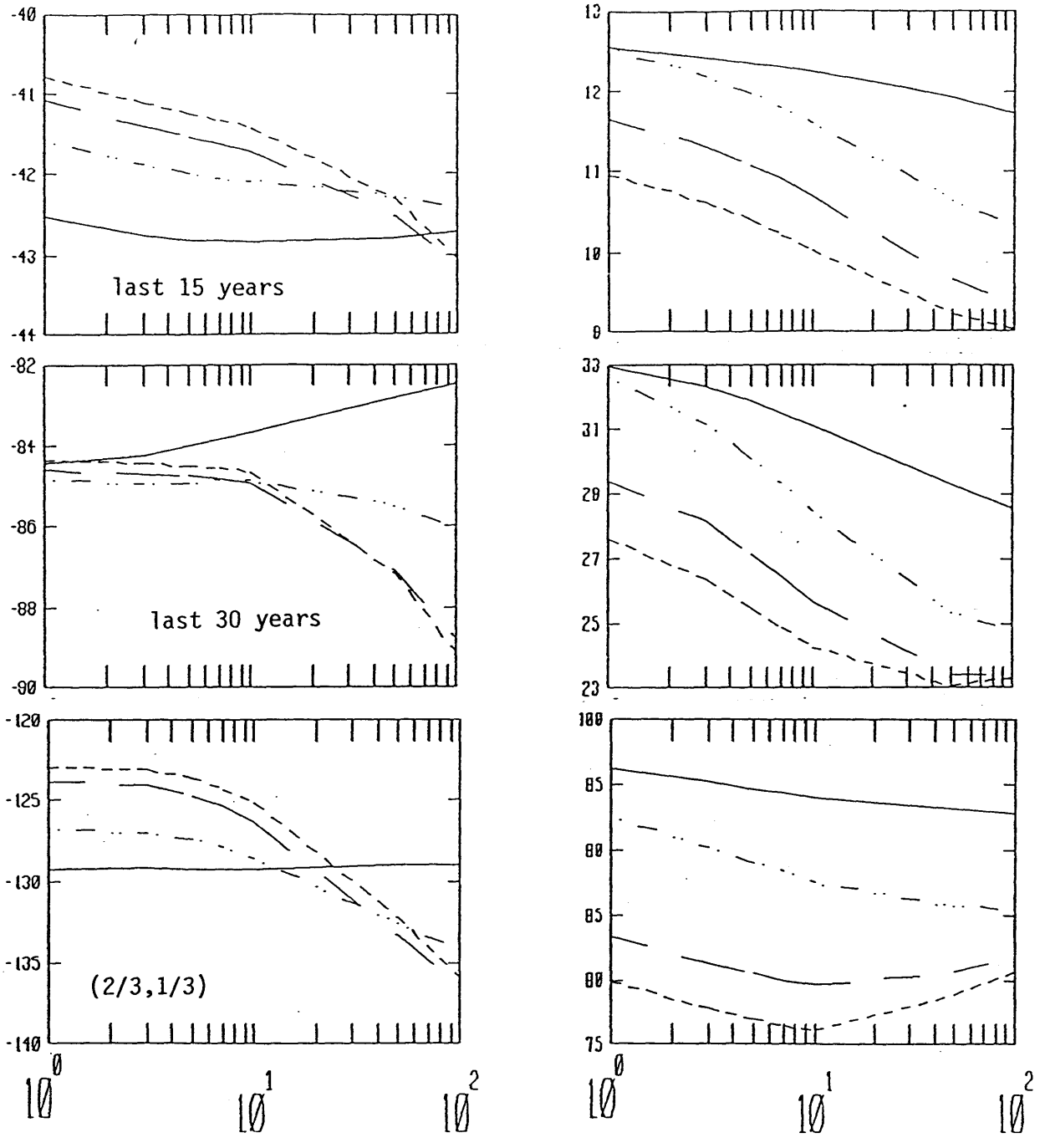


Figure 3-25: Selection of the optimal penalties P_a and P_b for large magnitude events ($4.5 \leq m \leq 7.5$) and the 3 partitions of the catalog.

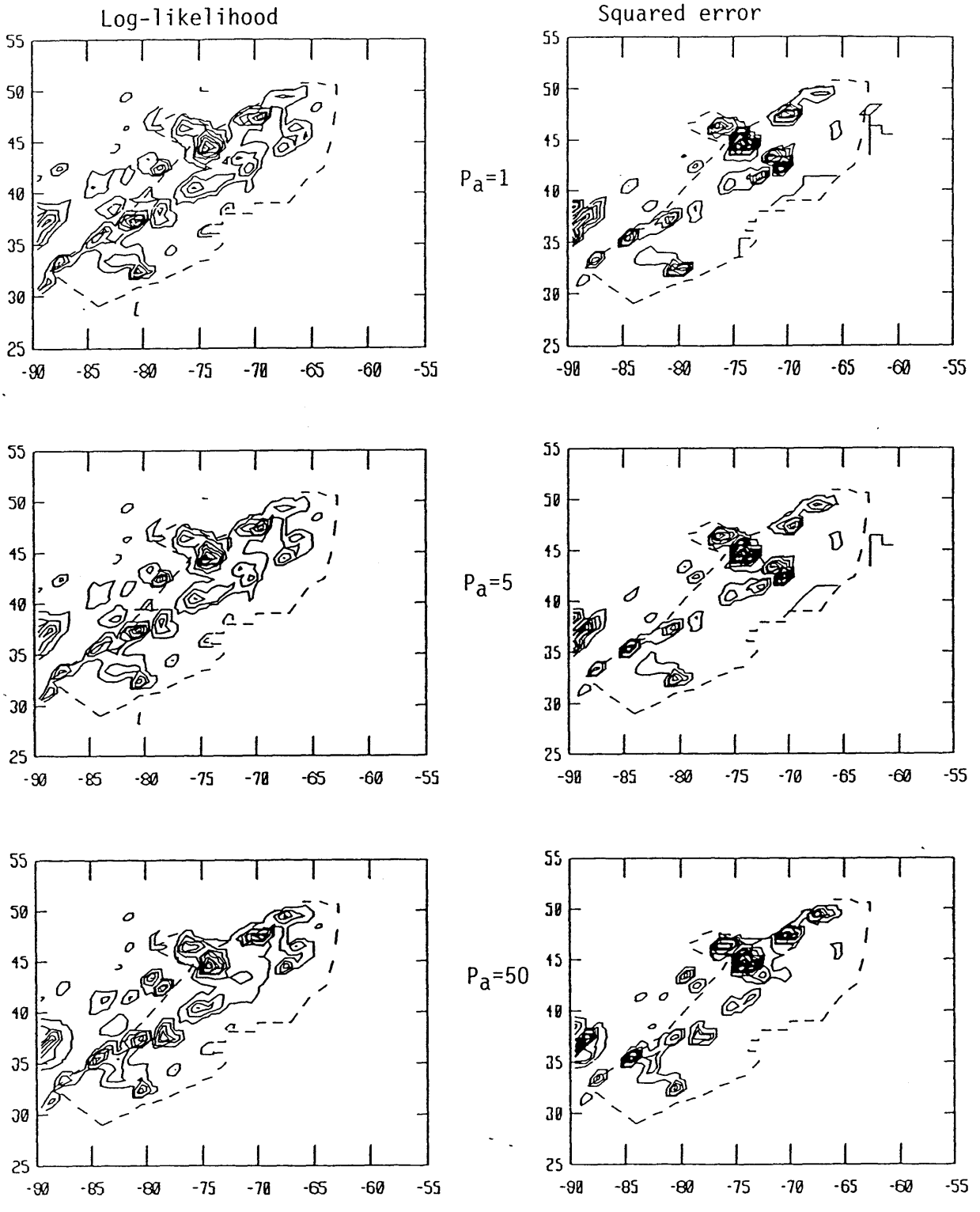


Figure 3-26: Decomposition of the cross-validated log-likelihood and squared error as a function of location for different penalties P_a (P_b=10000, (2/3,1/3) partition of the catalog).

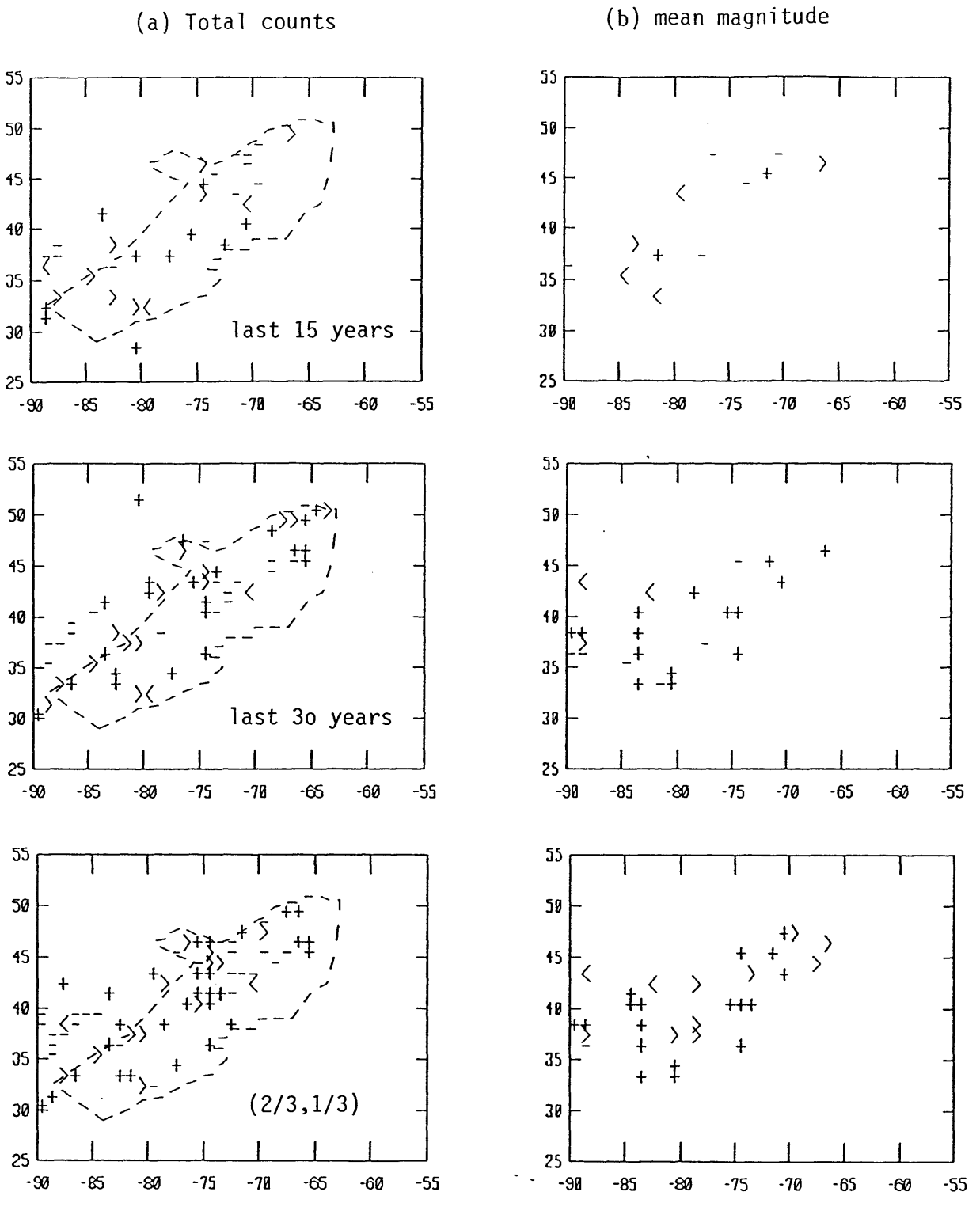


Figure 3-27: Flags for significant residuals for the total number of events and the total magnitude in each spatial cell.

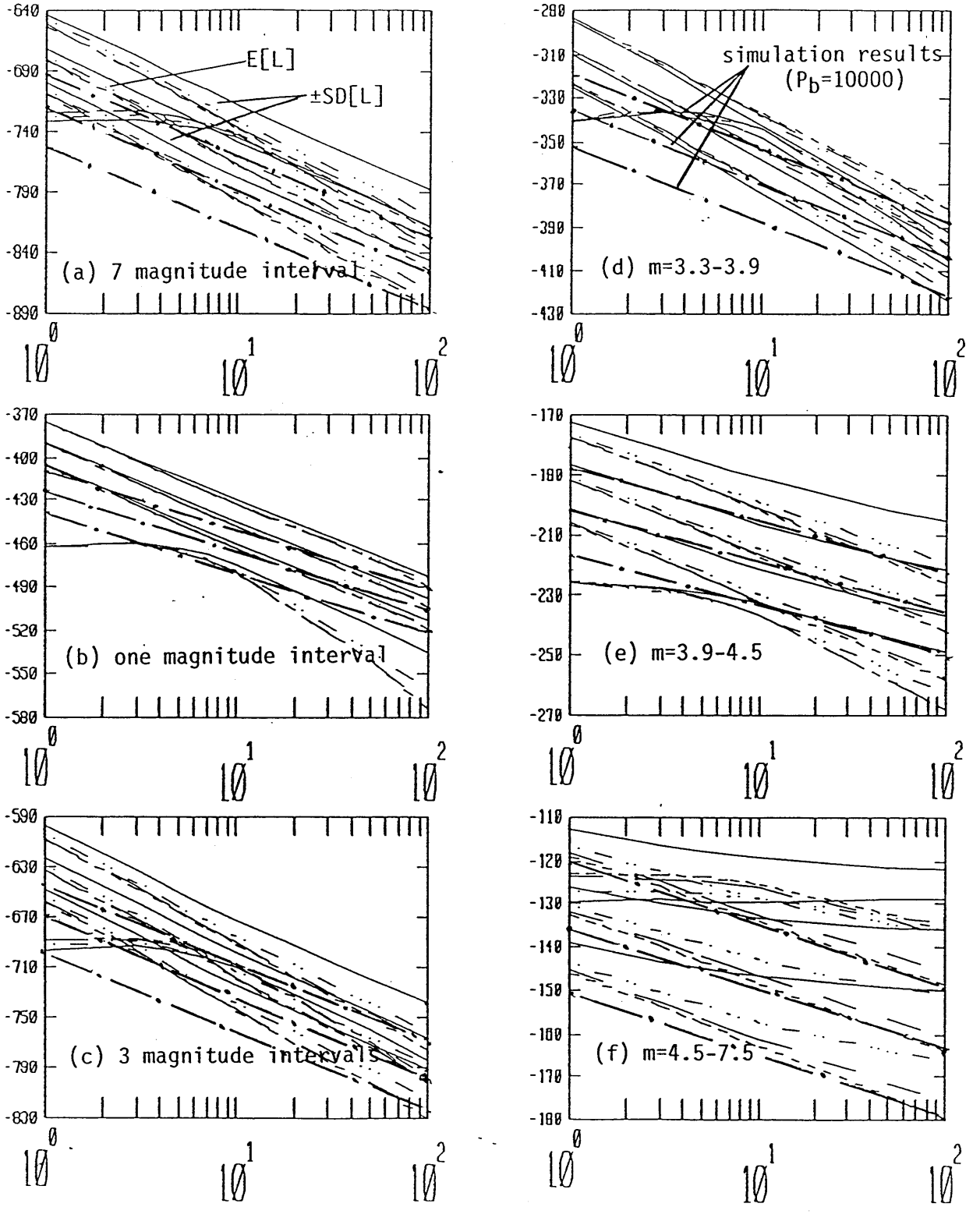


Figure 3-28: Cross-validated log-likelihood, its expected value and standard deviation for the (2/3,1/3) partition of the catalog, for different discretizations in magnitude.

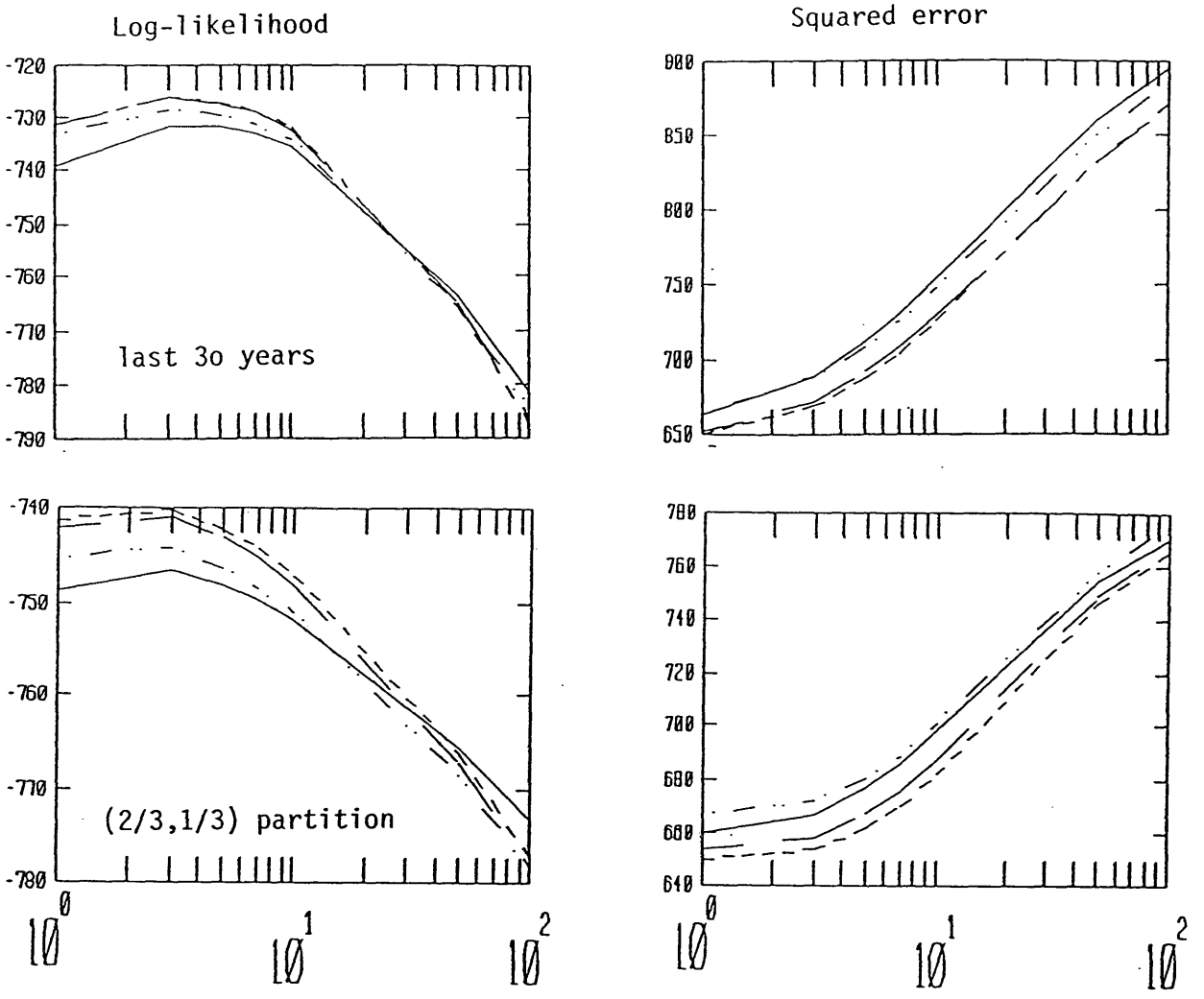


Figure 3-29: Selection of the penalty parameter P_a for the model with local neighborhoods (α=10 %).

Chapter 4

Combination of seismic source and historical estimates of earthquake hazard

4.1 Introduction

In the previous chapters, models of seismicity and estimation procedures were proposed. These procedures require a fair amount of computation but produce models which have optimal predictive characteristics. In this chapter, an alternative procedure is proposed which makes use of simpler historical and seismic source estimates of seismicity to produce estimates of seismic hazard which are equivalent to those from more sophisticated models of seismicity. The combined estiamtor is shown to be more precise than either the historical and seismic source estimators of hazard.

4.2 Characteristics of seismic-source and historical estimates of hazard

As was mentioned in section 2.1, a frequently used method for earthquake hazard estimation (Cornell , 1968) partitions the geographical region around the site of interest into provinces (sources) and assumes that, within each source, earthquakes occur according to a stationary and homogeneous Poisson process. Another frequent assumption, which is however not essential to the method, is that earthquake magnitude m has truncated exponential distribution, hence that earthquakes inside source i have a recurrence law of the type

$$\nu_i(m) = \begin{cases} 10^{a_i - b_i m} - 10^{a_i - b_i M_i}, & m \leq M_i \\ 0, & m > M_i \end{cases} \tag{4.1}$$

where $\nu_i(m)$ is the expected number of events per unit time and unit area with magnitude larger than m and M_i, a_i, and b_i are source-specific parameters.

For the calculation of earthquake hazard, one needs also know the attenuation law, i.e. the probability distribution of earthquake intensity at the site Y, as a function of the magnitude m and the location $\underline{x}_0$ of the earthquake. This attenuation law is written as

$$Y = g(m,\underline{x}_o,\varepsilon) \tag{4.2}$$

where ε is a random variable.

From the recurrence model in Equation 4.1 and the attenuation law in Equation 4.2 one can calculate the exceedance rate function (seismic hazard function) at the site, ($\lambda(y)$ = rate of events with intensity higher than y)

$$\lambda_{SS}(y) = \int_{\underline{x}} d\underline{x} \int_m \nu_i(m) P[Y>y|\underline{x},m] \; dm \tag{4.3}$$

The above method of seismic hazard analysis is called the seismic-source (SS) method and $\lambda_{SS}(y)$ denotes the associated estimator of $\lambda(y)$.

As an alternative to the SS method, one may use historic (H) procedures. These procedures estimate $\lambda(y)$ directly from a catalog of historic events, for example, as

$$\lambda_H(y) = \frac{1}{T} \sum_{\text{historic events, } i} \frac{1}{P_D(m_i,\underline{x}_{o_i})} P[Y_i > y \mid m_i,\underline{x}_{o_i}] \tag{4.4}$$

where T is the time period covered by the catalog and $P_D(m,\underline{x}_0)$ is the probability that a generic event in T with characteristics $(m,\underline{x}_0)$ is recorded in the catalog. Hence, the product $T\bullet P_D(m,\underline{x}_0)$ is the equivalent period of complete recording at location $\underline{x}_0$, for events with magnitude m.

The relative accuracy of the estimators $\lambda_{SS}(y)$ and $\lambda_H(y)$ depends on the value of y. One should notice in particular that $\lambda_H(y)$ is a nonparametric estimator and is unbiased, irrespective of the spatial variation of earthquake activity and of the probability distribution of magnitude. The variance of $\lambda_H(y)$ is small at low intensities, but it becomes large at high intensities, especially for values of y such that $\lambda(y)<1/T$. The estimator $\lambda_{SS}(y)$ has a smaller variance. However, if the geometry of the earthquake

sources or the type of magnitude distribution are incorrectly specified, $\lambda_{SS}(y)$ is biased. The net result is that, in typical applications, the mean squared error MSE (variance plus squared bias) of λ_H is smaller than that of λ_{SS} for small y, whereas the reverse is true for large y.

Figure 4.1 shows the historical estimates of Equation 4.8 for a site at location (74°W, 45°N). Two cases are illustrated, one with a median attenuation (or equivalently with $\sigma_\varepsilon=0$ in Eq. 4.2) (Figure 4.1a) and the other with a random attenuation function ($\sigma_\varepsilon=0.6$) (Figure 4.1b). Also showm are decompositions of hazard as a function of distance from the site and as a function of epicentral intensities. With respect to location, most of the hazard is contributed by seismicity within 200 km of the site. Remote events only contribute to the hazard for low site intensities. The hazard at higher site intensities is contributed mostly by closely located large events indicating the importance of a properly specified model with respect to the large magnitude events (both in terms of rate and of maximum epicentral intensity). Seismic hazard results are shown for the complete (full line) and the incomplete (dashed line) hazard functions in Figure 4.1d. Incompleteness is shown to be large (about 30%) for small site intensities and decreases momotonically with larger intensities.

The hazard functions are also decomposed for the seismic-source estimator (Figure 4.1c) for the source configuration of Figure 4.4. The decomposition with respect to epicentral intensities shows that the individual hazard functions have a similar shape but are shifted vertically and horizontally with respect to each other. Similarity of the shapes indicates that seismicity is identically distributed in space as a function of epicentral intensity, The amount of vertical shift of the hazard function decomposed for each epicentral intensity results from the assumption of exponentiality and the local estimates of $b(\underline{x})$. The latter parameter in combination with the maximum epicentral intensity also control the slope of the total hazard function at larger site intensities. In

the case of the historical estimates, equal vertical spacing of the decomposed hazard for low site intensities (all events contribute to the hazard) indicates that the exponential model holds globally except maybe for events with I_0=4.5-5.5 (Figure 4.1b). For larger site intensities, there is evidence that the exponential model may not hold in the vicinity of the site or equivalently, that there is a change in the spatial distribution of epicentral events as a function of intensity because of the change in the spacing of the curves. Finally, the influence of the source configuration on the estimates of $\lambda(y)$ can be very large, Figure 4.1e illustrates extreme source configurations for which seismic hazard may vary by as much as a factor of 4 depending on the location.

It is proposed to use $\lambda_{SS}(y)$ and $\lambda_H(y)$ in combination to form estimators $\lambda_{SS\text{-}H}(y) = C\bullet \lambda_{SS}(y)$ that are more accurate than either λ_{SS} or λ_H over the (high) intensities of interest for earthquake risk assessment. The basic idea is to choose the constant C such that, at some low intensity, the combined estimate coincides with the historical estimate. Different definitions of C produce different combined estimators. The estimators $\lambda_{SS\text{-}H}(y)$ are suggested as practical alternatives to more sophisticated local estimators $\lambda_L(y)$ that result from allowing the parameters a and b in Equation 4.1 to vary in space within each earthquake source (see Chapter 2 and 3). In order to evaluate the combined estimators, λ_{SS} and $\lambda_{SS\text{-}H}$ with λ_L are compared at many sites in the northeastern U.S.

4.3 Combined estimators

The estimators $\lambda_{SS}(y)$ and $\lambda_H(y)$ give the rate at which any specified intensity y is exceeded at the site. Both estimators make corrections for catalog incompleteness and account for attenuation uncertainty. For one of the combined estimators introduced below, the calibration factor C is defined in terms of the functions λ_{SS} and λ_H , but other combined estimators studied here require calculation of the hazard for the

incomplete earthquake sequence and for the case when the random attenuation law $g(m,\underline{x}_0,\varepsilon)$ is replaced with the median law $\bar{g}(m,\underline{x}_0)$. the function $\bar{g}$ is such that earthquakes with characteristics $(m,\underline{x}_0)$ produce site intensities above and below $\bar{y} = \bar{g}(m,\underline{x}_0)$ with equal probability. The hazard estimators for the case of incomplete catalog and median attenuation are denoted by $\lambda'_L(y)$, $\lambda'_{SS}(y)$, and $\lambda'_H(y)$, depending on the method of estimation. The first two such estimators are given by

$$\lambda'_L(y) \text{ or } \lambda'_{SS}(y) = -\int_{\underline{x}_o}\int_m P_D(m,\underline{x}_o)\frac{\partial \nu(m,\underline{x}_o)}{\partial m}K(y,m,\underline{x}_o)dm\, d\underline{x}_o \tag{4.5}$$

where $\nu(m,\underline{x}_0)$ is the function in Equation 4.5 at the geographical point $\underline{x}_0$ and $K(y,m,\underline{x}_0)$ is an indicator function with value 1 if $\bar{g}(m,\underline{x}_0) > y$ and value 0 otherwise. The function $\nu(m,\underline{x}_0)$ is estimated locally for λ'_L (Chapter 2) and is found under the assumption of homogeneous seismic sources for λ'_{SS}.

Consistently with Equation 4.8, the estimator λ'_H might take the form

$$\lambda'_H(y) = \frac{n(y)}{T} \tag{4.6}$$

where n(y) is the number of historic events with median attenuated intensity $\bar{y}_i = \bar{g}(m_i,\underline{x}_{0_i})$ in excess of y. Another possibility is to use

$$\lambda''_H(\bar{y}_i) = \lambda''_{H_i} = \frac{i}{T}\,\frac{n}{n+1} \tag{4.7}$$

This last estimator is defined only at the median historic intensities $\bar{y}_i$, which are ordered such that $\bar{y}_1 > \bar{y}_2 > \ldots > \bar{y}_n$.

The estimators λ_{SS}, λ_H, λ'_{SS}, λ'_H, and λ''_H are used to form three combined estimators of $\lambda(y)$:

$$\lambda^{(1)}_{SS-H}(y) = \frac{\lambda_H(y^*)}{\lambda_{SS}(y^*)}\lambda_{SS}(y) \tag{4.8}$$

$$\lambda^{(2)}_{SS-H}(y) = \frac{\lambda'_H(y^*)}{\lambda'_{SS}(y^*)}\lambda_{SS}(y) \tag{4.9}$$

$$\lambda^{(3)}_{SS-H}(y) = \frac{\lambda''_H(y^*)}{\lambda'_{SS}(y^*)}\lambda_{SS}(y) \qquad (4.10)$$

In all cases, y^* is a calibration intensity, which is chosen as described next.

4.4 Choice of the calibration intensity and evaluation of the combined estimators

The calibration should not be done at low recurrence rates. First, there is a lot of uncertainty on the amount of incompleteness in that range, and the hazard is contributed mostly by small or distant events which are only remotely related to the events that contribute to the hazard in the range of interest for seismic design. In addition, the incomplete historical and seismic source hazard functions both converge asymptotically to N/T for very low events, where N is the total number of events in the catalog and T is the total period of observation for the catalog. Similarly, the calibration should not be done with respect to the smaller historical rates because of the large uncertainty due to the small sample size. Figure 4.2 shows the variation of the calibration C as a function of the site intensity for a site located at (72°W,45°N) and the estimator of Equation 4.10. The calibration measures the vertical seperation of the historical and seismic source hazard functions as a function of y. Also shown are one standard deviation envelopes obtained using a Gamma distribution with parameter N(y) where N(y) is the number of historical events with site intensities smaller than y. Note that the calibration is equivalent to a local adjustment of $a(\underline{x})$ through the addition or substraction of a constant term Δa at each location surrounding the site, and does not affect the shape of the seismic source estimate of the hazard function.

It is convenient not to specify y^* externally and rather set y^* equal to one of the order statistics $\bar{y}_i$. [This is a necessity for the estimator $\lambda^{(3)}_{SS-H}$, which is defined only at the points $\bar{y}_i$.] The criterion used to select the calibration intensity $\bar{y}_i$ is to minimize the mean squared error of the log exceedance rate, which for the k^{th} combined estimator (k=1,2,3) is

$$MSE^{(k)}(y;i) = E\{[Log\lambda^{(k)}_{SS-H}(y;y^* = \bar{y}_i) - Log\lambda_L(y)]^2\} \quad (4.11)$$

In practice, it is impossible to calculate the mean squared error in Equation 4.11, because only one earthquake catalog is available for a given region. One might resort to Monte Carlo simulation and for example assume that the earthquake process is Poisson with the recurrence law used in the calculation of λ_L. A drawback of the simulation method is that the geometry of the sources usually reflects the spatial distribution of historical seismicity. Therefore, one should redefine the sources for each simulation.

It was found preferable to replace the expectation in Equation 4.11 with the average of the squared log error over a grid of sites. Regional variations of seismicity are further accounted for by setting y to the intensity that is exceeded at each site with a given frequency; i.e. we fix $\lambda_L(y)= \lambda$ and minimize with respect to i the quantity

$$ASE^{(k)}(\lambda;i) = \text{spatial average of } [Log\,\lambda^{(k)}_{SS-H}(y;y^*) = \bar{y}_i) - Log(\lambda)]^2 \quad (4.12)$$

where y is an intensity that varies from site to site and satisfies $\lambda_L(y)=\lambda$. Figure 4.3 illustrates the calculation of $ASE^{(1)}$. Similar procedures apply to $ASE^{(2)}$ and $ASE^{(3)}$.

Numerical results are obtained using the (Chiburis, 1981) catalog for the northeastern U.S. in the region (39-46°N, 69-77°W). A plot of main events for the period 1627-1981 is shown in Figure 4.6. Because for most of the large earthquakes the only available size measure is MM epicentral intensity I_o, in all calculations I_o is used in place of m and magnitude is converted when needed using the formula proposed by Chiburis, $I_o=(m-1)/0.6$. The maximum possible intensity, which is the equivalent for I_o of M_i in Equation 4.1, is taken everywhere to be IX-X.

The seismicity model for $\lambda_L(y)$ and $\lambda'_L(y)$ has spatially varying a nad b coefficients,

shown in Figure 4.5. These coefficients have been obtained through the local neighborhood method described in Chapter 2, and refer to a recurrence relationship of the type

$$\nu(\underline{x},m) = 10^{a(\underline{x})-b(\underline{x})(I_o-3.5)}-10^{a(\underline{x})-b(\underline{x})(9.5-3.5)}. \quad (4.13)$$

The unit area in the definition of a is that of a square equatorial degree (i.e. (111.11 km)^2).

For $\lambda_{SS}(y)$, two alternative source configurations are considered : In one case the region is partitioned into 11 sources, which closely reflect the spatial variation of historical seismicity. The sources are shown in Figure 4.6, which is an adaptation from Figure 1 of (WGC, 1980). In the other case a simple homogeneous source is used for the entire region. The latter assumption is unrealistic, but is useful to generate an upper bound to $ASE^{(K)}$ over all reasonnable choices of the seismic sources and to compare the robustness of the estimators λ_{SS} and $\lambda^{(K)}_{SS\text{-}H}$ with respect to source geometry.

In all calculations, y is taken to be the peak ground acceleration (cm/sec^2) and the attenuation law is that proposed by (Heidari, 1987) for peak horizontal acceleration on rock in the eastern and central U.S., i.e.

$$y = exp\{2.00 + 1.14m_{Lg} - 1.03lnR - 0.003R+\varepsilon\} \quad (4.14)$$

where R is hypocentral distance in kilometers for a focal depth of 10 kilometers and ε is a normal random variable with zero mean, standard deviation 0.6, and symmetrical truncation at $\pm$ 1.8. The L_g magnitude is obtained from I_0 using $m_{L_g} = 1+0.6I_0$ and median attenuated values are generated by setting ε=0.

The squared error is averaged over the 19 sites shown as stars on the grid of Figure 4.4 and calculations are repeated for $\lambda=10^{-2}$, 10^{-3}, and 10^{-4} events/year. Other sites of the grid are excluded from averaging, because the historical seismicity at those sites does not conform to the assumptions of the model. Lack of fit of the model has been detected by applying the Kolmogorov-Smirnov test at a significance level of 10% to

the median historic intensities $\bar{y}_i$, regarded as a random sample from the Poisson process with exceedance rate $\lambda'_{SS}(y)$ in Equation 4.5 (11 source solution). The tests were performed with respect to the upper-tail of the seismic hazard functions to minimize the effect of uncertainty on incompleteness. Table 4.I shows the empirical rate at which normalization is performed, the maximum separation between the two functions, the site acceleration at which it occurs, and the associated exceedance probability. Note that the test is mainly a test of goodness-of-fit with respect to $b(\underline{x})$. Figure 4.7 shows the upper-tails of the historical and seismic source (incomplete) estimates of seismic hazard. Significant differences occur at sites (76°W,45°N); 72°W,45°N; 76°W,43°N; 73°W,43°N; 76°W,42°N; 75°W,42°N; 74°W,42°N; 73°W,41°N) which are typically at the boundary between active and less active regions. When the historical estimates are larger than the seismic source estimates (i.e. the historical probability of exceedance for a given site intensity is smaller than what is predicted), the b parameter is locally overestimated by the seismic source model. Note that the larger (and more uncertain) events do not influence the outcome of the test because they are located in the upper tail of the distribution.

Results are presented in Figure 4.8a for the 11-sources configuration and in Figure 4.8b for the single-source case. For each combination of seismic source geometry and exceedance rate, the average squared errors in Equation 4.12 are plotted against i (and against λ''_{H_i} in Equation 4.7, where for the present catalog T=354 years and n=423) and are compared with the average squared error of the seismic source estimator,

$$ASE_{SS}(\lambda) = \textit{Spatial average of } [Log\lambda_{SS}(y) - Log\lambda]^2 \qquad (4.15)$$

In analogy with Equation 4.12, the intensity y in Equation 4.15 varies from site to site to satisfy $\lambda_L(y)=\lambda$. Notice that ASE_{SS} in Equation 4.15 does not depend on the calibration intensity $\bar{y}_i$ and therefore plots in Figure 4.8 as a horizontal line.

Figure 4.8 indicates that the combined estimators $\lambda^{(2)}_{SS\text{-}H}$ and $\lambda^{(3)}_{SS\text{-}H}$ have similar

average squared errors with respect to λ_L and are better than λ_{SS} if the calibration is chosen appropriately. The optimum value of i, i^* decreases slightly (the calibration intensity increases slightly) as the rate λ at which hazard is estimated decreases. Also, i^* is slightly smaller (the calibration intensity is slightly higher) for a poorer choice of the earthquake sources. These variations as well as the variation of i^* with the type of combined estimator, are however small and one may in all cases use a value of i around 15, which corresponds for the present catalog to a historical exceedance rate λ''_{H_i} of about one event in 25 years. Over different seismicity conditions, the optimum value of i is expected to remain stable and the optimum calibration rate is expected to vary as 15/T, where T is the period covered by the catalog.

The estimator $\lambda^{(1)}_{SS\text{-}H}$ is slightly less accurate than either $\lambda^{(2)}_{SS\text{-}H}$ or $\lambda^{(3)}_{SS\text{-}H}$, but is still superior to λ_{SS}, especially for low prediction rates and poor source configurations. The best value of i for $\lambda^{(1)}_{SS\text{-}H}$ is somewhat smaller than for the other combined estimators, but the choice i^*=15 is still nearly optimal.

Table 4.II gives estimates of the error factors

$$EF^{(k)} = 10^{(ASE^{(K)})^{0.5}} \tag{4.16}$$

which expresses the degree of dissimilarity between $\lambda^{(k)}_{SS\text{-}H}$ and λ_L and the analogous error factor for λ_{SS}. Different values are given for accurate and poor source configurations, by which is meant source geometries that respectively do and do not reflect the spatial distribution of historic seismicity. The values for poor configurations are intermediate between those of Figures 4.8a and 4.8b, in consideration of the very crude assumption of complete homogeneity in Figure 4.8b. An important conclusion from Table 4.II and Figure 4.8 is that the combined estimators $\lambda^{(k)}_{SS\text{-}H}$ are more robust than λ_{SS} with respect to the specification of the earthquake sources. Therefore, combined estimators reduce the consequences of errors in the source configuration and are particularly recommended when the interpretation of historical seismicity is controversial or when homogeneous earthquake sources do not exist.

The previous analysis is based on the comparison of various hazard estimators with the local estimator λ_L. In the following section, a semi-theoretical analysis is made of the error of $\lambda^{(3)}_{SS\text{-}H}(y)$ with respect to the true hazard $\lambda(y)$. This analysis indicates that the optimum calibration rate $\bar{y}_i$ is probably closer to $\bar{y}_{(20)}$ than to $\bar{y}_{(15)}$ and that the error factors of the combined estimator $\lambda^{(3)}_{SS\text{-}H}$ with respect to the true hazards are about 10% higher than the values reported in Table 4.II.

4.5 Mean squared error of the combined estimators with respect to the true rate

In the previous section, different hazard estimators $\hat{\lambda}(y)$ were compared on the basis of the difference between Log $\hat{\lambda}(y)$ and the logarithm of the local hazard estimator $\lambda_L(y)$. the justification for this criterion is that $\lambda_L(y)$ is an accurate estimator of the true hazard function $\lambda(y)$. In reality, λ_L is itself random and is positively correlated with all the other estimators $\hat{\lambda}(y)$, because all estimators use the same earthquake data.

Here, some results are derived for the mean squared error of $\lambda^{(3)}_{SS\text{-}H}$, when the error is defined as the difference between Log $\lambda^{(3)}_{SS\text{-}H}(y)$ and the logarithm of the true hazard, $\lambda(y)$. Hence the interest is in

$$MSE^{(3)}_T(\lambda;i) = E\{[Log\lambda^{(3)}_{SS-H}(y;y^*=\bar{y}_i)-Log\lambda]^2\} \qquad (4.17)$$

where y is such that $\lambda(y)=\lambda$ and the subscript T denotes true. Analogous quantities for $\lambda^{(1)}_{SS\text{-}H}$ and $\lambda^{(2)}_{SS\text{-}H}$ in Equation 4.4 are tedious to calculate, but they should be close to $MSE^{(3)}_T$.

First, the logarithmic difference in Equation 4.17 is written as the sum of two terms,

$$\Delta_a(\bar{y}_i) = Log\lambda''_H(\bar{y}_i) - Log\lambda'(\bar{y}_i) \qquad (4.18)$$

and

$$\Delta_b(y,\bar{y}_i) = [Log\lambda_{SS}(y)-log\lambda] - [Log\lambda'_{SS}(\bar{y}_i)-Log\lambda'(\bar{y}_i)] \qquad (4.19)$$

so that Equation 4.17 becomes

$$MSE_T^{(3)}(\lambda;i) = E\{[\Delta_a(\bar{y}_i)+\Delta_b(y,\bar{y}_i)]^2\} = (m_a+m_b)^2 + \sigma_a^2 + \sigma_b^2 + \rho\sigma_a\sigma_b \quad (4.20)$$

where m_a and σ_a^2 are the mean and variance of $\Delta_a(\bar{y}_i)$, m_b and σ_b^2 are the mean value and variance of $\Delta_b(y,\bar{y}_i)$ and ρ is the correlation coefficient betweem $\Delta_a(\bar{y}_i)$ and $\Delta_b(y,\bar{y}_i)$. The term $\Delta_a(\bar{y}_i)$ is the error of the historic estimator λ''_H in Equation 4.7 at the calibration intensity $\bar{y}_i$ and the term $\Delta_b(y,\bar{y}_i)$ is the error of prediction of the seismic source estimator if λ'_{SS} is calibrated to the exact incomplete rate λ'. The first two moments of these errors, which are needed for the calculation of $MSE_T^{(3)}$, are obtained in a semi-empirical way, as follows:

The mean value m_a and the variance $\sigma^2{}_a$ can be calculated theoretically : The error $\Delta_a(\bar{y}_i)$ is random because $\lambda'(\bar{y}_i)$ is random. This incomplete rate can be written as

$$\lambda'(\bar{y}_i)=\lambda'_o[1-F'(\bar{y}_i)] \quad (4.21)$$

where λ'_o is the total rate of events for the incomplete catalog and F' is the cumulative distribution function of site intensity for the generic event of the same catalog. The total rate λ'_o may be considered known with value N/T, where N is the total number of events in the catalog and T is the period of recording. Therefore, the term $[1-F'(\bar{y}_i)]$ is the only important source of randomness for $\Delta_a(\bar{y}_i)$. The distribution of $[1-F'(\bar{y}_i)]$ is known to be Beta, with parameters (i,N-i+1); see (Johnson, 1970), p38. This result can be used to calculate m_a and σ_a^2 for given N,T, and i.

Theoretical calculation of the other terms in Equation 4.20 is much more complicated. For them, numerical estimation was used under the assumption that Δ_b should not vary much if one replaces the true rates $\lambda(y)=\lambda$ and $\lambda'(\bar{y}_i)$ in Equation 4.19 with the corresponding local estimates $\lambda_L(y)$ and $\lambda'_L(\bar{y}_i)$. $(\bar{y}_i)$. With this replacement, m_b, σ_b^2 and ρ were obtained as sample values from the 19 sites used earlier to rank various estimators; see Figure 4.4. The quantities

$$E[\Delta_a^2(\bar{y}_i)] = m_a^2 + \sigma_a^2 \quad (4.22)$$

$$E[\Delta_b^2(y,\bar{y}_i)] = m_b^2 + \sigma_b^2$$

$$MSE_T^{(3)}(\lambda;i) = E\{[\Delta_a(\bar{y}_i) + \Delta_b(y,\bar{y}_i)]^2\}$$

are plotted in Figure 4.9. For comparison, the last quantity in Equation 4.22 for the case when $\rho=0$ and the averaged squared error $ASE_{SS\text{-}H}^{(3)}$ in Equation 4.12 are also shown.

As one would expect considering the correlation between $\lambda_{SS\text{-}H}^{(3)}$ and λ_L, the mean squared error in Equation 4.17 is larger than the average squared error $ASE_{SS\text{-}H}^{(3)}$. The difference between the two quantities increases with decreasing calibration rate (with increasing i). As a consequence, the value of i that minimizes $MSE_T^{(3)}$ in Equation 4.17 increases, from about 15 to about 30 for the 11-source case. For the case of a single source, the optimum value of i remains around 15-20. The increase in the optimal calibration rate for the 11-source case is probably exaggerated by the fact that the replacement of $\lambda(y)$ and $\lambda'(y)$ with $\lambda_L(y)$ and $\lambda_L'(y)$ reduces the value of $E[\Delta_b(y,\bar{y}_i)]$ and hence increases the optimum value of i. In consideration of this fact, a calibration value of i around 20 is recommended.

The correlation ρ is small in the case of 11 sources, but is non-negligible and negative in the 1 source case, adding to the robustness of the estimator.

The negative correlation is caused by the difference in the slope parameter b among various seismicity models. The parameter b has a direct influence on the slope of the hazard curves at the site. Because of the various degrees of spatial smoothing of b, one typically observes that the slope of the local hazard estimator λ_L is intermediate between the slope of λ_{SS} for a single source and the slope of the historical hazard λ_H. Figure 4.10 illustrates typical situations and the resulting negative correlation between Δ_a and Δ_b.

4.6 Conclusions

Historic estimates of earthquake hazard, $\lambda_H(y)$, have the desirable properties of being unbiased, of requiring little external information, and of being accurate at low intensities. However, for the high intensities of interest in earthquake risk mitigation, estimators $\lambda_{SS}(y)$ based on homogeneous earthquake sources and on parametric magnitude distributions are in most cases preferable. A problem with the latter estimators is that they are biased if the earthquake sources or the distribution or earthquake size are chosen incorrectly.

The bias of $\lambda_{SS}(y)$ can be reduced by scaling this estimator so that it coincides with $\lambda_H(y)$ at a specified site intensity y^* (a few variants of this idea are considered in this chapter). It is found that the resulting combined estimators perform best if y^* is an intensity that has been exceeded at the site about 20 times, according to the historical catalog. The optimum calibration intensity depends somewhat on the exceedence rate of interest (it is higher if one wants to estimate lower exceedance rates) and on the accuracy of the earthquake sources (it is smaller for source configurations that closely reflect the pattern of historical seismicity), these variations are however not large.

Optimally calibrated combined estimators are superior to uncalibrated seismic-source estimators, in the sense of being closer to the exceedance rates obtained from detailed local models of seismicity. Another important property of the combined estimators is that they are robust with respect to misspecification of the earthquake sources. Therefore, these estimators are useful when the source boundaries cannot be estimated accurately and even more useful, when the very existence of homogeneous earthquake sources is in doubt.

SITE	N	RMAX	ACC	PDIF
1	35	0.2456528	5.5508688E-03	2.9237509E-02
2	35	8.9803219E-02	1.8975601E-02	0.9281116
3	35	0.2445006	4.4321716E-03	3.0416887E-02
4	33	0.2335482	2.5469183E-03	5.4692954E-02
5	35	0.1667674	4.8329304E-03	0.2844647
6	33	0.1470493	8.2659582E-03	0.4734901
7	35	0.1923276	4.2078495E-03	0.1500284
8	34	0.1479676	1.1686089E-02	0.4463950
9	29	0.1525304	9.3309293E-03	0.5099026
10	33	9.2692405E-02	9.0133706E-03	0.9274294
11	34	0.1140721	1.6809855E-02	0.7677401
12	35	0.1950848	3.7926908E-03	0.1392164
13	35	0.1793252	6.3754572E-03	0.2101933
14	35	0.1463561	5.2699270E-03	0.4413616
15	34	9.3122661E-02	7.9846457E-03	0.9200233
16	35	0.1561385	5.3619575E-03	0.3606302
17	34	0.1643441	6.1584823E-03	0.3176947
18	35	0.1342877	5.8467882E-03	0.5529675
19	33	0.2135632	4.4321716E-03	9.8627299E-02
20	35	0.2511505	4.6684528E-03	2.4148036E-02
21	34	0.2550245	5.2699270E-03	2.4057195E-02
22	32	0.2322709	5.6478051E-03	6.3303739E-02
23	35	0.1386299	1.2524039E-02	0.5115321
24	35	0.1681753	6.9519286E-03	0.2753009
25	34	0.1075666	8.4103113E-03	0.8248645
26	35	0.1007877	5.0032036E-03	0.8657724
27	35	0.2117304	1.0904205E-02	8.6649381E-02
28	35	8.3403766E-02	8.5571846E-03	0.9438092
29	34	0.1827569	1.3422073E-02	0.2063685

Table 4-I: Results from the Kolmogorov-Smirnov tests on the upper-tails of the historical and seismic-source hazard functions for the 29 sites of Figure 4.4.

	Estimator	Prediction rate, λ(events/year)		
		10^{-2}	10^{-3}	10^{-4}
Accurate source configuration	$\lambda^{(1)}_{SS-H}$	1.55	1.60	1.75
	$\lambda^{(2)}_{SS-H}, \lambda^{(3)}_{SS-H}$	1.35	1.40	1.60
	λ_{SS}	1.55	1.65	1.90
Poor source configuration	$\lambda^{(1)}_{SS-H}$	1.55	1.65	1.85
	$\lambda^{(2)}_{SS-H}, \lambda^{(3)}_{SS-H}$	1.40	1.50	1.70
	λ_{SS}	2.00	2.20	2.50

Table 4-II: Estimated error factors with respect to the local estimator λ_L.

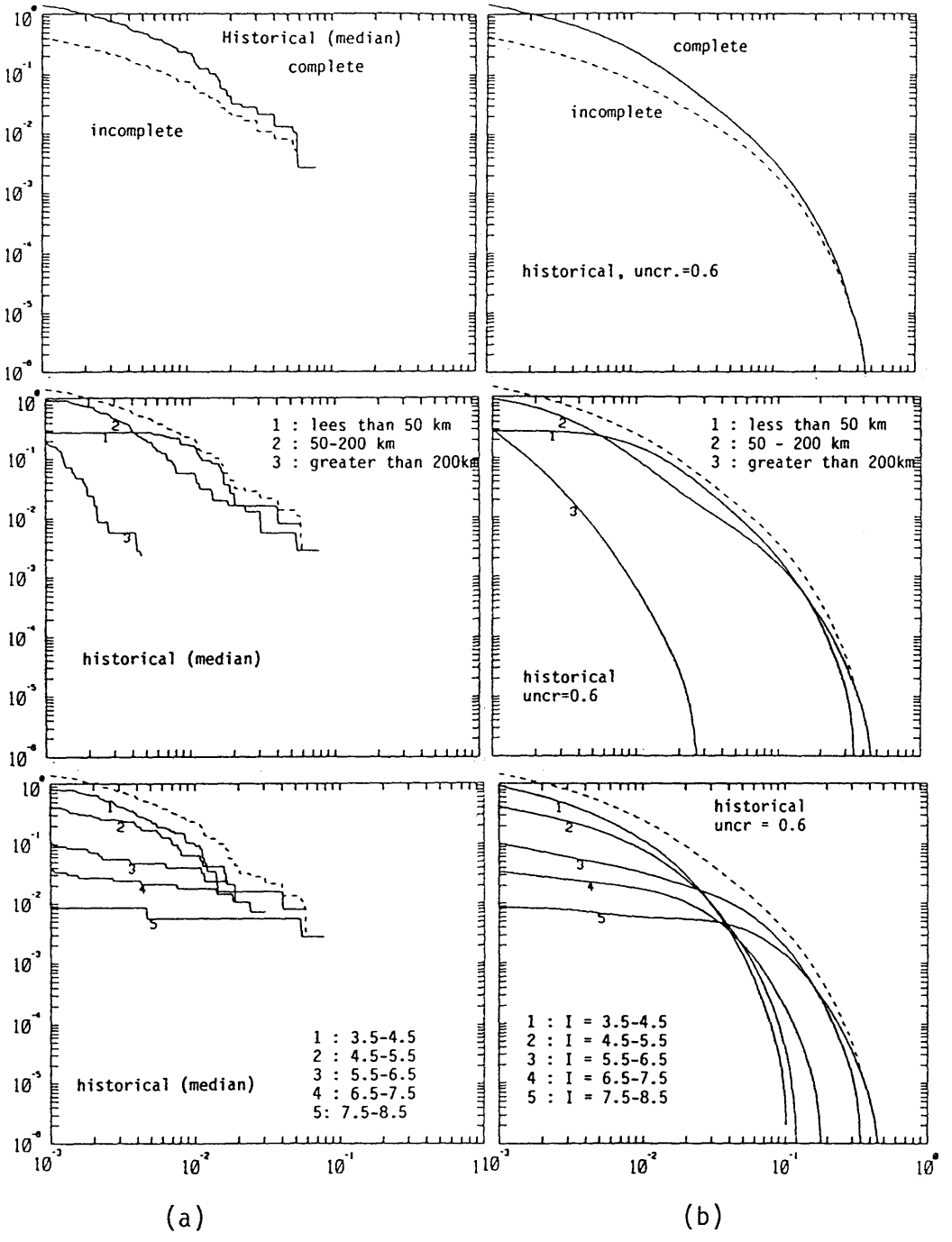
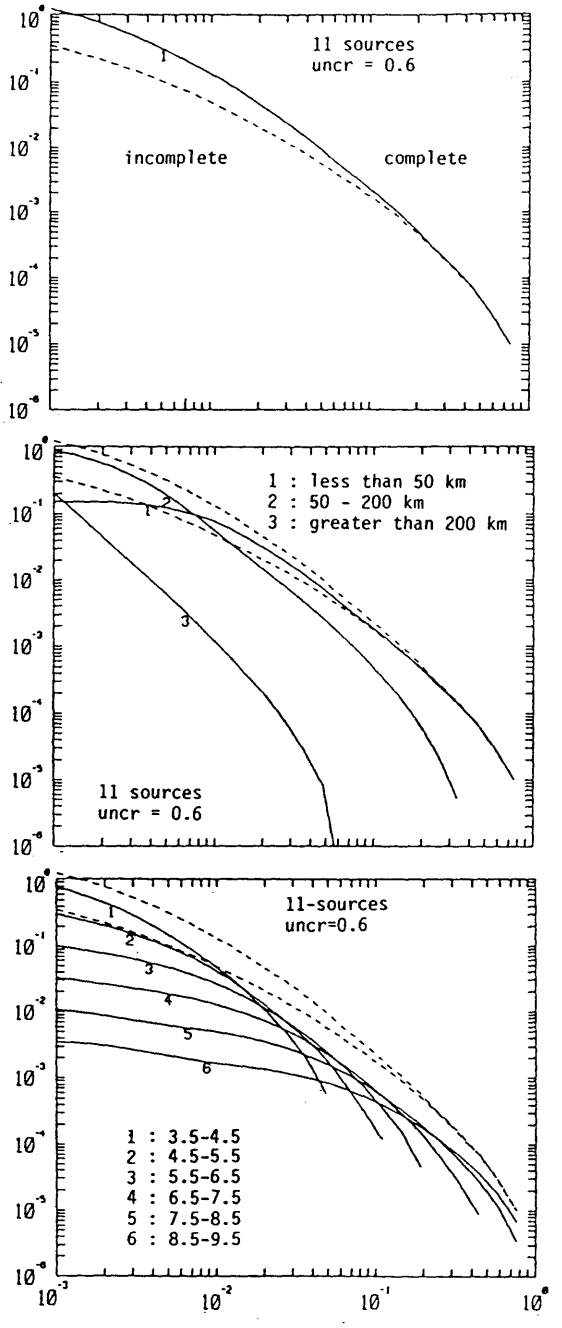
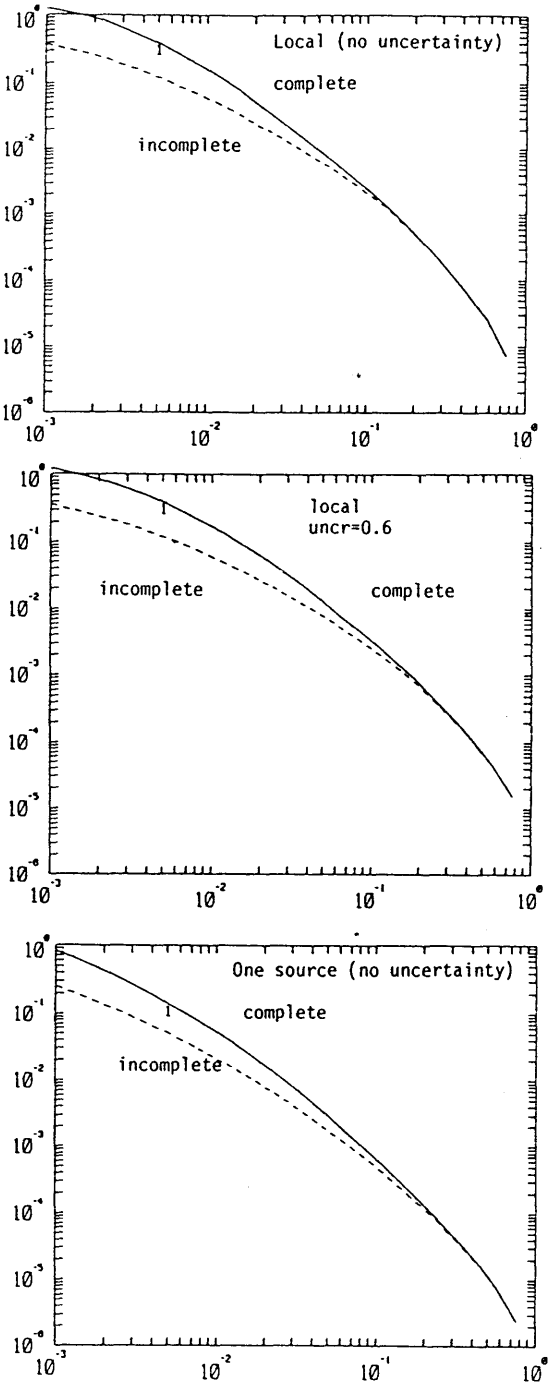


Figure 4-1: Seismic hazard estimates for a site located at 74°W 45°N.


11 sources
uncr = 0.6
incomplete
complete
1 : less than 50 km
2 : 50 - 200 km
3 : greater than 200 km
11 sources
uncr = 0.6
11-sources
uncr=0.6
1 : 3.5-4.5
2 : 4.5-5.5
3 : 5.5-6.5
4 : 6.5-7.5
5 : 7.5-8.5
6 : 8.5-9.5

(c)


Local (no uncertainty)
complete
incomplete
local
uncr=0.6
incomplete
complete
One source (no uncertainty)
complete
incomplete

(d)

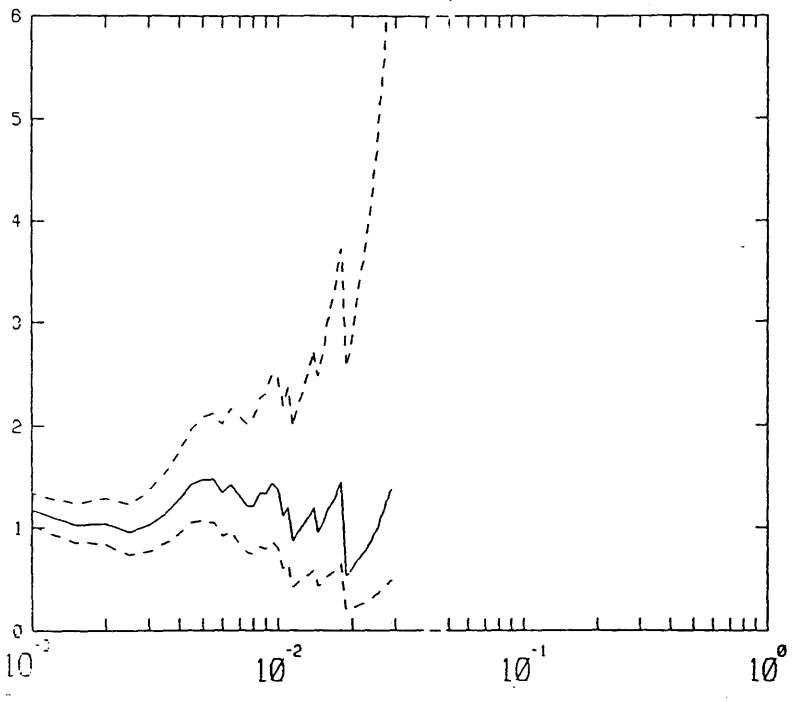


Figure 4-2: Calibration factor and associated uncertainty for a site at location 72°W 45°N and the combined estimator of Equation 4.14.

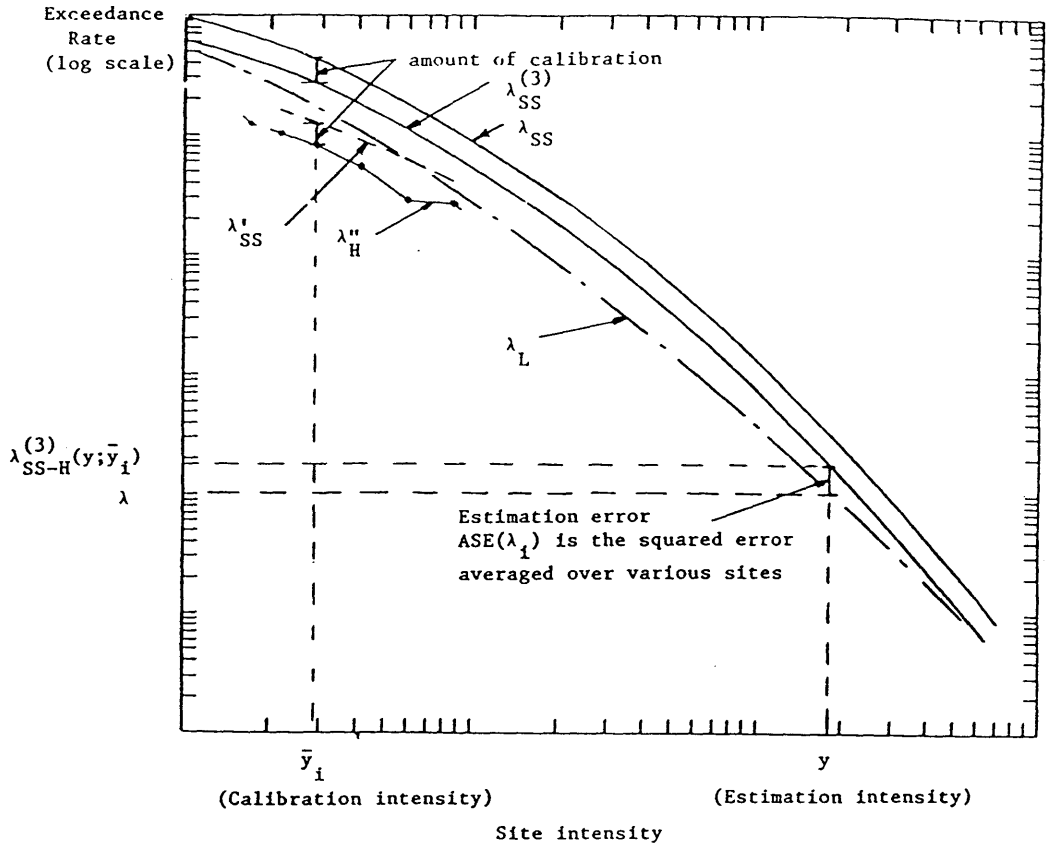


Figure 4-3: Illustration of the combined estimator $\lambda^{(3)}_{SS-H}$ and its error with respect to λ_L.

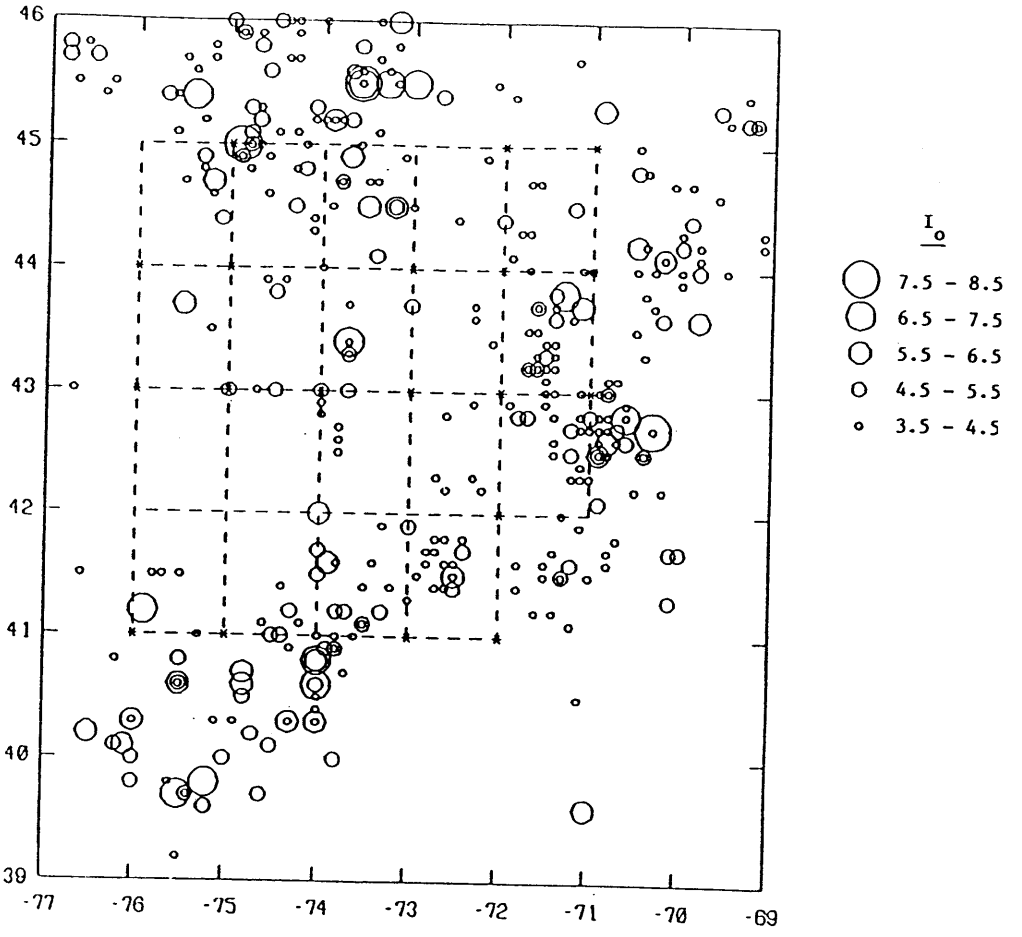
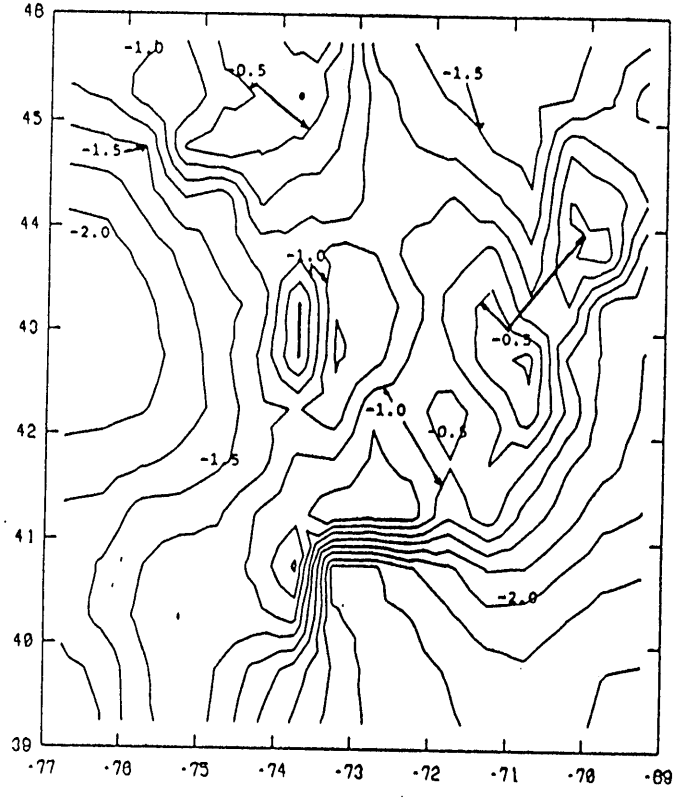
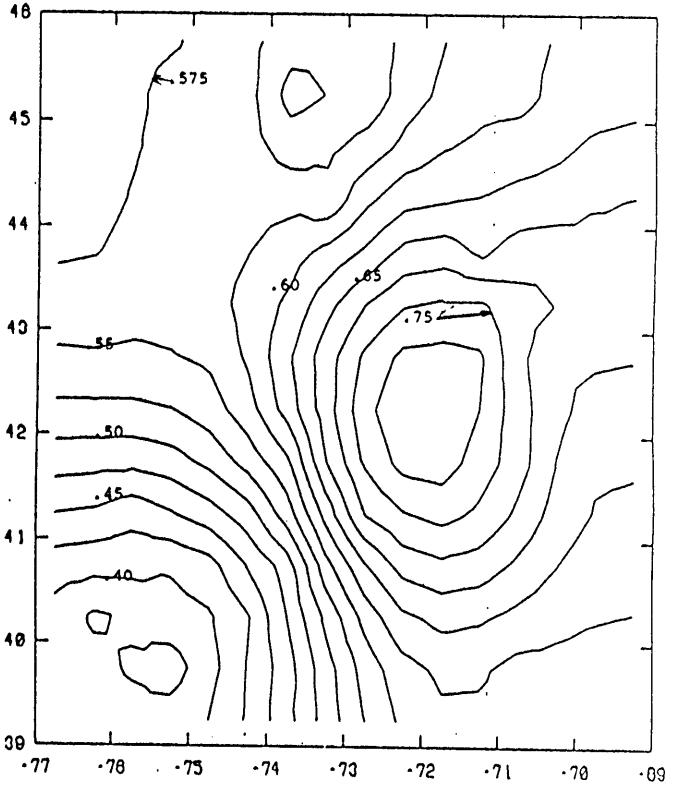


Figure 4-4: Earthquakes with MM intensity greater than 3.5 from 1627 to 1981 (Chiburis 1981). The starred points on the grid are used to estimate and rank different hazard estimators.

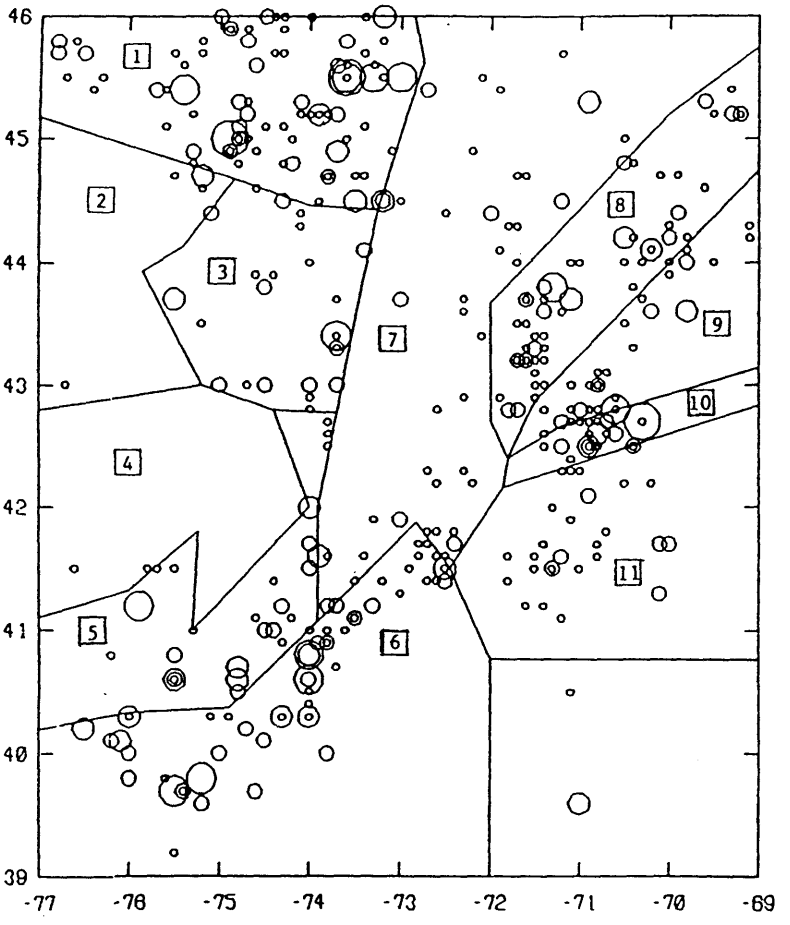


a(x)



b(x)

Figure 4-5: Contour plots of the seismicity parameters a and b in Equation 4.5 used for the local estimator λ_L.



11 Sources

Source	a	b
1	-.562	.552
2	-1.645	.651
3	-.930	.565
4	-2.139	.869
5	-1.100	.452
6	-1.155	.447
7	-1.161	.730
8	-.713	.643
9	-.705	.790
10	-.574	.543
11	-1.203	.782

1 Source

a	b
-1.037	.585

Figure 4-6: Estimates of seismicity parameters for the two source configurations used in the calculation of mean squared errors.

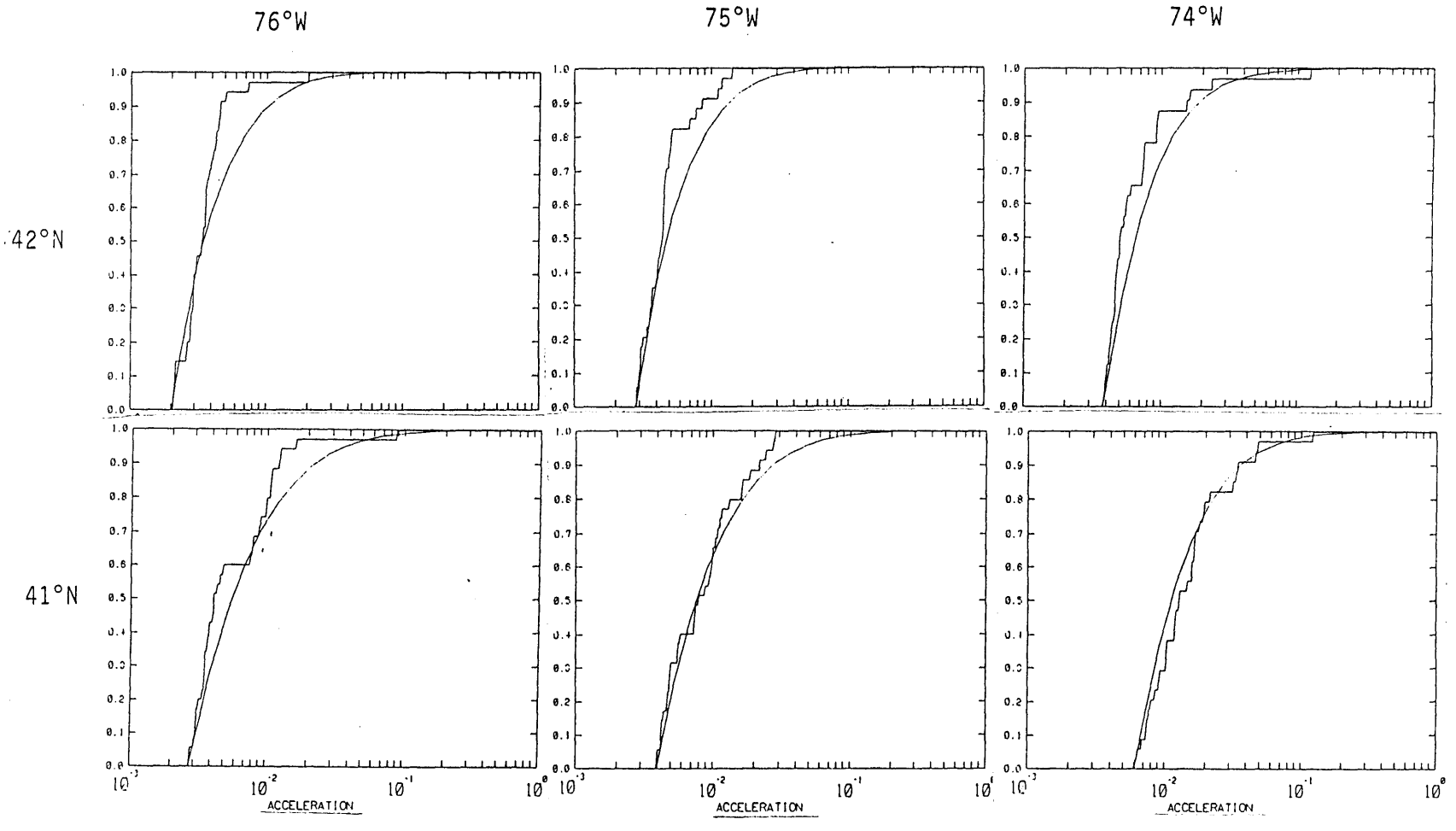
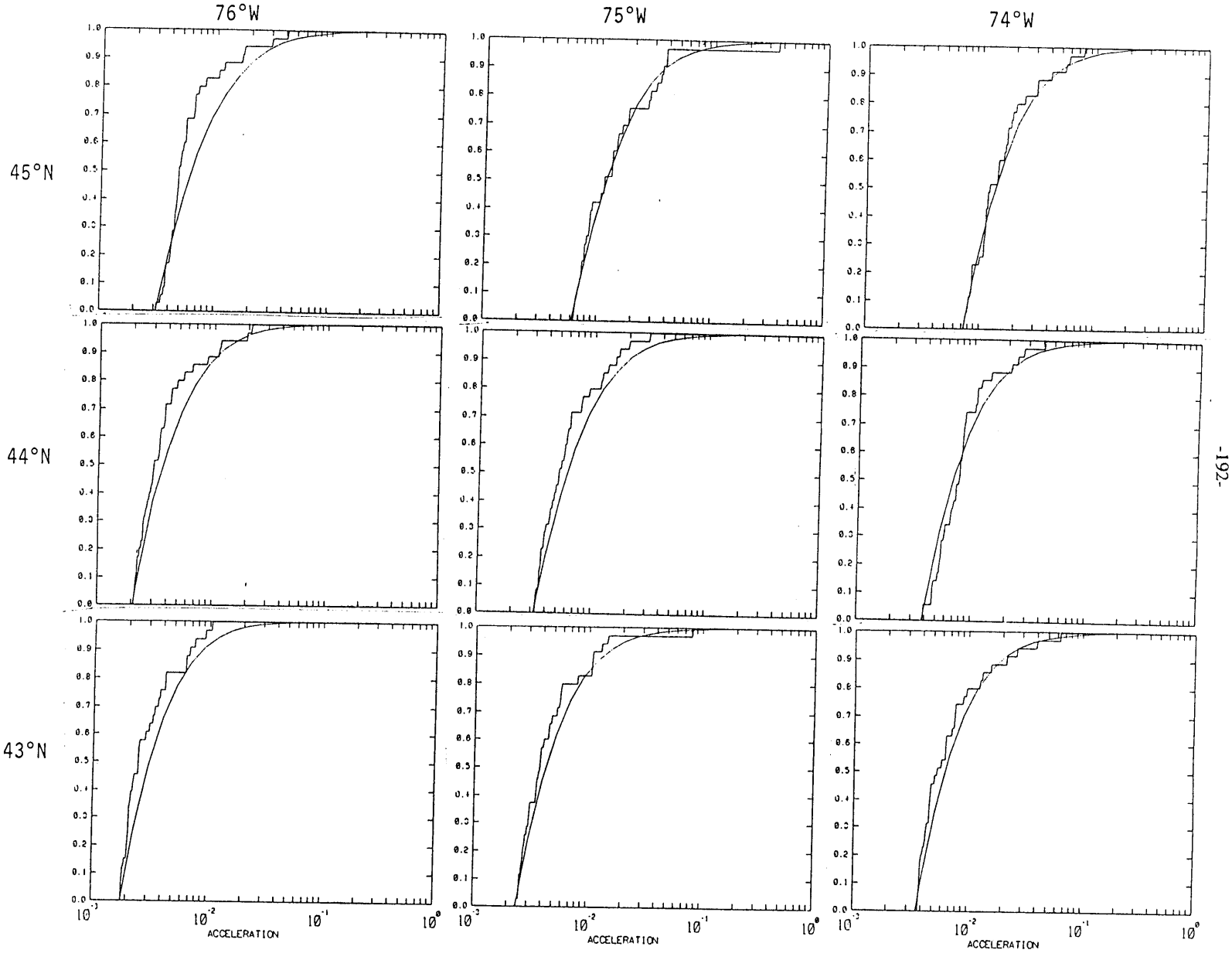
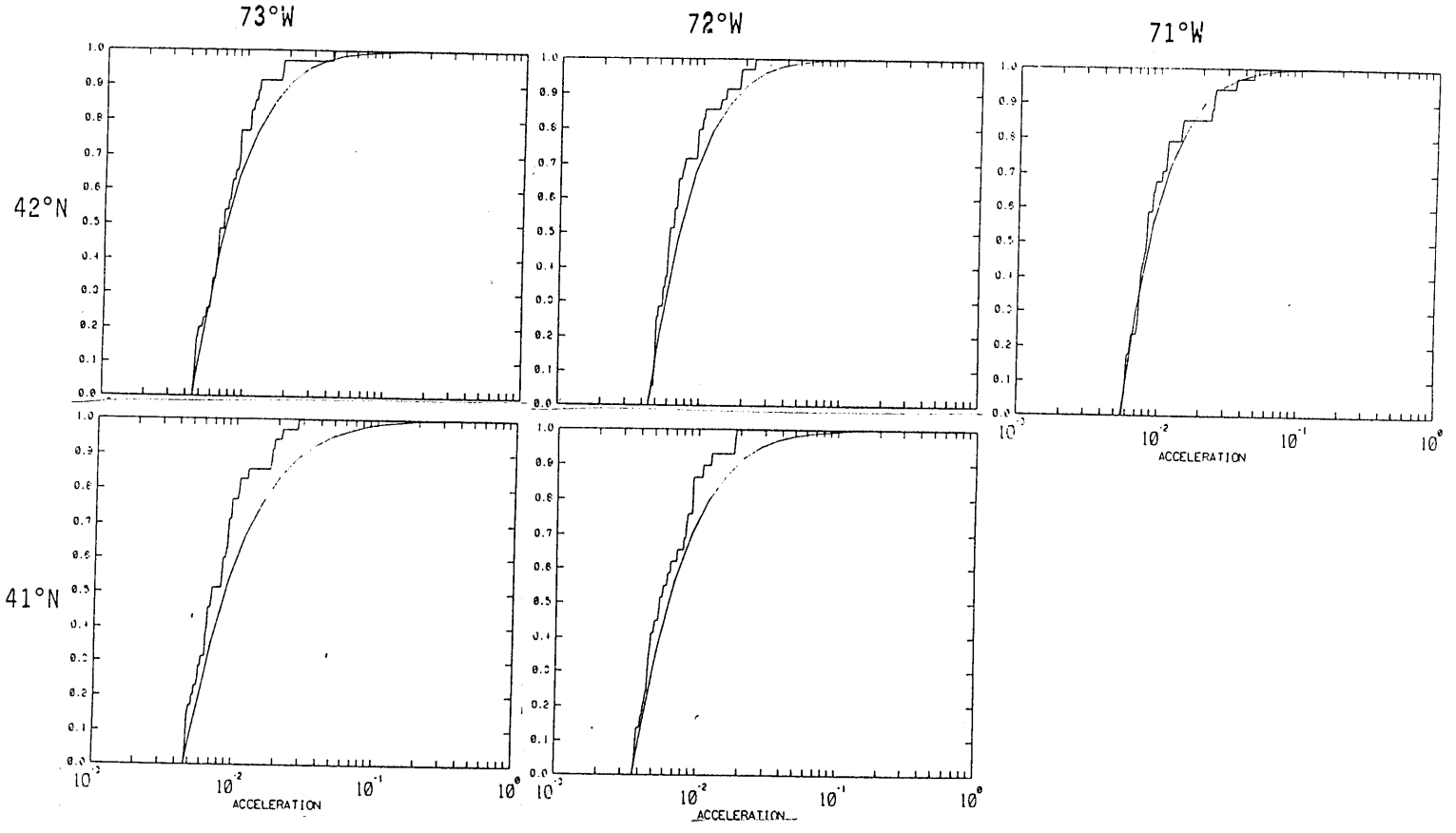
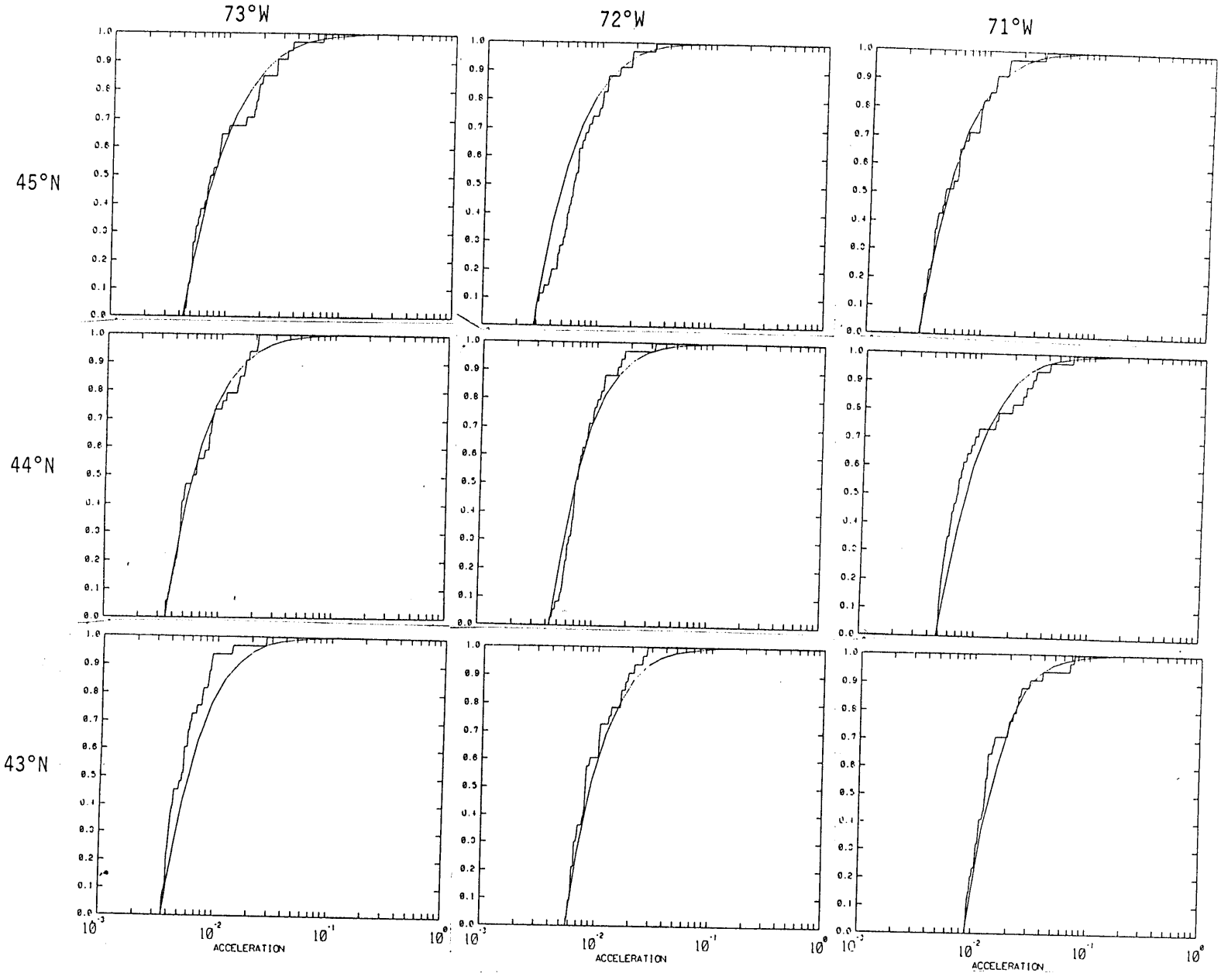


Figure 4-7: Comparison of the upper tails of the historical and seismic-source hazard functions for the 29 sites of Figure 4.4


76°W
75°W
74°W
45°N
44°N
43°N
ACCELERATION
ACCELERATION
ACCELERATION


73°W
72°W
71°W
42°N
41°N
ACCELERATION
ACCELERATION
ACCELERATION


73°W
72°W
71°W
45°N
44°N
43°N
ACCELERATION
ACCELERATION
ACCELERATION

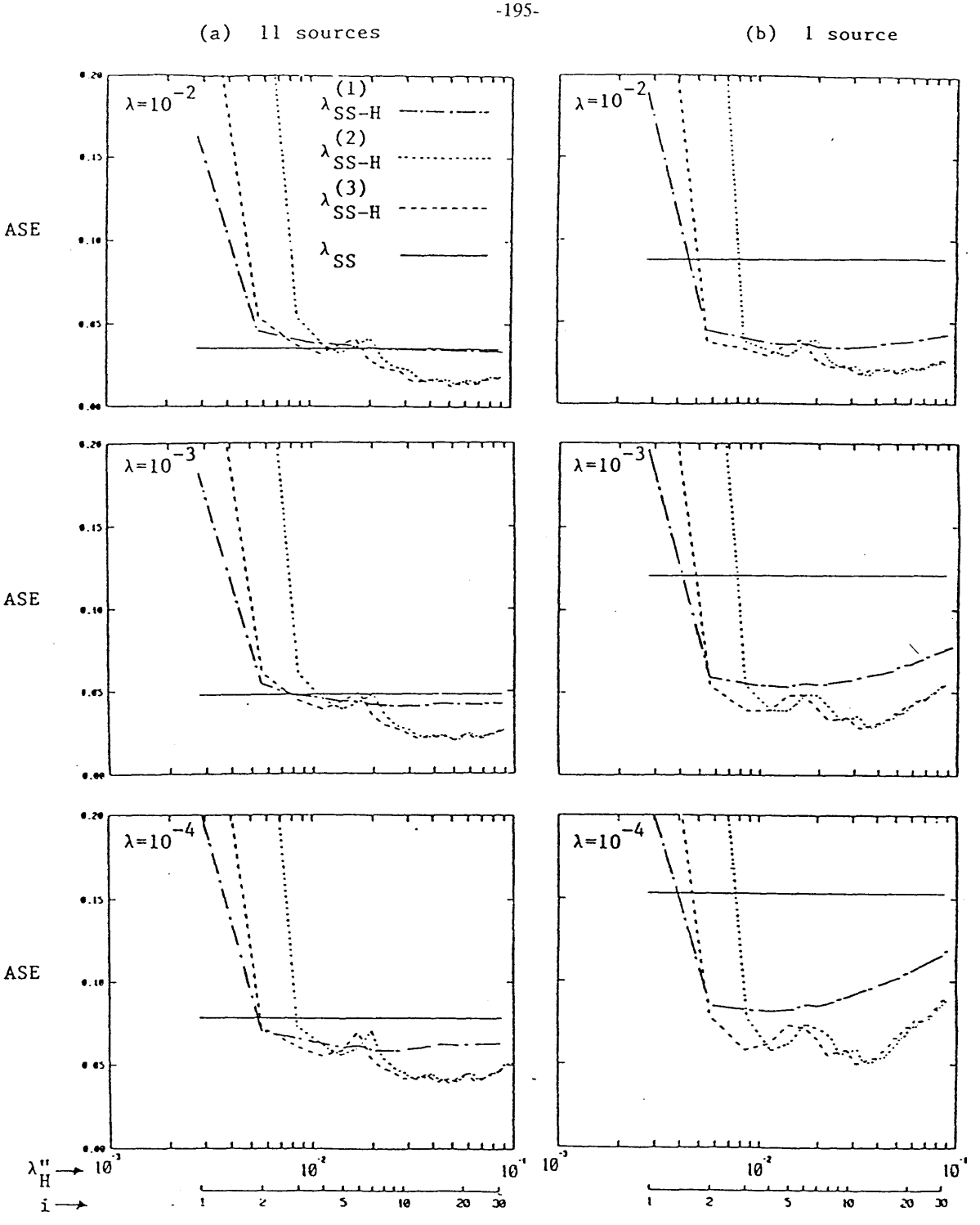


Figure 4-8: Comparisons of ASE (Equations 4.12 and 4.15) for two source configurations and three exceedance rates.

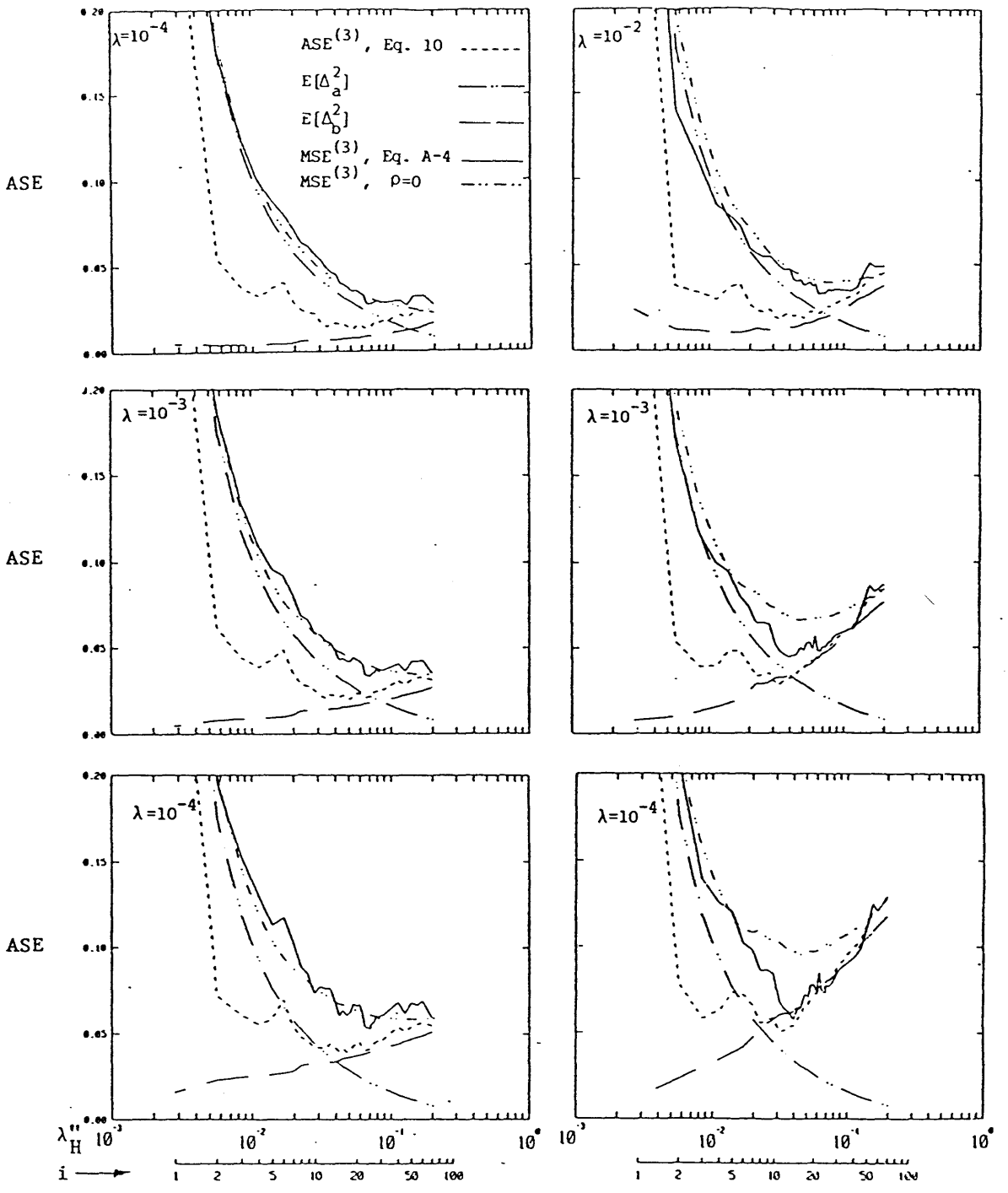
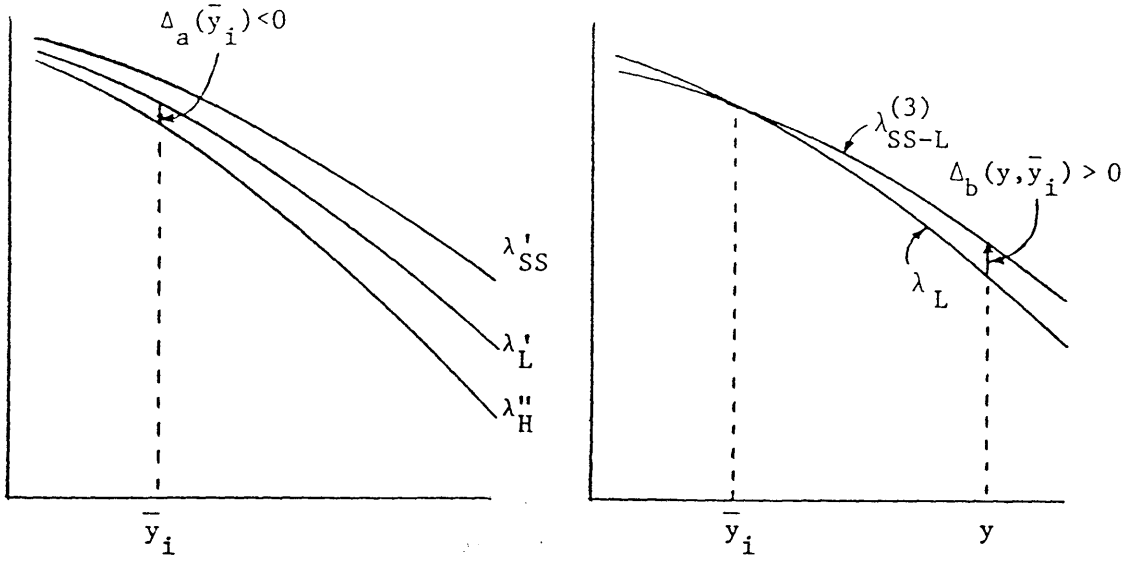
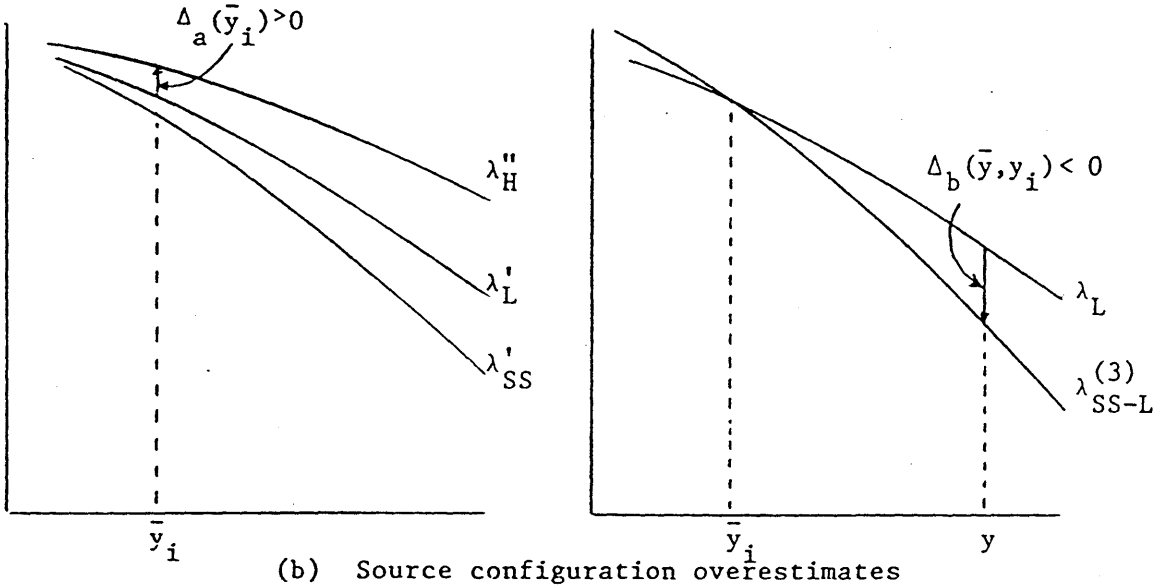


Figure 4-9: Mean sqared error of $\lambda_{SS}^{(3)}$ with respect to the true rate λ.



(a) Source configuration underestimates $b(\underline{x})$ locally



(b) Source configuration overestimates $b(\underline{x})$ locally

Figure 4-10: Typical situations resulting in negative correlation between Δ_a and Δ_b. Notice that the estimators λ''_H, λ'_L and λ'_{SS} are similar for site intensities smaller than the calibration intensity $\bar{y}$ (plots on the left). The symbol $\lambda^{(3)}_{SS-L}$ is used for the SS curve calibrated to the local estimator.

Chapter 5

Conclusions and Recommendations

The problems addressed in the thesis can be classified into three groups, 1. the identification and preservation of significant discontinuities in the estimation of models for intra-plate seismicity, 2. the selection of model parameters, 3. and the estimation of seismic hazard.

The main contribution with respect to the estimation of seismicity models is in the development of procedures which objectively identify and preserve significant changes in the spatial variation of the rate of activity. This is an improvement over present day procedures which require the external specification of seismic sources inside which seismicity is assumed constant. In the proposed procedure, seismic sources can optionally be used in the identification of significant features but influence the estimates only if validated by the data.

With respect to the selection of model parameters, one of the present day approach is to select penalties such that the number of observed and expected significant residuals are equal for different partitions of the catalog (i.e. the flagging procedure, section 3.2.4). The implementation of this procedure is easy, its application not computer-intensive, and the visual display of the results is informative with respect to possible causes of lack-of-fit. However, the test lacks power when there are few observations in each cell. Two new selection procedures are proposed : either certain observed statistics are set equal to their mean or median values under the model, or cross-validated measures such as the likelihood are maximized. The first is an extension of the flagging procedure with new statistics (section 3.2). However, this method lacks the predictive interpretation of the second procedure (cross-validation), which can be used to simultaneously select several model parameters and compare competing models.

A final contribution is with respect to the estimation of seismic hazard. The problem with present estimators, (i.e. the seismic-source and historical estimators) is that they are potentially inaccurate: the seismic-source estimates can be biased if the seismic source configuration is not properly specified and the historical estimates may have a large variance for small recurrence rates given the small sample size. Using a more sophisticated model of seismicity as the one proposed above is a solution but requires substantial work. The combined estimator proposed in section 4.2 is a much simpler alternative, is a significant improvement over the seismic source and historical estimators, and is shown to be robust with respect to badly specified seismic source configurations.

The conceptual results of the research for the estimation of seismicity models and seismic hazard are reviewed next, followed by a discussion of applications to New England and the Eastern United States and recommendations for future work.

• Formulation and estimation of the model

The models considered in this thesis represent seismicity through a Poisson process, non-homogeneous in space and (locally) stationary in time. It is assumed that the size distribution of events is exponential and location dependent. Estimation procedures are considered, which weigh the observations differently as a function of location, and size.

Spatial variation is allowed and estimates of the parameters a and b in Eq. 1.1 are obtained by spatial smoothing of the estimates (section 2.2). The level of smoothness is controlled through a penalty parameter and smoothness is measured as the difference between the estimate at a given location and the average of estimates at neighboring locations (the so-called local neighborhood). For estimating $a(\underline{x})$, procedures which are recommended are those that smooth the estimates within local neighborhoods having similar levels of activity. The local neighborhoods for a given cell are identified

through a test of equality of the recurrence rate (assuming a Poisson process) with each of the neighboring cells. For smoothing estimates of $b(\underline{x})$, it was found preferable to keep a fixed neighborhood (the eight neighboring cells) given the typically small number of observations in each cell (section 3.3). An alternative which has been explored is to use for $b(\underline{x})$ the same neighborhoods as the ones identified for $a(\underline{x})$. The assumption for such a procedure is that one may expect that given a larger sample, one would identify similar homogeneous neighborhoods for $a(\underline{x})$ and $b(\underline{x})$. However, this estimator is found not to be as accurate as when fixed neighborhoods are used.

In the application to the Chiburis catalog, the procedure was modified to allow the significance level for the test of equality of the recurrence rates to vary as a function of space. However, the effect of such a modification on the estimates was found to be minimal (section 3.3.1). Information on seismic source configuration is included in the estimation of local neighborhoods through a modification of the significance level for the test (section 2.4). The internal homogeneity of a source is measured by the odds ratio that two cells within the same source are homogeneous, and is estimated as the ratio of the number of times the null hypothesis is accepted to the number of rejections (at the given significance level). If this odds ratio is larger than the odds ratio obtained when the cells are not classified according to source ($\hat{R}_0$), the source identifies a zone of homogeneous seismicity, and the significance level of the test for pairs of cells within that source is lowered to allow greater internal smoothing. If the odds ratio for pairs of cells in neighboring sources is smaller than $\hat{R}_0$, the boundary between the two sources identifies a significant discontinuity in the rate of seismicity and the significance level for similar tests is increased to lower the likelihood that local neighborhoods are identified across the boundary. If an anomaly (e.g. a very active cell) is found within an hypothesized homogeneous source, the odds ratio for the zone decreases significantly and the anomaly is extracted in the fitting of the model. It is

interesting to note that, if a source configuration is non-informative, the estimates are identical to those which would be found in the absence of the zonation. In consequence, the procedure is robust with respect to misspecification of the seismic sources. Neighboring sources found to have similar seismic characteristics are merged in a preprocessing step. In all the present applications, the effect of the inclusion of expert opinion on the estimates was found influential only when a boundary is locally associated with a large gradient in the observed rate of activity (section 2.5).

Finally, analyses show that a grid of half degree cells offers a good level of discretization and that there is no gain in the accuracy of predictions for smaller discretizations.

• Selection of model parameters

The main parameters of the previous models which need to be selected are the penalties on the estimates of $a(\underline{x})$ and $b(\underline{x})$ (P_a, P_b). Their selection is performed through two different approaches: target-statistics and cross-validation.

The target-statistics procedure is suggested for obtaining quick estimates of the optimal penalty for $\hat{a}(\underline{x})$. The target-statistics procedure compares the total number of observed and estimated events in each cell through various goodness-of-fit statistics. The recommended targets are the expected value or the median of the test statistics. The procedure is computationally less demanding than cross-validation and tends to identify optimal penalties slightly larger than those from cross-validation (section 3.2.8).

In cross-validation, the catalog is divided in time into non-overlapping estimation and validation samples, and optimal models are selected on the basis of statistics measuring the accuracy of the predictions. Extrapolation is recommended in the definition of the cross-validation samples, which means that the estimation sample associated with a given prediction sample contains only prior seismicity.

For most applications, cross-validation is not recommended for determining the parameters controlling the selection of the local neighborhoods (the extent of the local neighborhoods [M] and the level of significance of the test [α]) (section 3.3.1), because the outcome of the tests is greatly affected by the removal of the validation sample given the small number of observations usually available in each cell and migration of seismicity. It is recommended, based on the results for New England and the Eastern United States, to limit the size of local neighborhoods to immediately neighboring cells and to fix the level of significance of the test to either 10% or 15%. Similarly, the parameters controlling the variation of the probability of detection are not cross-validated and are kept fixed to estimates obtained using the whole catalog. This is justified by the fact that cross-validating the probability of detection has little effect on the selection of the optimal penalties P_a and P_b and is computationally much more demanding.

• Goodness-of-fit

Goodness-of-fit of the predictions can be assessed through an analysis of the residuals (for example, the flagging procedure of section 3.2.4). Alternatively, one may compute the distribution of the cross-validated log-likelihood through simulation to which is compared the observed statistic (section 3.3.2). The cross-validated likelihood and its expeted value can be spatially displayed to identify systematic lack-of-fit over extended regions.

For the size distribution of the events, the flagging procedure (section 3.2.4) can be used on the total number of observations in different size intervals for checking the assumption of exponentiality for the full catalog or for shorter periods of observation. The lack-of-fit of the exponential model within each spatial cell can also be analysed, but requires large numbers of observations (section 3.3.2). One may also resort to simulation to determine the distribution of the cross-validated likelihood for different magnitude intervals (section 3.3.2).

Finally, it is recommended to fit models which use different amounts of the most recent seismicity and to compare the estimates on the basis of the cross-validated statistics. Seismicity in the near future can resemble more the recent past than the average seismicity during long periods of time.

• Seismicity of the Eastern United States

In this application, the optimal penalties on $a(\underline{x})$ were small indicating that the locations which are most likely to be active in the future are those which have been active in the past. This localized pattern of predicted seismicity is true for both small and large magnitude events for the regions analysed and the time periods considered. For this data set and the partition considered, the likelihood that previously completely inactive areas become active in the future is small. However, the relative level of activity of previously active areas can fluctuate significantly from time to time. Positive and negative residuals are not randomly distributed spatially, but are predominant over extended regions, suggesting that regions are more or less active than others as a function of time. In contrast, the optimal penalty P_b is found to be large, resulting in almost constant estimates. Notice that the influence of P_b on the cross-validated statistics is much smaller than P_a because it affects mainly the estimates of the recurrence rate for the large magnitude events.

For the application to the Eastern United States, regions of significantly higher seismicity were identified around Newburyport, the Ottawa River Valley, the Charlevoix area, the Charleston area, Eastern and Western Tennessee, and Eastern Virginia (Figure 2.11). Local fluctuations of the recurrence rate are detected in the Charleston area, and Western Tennessee. Regional non-stationarities over larger areas are detected in the Ottawa River Valley, southern New Hampshire, the Charlevoix area, and along a ridge across Eastern Tennessee and Virginia (Figure 3.28a).

• Alternative estimator of seismic hazard

The seismicity estimates obtained using the above procedures can be combined with an attenuation function to determine the seismic hazard at different sites. An alterantive estimator of seismic hazard is proposed (Chapter 4) which combines estimates from much simpler models of seismicity (seismic source and historical models) and produces results similar to those using the previous estimates. Stationarity is assumed both for the historical and seismic-source estimators of seismic hazard. It is recommended to use a reasonable source configuration with respect to the observed seismicity and to calibrate the seismic source hazard estimates with respect to the 20^{th} largest historical event (section 4.2). The combined estimator has been shown to be robust with respect to badly specified source configurations. The seismic hazard estimates obtained with this procedure are close to the estimates obtained using the local model of seismicity.

• Recommendations for future work

The previous models incorporate many assumptions with respect to the size distribution of the events and the stationarity of the process in time. With respect to the probability of detection, there are additional assumptions with respect to its variation as a function of time and magnitude (constant within each incompleteness regions, monotonically increasing with time and magnitude).

Refinements could be made with respect to the magnitude distribution of the events, specially with respect to the larger events. Here, the simple exponential model may be unconservative in some instances, specially with respect to so-called characteristic events. In addition, the assumption in this application that the maximum magnitude is known with certainty and is equal at all locations is unrealistic. These are important issues given the infleunce of large size events on seismic hazard for small recurrence rates (e.g. 10^{-4} event/year). Extensions of the model should eliminate the need for spatial discretization, which introduces some bias in the estimates.

In future work, the non-parametric formulation of the model in space should be extended to time and magnitude. In particular, this will allow a better understanding of the phenomenon of migrating seismicity which was detected. This modification can be implemented though a kernel estimation procedure (section 2.6). Kernel estimation is computationally less demanding than maximum penalized likelihood and does not require discretization in space, time, or magnitude. A wide variety of kernel estimators is available, some of which can preserve discontinuities in the variation of the seismicity parameters (e.g. anisotropic and adaptive kernel functions).

References

Aki, K. (1965). Maximum Likelihood Estimate of b in the Formula LogN=a-bM and its Confidence Limits. *Bull. Earthquake Res. Inst., 43*, 237-239.

Anderson, J. G. (1979). Estimating the Seismicity from Geologic Structure for Seismic Hazard Analysis. *Bull. Seism. Soc. Am., 69*, 135-168.

Armbruster, J. and Seeber, L. (1987). Seismicity and Seismic Zonation Along the Apalachian and the Atlantic Seaboard from Intensity Data. *Symp. on Seismic Hazards, Ground Motions, Soil-Liquefaction and Engineering Practice in Eastern North America.* Sterling, New York, Technical Report NCEER-87-0025.

Barosh, P. J. (1986). *State-of-the-Art for Assessing Earthquake Hazards in the United States* (Tech. Rep. Miscellaneous Paper 5-73.1, Report 21). U. S. Army Engineering Waterways Experiment Station, Seismic Source Zoning of the Atlantic Seabord and Apalachian Regions.

Barstow, N.L., Brill, K.G., Nuttli, O.W., and Pomeroy, P.W. (1981). *An Approach to Seismic Zonation for Siting Nuclear Electric Power Generating Facilities in the Eastern United States* (Tech. Rep. NUREG/CR-1577). Washington, D.C.: U.S. Nuclear Regulatory Commission,

Bath, M. (1956). A Note on the Measure of Seismicity. *Bull. Seism. Soc. Am., 46*, 217-218.

Bender, B. (1983). Maximum Likelihood Estimation of b Values for Magnitude Grouped Data. *Bull. Seism. Soc. Am., 73*(3), 831-851.

Berger, O. J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag.

Bhapkar, V.P. (1980). ANOVA and MANOVA. Models for categorical data. In Krishnaniah, P.R. (Ed.), *Analysis of Variance, Volume 1.* North-Holland Publ. Co.

Bishop, Y. M. M.; Feinberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT press.

Borissof, B.A., Reisner, G.I. and Sholpo, V.N. (1977). A Geotectonic Method of Predicting the Maximum Magnitudes of Expected Earthquakes as Applied to the Northern Italy Area. *Boll. Geof. Teor. Appl., 20*(73), 19-26.

Bradley, J.V. (1968). *Distribution-Freee Statistical Theory*. Prentice-Hall.

Caputo, M. and Postpischl, D. (1974). Contour Mapping of Seismic Areas by Numerical Filtering and Geological Implications. *Annali di Geofisica, 27*(3-4), 619-639.

Catteneo, M.,Eva, C., and Merlanti, F. (1981). Seismicity of Northern Italy: A Statistical Approach. *Boll. Geof. Teor. Appl., 23*(89), 31-42.

Chiburis, E. F. (1981). *Seismicity, Recurrence Rates, and Regionalisation of the Northeastern U.S. and Adjacent Southeastern Canada* (Tech. Rep. NUREG/CR-2309). Washington, D.C.: Nuclear Regulatory Commission,

Chinnery, M. A. (1979). A Comparison of the Seismicity of Three Regions of the Eastern United States. *Bull. Seismol. Soc. Am.*, (69), pp. 757-772.

Consentino, M. (1978). Frequency-Magnitude Statistical Parameters with Reference to Some Zoning Problems. *Boll. Geof. Teor. Appl., 20*(78), 198-204.

Cornell, C.A. (1968). Engineering Seismic Risk Analysis. *Bull. Seism. Soc. Am., 54*(5), 1583-1606.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns*. John Wiley and Sons.

Ebel, J.E. (1987). The Seismicity of the Northeastern United States. *Symp. on Seismic Hazards, Ground Motions, Soil-Liquefaction and Engineering Practice in Eastern North America*. Sterling, New York, Technical Report NCEER-87-0025.

EPRI. (1985). *Seismic Hazard Methodology for Nuclear Facilities in the Eastern United States* (Tech. Rep. Research Project P101-29). Palo Alto, California: Electric Power Research Institute,

Epstein, B, and Tsao, C.K. (1953). Some Tests Based on Ordered Observations from Two Exponential Populations. *Annals of Mathematical Statistics,* (24), pp. 458.

Esteva, L. (1969). Seismicity Prediction: A Bayesian Approach. *Proceedings, Fourth World Conference on Earthquake Engineering,* , pp. A1-172,A1-184.

Gibbons, J.D. (1985). *Non-Parametric Methods for Quantitative Analysis (2nd Edition).* American Science Press. American Series in Mathematical and Management Sciences.

Good I.J. (1958). Significance Tests in Parallel and in Series. *Journal of the American Statistical Asociation,* (53), pp. 799-813.

Good, I. J. and Gaskins, R, A, (1980). Density Estimation and Bump Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteriorite Data. *Journal of the American Statistical Association, 75*(369), 42-72.

Greig-Smith, P. (1952). The Use of Random and Contiguous Quadrats in the Study of the Structure of Plant Communities. *Annals of Botany,* (16), pp. 293-316.

Hand, D. J. (1982). *Kernel Discriminant Analysis.* Research Studies Press (John Wiley and Sons).

Heidari, M. (1987). *Statistical Methods of Earthquake Attenuation.* Doctoral dissertation, Dept. of Civil Engineering, MIT,

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions-2.* Boston: Houghton-Mifflin Co.

Kaila, K.L. and Hari Narain. (1971). A New Approach for Preparation of Quantitative Seismicity Maps as Applied to Alpine Belt-Sunda Arc and Adjoining Areas. *Bull. Seism. Soc. Am., 61,* 1275-1291.

Kaila, K.L., Rao, N.M. and Hari Narain. (1974). Seismotectonic Maps of Southwest Asia Region, Comprising Eastern Turkey, Caucasus, Persian Plateau, Afghanistan and Hindu Kush. *Bull. Seism. Soc. Am., 64*, 657-669.

Krishnaiah, P.R and Sen, P.K. (1984). *Hankbook of Statistics, 4; Non-Parametric Methods*. North-Holland.

Larsen, R.J. and Marx, M.L. . (1981). *An Introduction to Mathematical Statistics and its Applications*. New Jersey: Prentice-Hall.

Lehman E. L. (1959). *Testing Statistical Hypothesis*. Wiley Series in Probability and Mathematical Statistics.

LLNL. (1985). *Seismicity, Recurrence Rates, and Regionalisation of the Northeastern U.S. and Adjacent Southeastern Canada* (Tech. Rep.). : Lawrence Livermore National Laboratory,

McGuire, R. K. (1979). Effects of Uncertainty in Seismicity on Estimates of Seismic Hazard for the East Coast of the United States. *Bull. Seism. Soc. Am., 67*, 777-792.

Medvedev, S.V. (Edr.). (1976). *Seismic Zoning of the USSR*. Jerusalem, Israel: Keter Publishing House.

Mitronovas, W. (1981). Temporal and Spatial Variation in Seismicity Within New York State and Eastern U.S. Beavers, J. E. (Ed.), *Proceedings of Earthquakes and Earthquake Engineering: The Eastern United States*. Knoxville, Tennesse.

Pielou. (1969). *The interpretation of Ecological Data, A Primer on Classification and Ordination*. John Wiley and Sons.

Przyborowski, J. and Wilenski, H. (1939). Homogeneity of Results in Testing Samples From Poisson Series. *Biometrika, 31*, 313-323.

Ripley, B. D. (1981). *Spatial Statistics*. John Wiley and Sons.

Shakal, A.F. and Toksoz, M.N. (1977). Earthquake Hazard in New England. *Science,* (195), pp. 171-173.

Silverman, B. W. (1985). *Density Estimation for Statistics and Data Analysis.* John Wiley and Sons.

St-Amand, P. (1956). Two Proposed Measures of Seismicity. *Bull. Seism. Soc. Am., 46*, 41-45.

Stepp, J.C. (1972). Analysis of the Completeness of the Earthquake Sample in the Puget Sound Area and Its Effect on Statistical Estimates of Earthquake Hazard. *Proceedings, International Conference on Microzonation, 2*, 897-910.

Thenhaus, M.; Perkins, D. M.; Algermissen, M. and Hanson, S. L. (1987). Earthquake Hazard in the Eastern United States: Consequences of Alternative Seismic Source Zones. *Earthquake Spectra, 3*(2), 227-261.

Titterington, D. M. and Bowman, A. W. (1985). A Comparative Study of Smoothing Procedures for Ordered Caterogical Data. *J. Statist. Comput. Simul., 21*, 291-312.

VanDyck, J. (1986). *Statistical Analysis of Earthquake Catalogs.* Doctoral dissertation, Dept. of Civil Engineering, MIT,

Veneziano, D. and Van Dyck, J. (1987). Statistical Analysis of Earthquake Catalogs for Seismic Hazard. In Y. K. Lin and R. Minai (Eds.), *Stochastic Approaches in Earthquake Engineering, Lecture Notes in Engineering.* New York: Springer-Verlag.

Weichert, D.H. (1980). Estimation of the Earthquake Recurrence Parameters for Unequal Observation Periods for Different Magnitudes. *Bull. Seism. Soc. Am., 70*, 1337-1346.

WGC. (1980). *Site Dependant Response- Spectra: Yankee Rowe* (Tech. Rep.). Westboro, Massachusetts: Weston Geophysical Corporation, Report to Yankee Atomic Electric Company.