Earthquake Modelling at the Country Level Using Aggregated Spatio-Temporal Point Processes

M.N.M. van Lieshout · A. Stein

Abstract The goal of this paper is to derive a hazard map for earthquake occurrences in Pakistan from a catalogue that contains spatial coordinates of shallow earthquakes of magnitude 4.5 or larger aggregated over calendar years. We test relative temporal stationarity by the KPSS statistic and use the inhomogeneous J-function to test for inter-point interactions. We then formulate a cluster model, and de-convolve in order to calculate the hazard map, and verify that no particular year has an undue influence on the map. Within the borders of the single country, the KPSS test did not show any deviation from homogeneity in the spatial intensities. The inhomogeneous J-function indicated clustering that could not be attributed to inhomogeneity, and the analysis of aftershocks showed some evidence of two major shocks instead of one during the 2005 Kashmir earthquake disaster. Thus, the spatial point pattern analysis carried out for these data was insightful in various aspects and the hazard map that was obtained may lead to improved measures to protect the population against the disastrous effects of earthquakes.

Keywords Aftershocks · Cluster point process · De-convolution · Hazard map · Inhomogeneous J-function · Intensity function · Relative stationarity test

In memory of Julian E. Besag.

M.N.M. van Lieshout
Probability, Networks and Algorithms, CWI, Amsterdam, The Netherlands

A. Stein (✉)
Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede,
The Netherlands
e-mail: stein@itc.nl

1 Introduction

Disasters such as earthquakes occur at apparently erratic seismic locations and at unexpected moments. Processes generating earthquakes are prominent in earthquake prone areas that are at least partly determined by geological faults and occur, in particular, close to subduction zones. An earthquake describes both a sudden slip on a fault and the resulting ground shaking and radiated seismic energy caused by the slip, by volcanic or magmatic activity, or other sudden stress changes in the earth. The release of energy at an unanticipated moment may then appear at the earth surface and is registered as the main shock. Main shocks are usually followed by aftershocks that are smaller than the main shock and within 1–2 rupture lengths distance from the main shock. Aftershocks can continue over a period of weeks, months, or years. In general, the larger the main shock, the larger and more numerous the aftershocks, and the longer they will continue. Other types of clusters known as swarms also occur. Such clusters are more diffuse and can be distinguished from aftershock sequences by their showing no clear correlation with a main shock, nor the typical decay in frequency and magnitude common to aftershock patterns.

In the analysis of earthquakes, spatio-temporal Hawkes processes, in particular, Ogata's Epidemic Type Aftershock-Sequences (*ETAS*) model, have become the standard first approximation for seismic catalogue data that come in the form of a list of earthquake locations, their times, and magnitudes (Ogata 1988). An excellent review is given by Ogata (1998) who also gives historical references. In such a Hawkes process, immigrants arrive according to a temporal Poisson process and are marked by their spatial location and other attributes as required. Each immigrant generates a finite marked Poisson process of 'offspring' with an intensity function that depends on the parent, independently of other immigrants. The offspring in their turn also generate offspring independently of all others, and so on. In other words, each immigrant produces a branching process of descendants. Therefore, a Hawkes process can also be described as a marked Poisson cluster process. An alternative to Hawkes processes is to use a Neymann–Scott process which restricts the branching to only one generation of descendants (Vere-Jones 1970; Adamopoulos 1976).

In the earthquake context, the offspring are swarms or aftershock clusters; their number typically depends on the magnitude of the parent, their temporal displacement follows the Omori (Pareto) power law. With regard to the marks, the magnitudes are widely assumed to follow a shifted exponential distribution; for the spatial displacements, various probability distributions have been tried. Examples, including the spherical Gaussian distribution we shall use in Sect. 6, are given in Ogata (1998), Holden et al. (2003), Daley and Vere-Jones (2003, 2008). See also Vere-Jones (1970), Vere-Jones and Musmeci (1992), Shurygin (1993), or Zhuang et al. (2002).

The advantage of focusing on the temporal dimension and treating other variables of interest (magnitude and spatial location) as marks is that a conditional intensity can be written down. Consequently, a likelihood function is available in closed form and can be used for inference. For details, see Ogata (1998). Moreover, many of the above papers consider the island states of Japan and New Zealand for which edge effects seem not to be an issue.

Early work in the earthquake literature includes a study into the clustering of earthquakes (Adamopoulos 1976) and the use of a Poisson process for the long-term prediction of seismicity for linear zones (Shurygin 1993). More recent contributions have been made by Holden et al. (2003) who developed a marked point process model for earthquakes, focusing on parameter estimation, by Lucio and Castelucio de Brito (2004), who use simulations to detect randomness in spatial point patterns with an application to seismology, by Pei et al. (2007) who studied vulnerability to earthquakes and aftershocks in a Bayesian framework, and by Pei (2011) who developed a non-parametric index for the numbers of events in an earthquake cluster on the basis of the nearest neighbour distance with an application to the earthquake catalogue of China. Our paper adds to the current literature in that non-homogeneity is taken into account explicitly both in the testing and in the spatial modelling.

The aim of this study is to explore spatial statistical techniques for data aggregated over time for which an explicit likelihood function is not available. Attention focuses on Pakistan, for which country annual patterns of earthquakes have been recorded for more than 35 years. During this period, two major earthquakes of magnitude larger than seven were recorded: one occurring in 1997 and the major Kashmir earthquake of 2005. Moreover, seismic activity in neighbouring countries may well influence occurrences inside Pakistan.

The plan of this paper is as follows. In Sect. 2 we discuss the data. In Sect. 3, we pool all normal years and calculate a kernel estimator for the pooled intensity map of earthquake occurrence. This map is then used to test for relative temporal stationarity (Sect. 4) and inter-point interactions (Sect. 5). We conclude that there is no evidence for a temporal trend, but that there does appear to be clustering in the data that cannot be attributed to geological inhomogeneity. These observations lead us to look at the patterns of aftershocks in the major earthquake years 1997 and 2005 in Sect. 6 in order to formulate a cluster model (Sect. 7), from which a hazard map can be derived. Such a map provides information about the likelihood of earthquake occurrences in an area of interest, but does not predict when the next disaster will strike. We verify in Sect. 8 that the hazard map is not dominated by exceptional years using the leave-one-out principle. The paper closes with a critical discussion and summary of the main findings.

2 Background and Data

Pakistan is a country that is regularly affected by earthquakes. The reason for the vulnerability of the country to earthquakes is the subduction of the Indo-Australian continental plate under the Eurasian plate with its two associated convergence zones. One such zone crosses the country from approximately its south-western border with the Arabic Sea to Kashmir in the North East. The other convergence zone crosses the northern part of the country in the east-west direction and is the direct cause of the Himalayan orogeny. Pakistan-administered Kashmir lies in the area where the two zones meet. The geological activity born out of the collisions is the cause of unstable seismicity in the region.

Earthquakes can be severe with a devastating effect on human life and property. Two major earthquakes were recorded in 1997 and 2005 with magnitudes of 7.3 and

7.6, respectively. The 1997 earthquake occurred along the convergence zone running from the South West to the North East and resulted in about seventy casualties. The 2005 Kashmir earthquake, however, was catastrophic with at least 86,000 casualties. The Pakistan Meteorological Department estimated a 5.2 magnitude on the Richter scale, whereas the United States Geological Survey (USGS) measured its magnitude as at least 7.6 on the moment magnitude scale, classifying the quake as major. Its epicentre lay about 19 km north-east of the city of Muzaffarabad, its hypocenter was located at a depth of 26 km below the surface.

Such big earthquakes are accompanied by many aftershocks. For example in 2005, the city of Karachi (more than 1,000 km away from the epicentre) experienced a minor aftershock. There were many secondary earthquakes in the region, mainly to the north-west of the epicentre. A total of 147 aftershocks were registered in the first day after the initial quake. On October 19, a series of strong aftershocks occurred about 65 km north-northwest of Muzaffarabad. As an aside, note that aftershocks can even be stronger than the main earthquake itself.

In addition to such major shocks, that are still relatively rare, many smaller shocks have been recorded (see http://earthquake.usgs.gov/earthquakes for a list of earthquakes since 1973). The majority of tectonic earthquakes originate at depths not exceeding tens of kilometres. Those occurring at a depth of less than 70 km are classified as shallow. Earthquakes that originate below this upper crust are classified as intermediate or deep. See Molnar and Chen (1982) for further details. Clearly, the impact of an earthquake depends on its epicentre, its depth as well as its magnitude. Minor earthquakes occur very frequently and may not even be noticed or recorded. Therefore, we focus on those having a magnitude of at least 4.5 for which records are believed to be exhaustive (Van der Meijde, pers. comm.).

To summarize, our data (available from the authors on request) consist of the annual patterns of shallow earthquakes of magnitude 4.5 or higher in Pakistan during the period 1973–2008. The country level is appropriate, as most political and risk management actions are taken at this level. However, to avoid edge effects, we must sometimes refer to data on earthquakes across the border. For each event, its location and magnitude is recorded. For the major earthquake years 1997 and 2005, also the times at which shocks occurred in the month following the main one are available.

In Fig. 1, we plot the annual number of such earthquakes per square degree latitude-longitude (y-axis) over the period 1973–2008 (x-axis). Note that the clearly visible outlier corresponds to the Kashmir earthquake in 2005 that generated a large number of aftershocks. The number of aftershocks in 1997 was considerably less and their pattern more diffuse.

A scaled histogram of the observed magnitudes is presented in Fig. 2. In accordance with the Gutenberg–Richter power law, we fit a shifted exponential probability density $\beta e^{-\beta(m-4.5)}$ for $m \geq 4.5$ and 0 elsewhere. The rate parameter β can be interpreted as follows: $1/\beta$ is the expected excess magnitude with respect to the threshold value 4.5. For our data, the mean excess is 0.354 so the unknown parameter can be estimated by $\hat{\beta} = 2.28$. Comparing the graph of the fitted shifted exponential probability density to the histogram indicates an adequate fit (Fig. 2).

Fig. 1 Annual total number of shallow earthquakes per square degree latitude–longitude
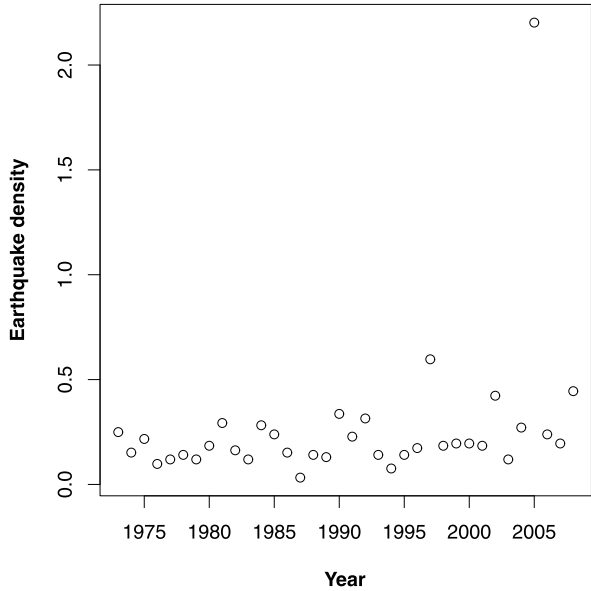

Fig. 2 Scaled histogram of earthquake magnitudes and fitted shifted exponential density (*black line*)
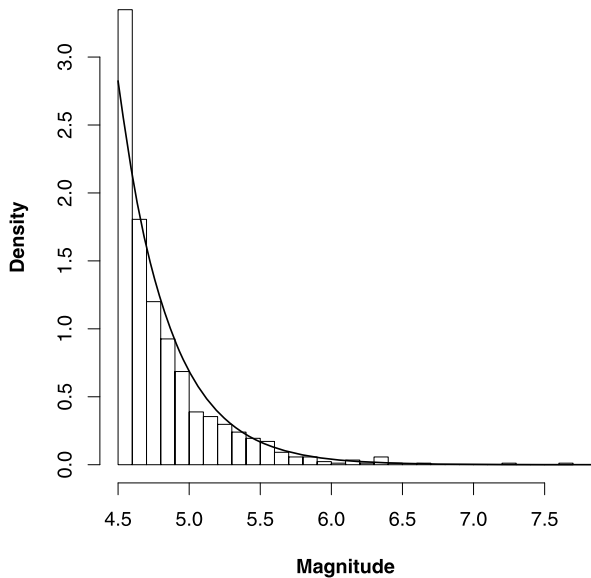
3 Spatial Intensity

Our first goal is to find the pooled intensity map of earthquake occurrences in Pakistan. To avoid edge effects, earthquake locations in neighbouring countries within a distance of about one degree from the Pakistan border are also taken into account. The aggregated patterns are shown in Fig. 3.
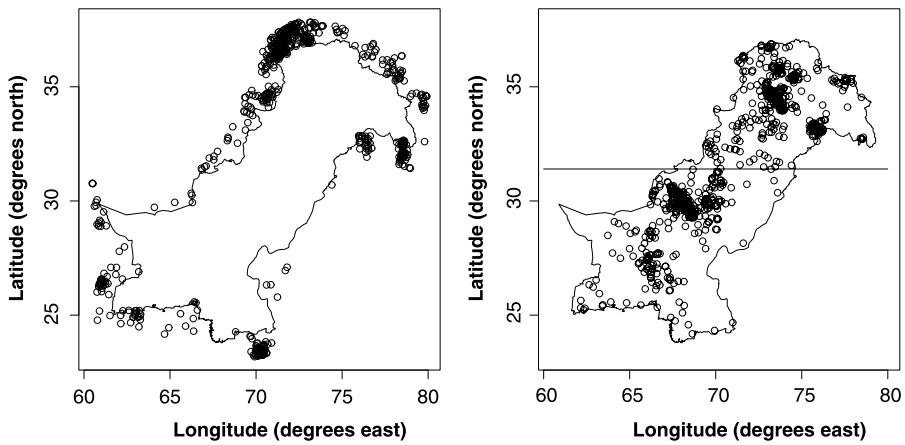
Fig. 3 Earthquake locations in (*left*) a zone up to about one degree removed from the Pakistan border and (*right*) within the Pakistan territory during the years 1973–2008. The line in the rightmost panel is at 31.4 degrees latitude
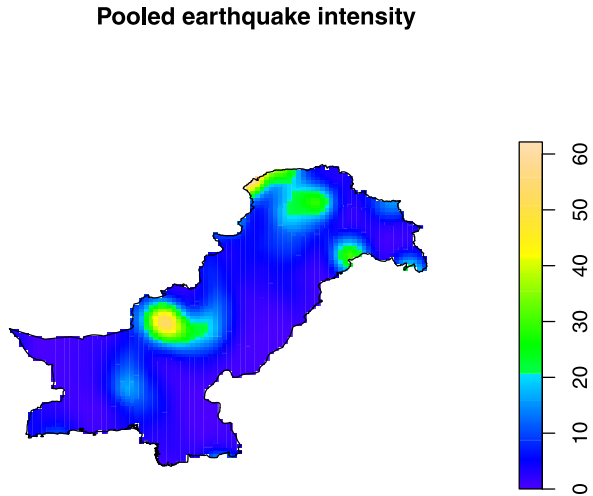
We proceed as follows. We first exclude the major earthquake years 1997 and 2005 and pool the locations of earthquakes in Pakistan with those in neighbouring countries closer than approximately one degree to the Pakistan border over the remaining 34 years together. Next, we calculate the kernel estimator of intensity (Diggle 1985; Gelfand et al. 2010, Chap. 18.5) using an isotropic Gaussian kernel with standard deviation 0.5. As always, choosing the standard deviation is a trade-off between bias and variance. Exploring values of 1.0 and 0.25, we noticed the absence of the earthquake prone area in Kashmir and a local emphasis on small isolated areas, respectively. Too large a value hence obscures important features, whereas a too small value may lead to spurious results. Our choice corresponds to a range of approximately 50 km, a value that is well in line with the spatial extent of zones affected by a major earthquake. We finally show the result in Fig. 4, which maps the estimated intensity of earthquakes during the years 1973–2008\{1997, 2005} per square degree latitude–longitude in the Pakistan territory.

High intensity parts occur in the north of the country, in the region bordering Afghanistan and Tajikistan, near the junction of plate boundaries, and in a smaller region in the east. A second zone of high earthquake activity lies in the mid-west of the country. In fact, the epicentre of the 1997 earthquake is located in this area.

4 Annual Relative Earthquake Rates

Geological considerations as well as Fig. 4 suggest enhanced earthquake intensity in the northern and mid-western parts of the country. In this section, we test whether this pattern persists over the years. To do so, write $W \subset \mathbb{R}^2$ for the set representing the Pakistan territory and divide W in two disjoint regions A_N and A_S of equal estimated intensity mass (cf. Fig. 4). More precisely, A_N is the subset of W lying north of the 31.4° latitude line, whereas $A_S = W \setminus A_N$ is the subset of W lying south of that line.

Pooled earthquake intensity

Next, we introduce the random variables Y_i as the fraction of shocks above $31.4°$ latitude in year $1972 + i$ for $i \in \{1, \ldots, 36\}$ and apply a test developed by Kwiatkowski et al. (1992) known by the acronym KPSS referring to the first characters of the authors' surnames. Let $\bar{Y}_n$ denote the mean $\sum_{i=1}^{n} Y_i / n$ of the time series $(Y_i)_i$ of length n (in our case, $n = 36$). Then the KPSS test statistic T_n is based on the partial sums process $S_n(i) = \sum_{j=1}^{i} (Y_j - \bar{Y}_n)$, $i \in \{1, \ldots, n\}$, as follows

$$T_n = \frac{1}{n^2 s_n^2} \sum_{i=1}^{n} S_n(i)^2.$$

The factor s_n^2 is the long run expectation of $S_n(n)^2/n$. See Herrndorf (1984), Newey and West (1987), Phillips and Perron (1988), or De Jong (2000) for the theoretical properties of T_n. In particular, for large n the test statistic T_n is approximately equal in distribution to the integral of the squared Brownian bridge. Critical values of the test are reported in Table 1 in Kwiatkowski et al. (1992).

For the data discussed in Sect. 2, the test statistic takes the value 0.1297 with a p-value exceeding 10%, so the null hypothesis cannot be rejected. We conclude that there is no statistical evidence of a temporal trend in intensity patterns of shallow earthquakes based on the north-south divide during the 36 years of study. Clearly, we could have divided W into more than two subsets, but in some years the number of events is so low (cf. Fig. 1) that we did not choose to do so. We also refrained from pooling years together to boost the counts as this would lead to a small value of n.

5 The J-Function for Inhomogeneous Point Processes

The next step is to investigate whether a series of inhomogeneous Poisson point processes could explain the annual patterns of earthquakes. For this purpose, we use the Monte Carlo approach (Besag and Diggle 1977; Diggle 1979) in combination

with the inhomogeneous J-function (Van Lieshout 2011). For alternative exploratory functions, see Baddeley et al. (2000), Adelfio and Schoenberg (2009), Gabriel and Diggle (2009), Adelfio and Chiodi (2010), or Gelfand et al. (2010).

The J-function is based on the idea to compare the distances to the nearest (other) point as seen from respectively a point in the mapped pattern and an arbitrarily chosen origin in space in order to gain an understanding of the interaction structure of the point process that generated the data. For example, a clustered pattern is characterized by small inter-point distances but relatively large distances to the origin due to the gaps in between the clusters. Let X be a homogeneous point process on $\mathbb{R}^2$ without multiple points observed in some bounded set W. Write $d(x, X)$ for the distance from $x \in \mathbb{R}^2$ to X. The idea is then to compare $\mathbb{P}(d(0, X) > t)$ to $\mathbb{P}(d(0, X \setminus \{0\} > t \mid 0 \in X)$, where because of stationarity 0 may be replaced by any other point. A naive estimator of the first term would be to take each point l_k of a finite point grid L in W as origin and to compute the fraction that have no nearest neighbour within distance t, that is,

$$\frac{\sum_{l_k \in L} 1\{d(l_k, X) > t\}}{\#L} = \frac{\sum_{l_k \in L} \left(\prod_{x \in X \cap B(l_k, t)} 0 \right)}{\#L} \tag{1}$$

under the convention that an empty product is equal to one. Here, $\#L$ denotes the number of grid points, $B(l_k, t)$ the closed ball of radius t centred at l_k. To see (1), observe that the product in the hand right hand side is equal to one if and only if $X \cap B(l_k, t) = \emptyset$ if and only if $d(l_k, X) > t$. Note though that for l_k closer than t to the boundary ∂W of W it is not possible to decide whether its nearest neighbour is at least a distance t away. Therefore, we have to restrict ourselves to those l_k that lie in the eroded set $W_{\ominus t} = \{(x, y) \in W : d((x, y), \partial W) \geq t\}$ of points within W that are at least a distance t away from the boundary. Similarly, $\mathbb{P}(d(0, X \setminus \{0\} > t \mid 0 \in X)$ can be estimated by summing over the points of X rather than L in so far as they lie in $W_{\ominus t}$.

The point processes we consider are not homogeneous. Thus, we must modify the J-function so that it quantifies the spatial interaction after accounting for spatial variation. This is necessary as apparent clustering may be due to both spatial variation (earthquakes tend to happen near fault zones) and to inter-point attraction caused by aftershock sequences and swarms. To separate the two, replace the term 0 in the product in (1) by $[1 - \frac{\mu_0}{\mu(x,y)}]$ where $\mu(x, y)$ is the intensity function of X. We assume that the intensity function is bounded away from zero by $\mu_0 > 0$. Under this assumption one never divides by zero and the product is over terms that lie between 0 and 1. Moreover, in the homogeneous case $\mu(x, y) \equiv \mu_0$ and we retrieve the original formula. In summary,

$$\widehat{J_{\text{inhom}}(t)} = \frac{\frac{1}{\#X \cap W_{\ominus t}} \sum_{(x_k, y_k) \in X \cap W_{\ominus t}} \prod_{(x,y) \in X \setminus \{(x_k, y_k)\} \cap B((x_k, y_k), t)} [1 - \frac{\mu_0}{\mu(x,y)}]}{\frac{1}{\#L \cap W_{\ominus t}} \sum_{l_k \in L \cap W_{\ominus t}} \prod_{(x,y) \in X \cap B(l_k, t)} [1 - \frac{\mu_0}{\mu(x,y)}]}. \tag{2}$$

Under the Poisson null hypothesis, $J_{\text{inhom}} \equiv 1$. For further details, see Van Lieshout (2011).

In our context and based on the results of Sect. 4, we assume that the intensity function of X_i, the pattern of earthquakes happening in the ith year $1972 + i$, $i =$

Fig. 5 $\widehat{J_{\text{inhom}}}(t)$ (y-axis) plotted as a *solid line* against t (x-axis, in degrees) for the pattern of all earthquake locations in normal years with upper and lower envelopes (*broken lines*) based on 19 independent realisations of an inhomogeneous Poisson process
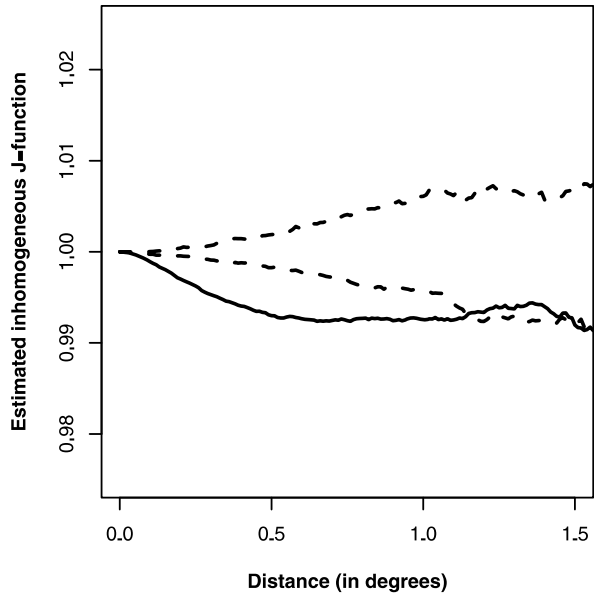

$1, \ldots, 36$, is of the form

$$\mu_i(x, y) = w_i \, \tilde{\mu}(x, y),$$

where $\tilde{\mu}$ is supposed to be normalized, that is, $\int_W \tilde{\mu}(x) \, dx = 1$. Under this assumption, the weights w_i can be estimated by the number of shocks in the ith year. Furthermore, for a set I of year indices, the intensity function μ_I of $X_I = \bigcup_{i \in I} X_i$ is $\sum_{i \in I} w_i \, \tilde{\mu}(x, y)$. Note that Fig. 4 depicts an estimator $\hat{\mu}_I$ for the set of indices I representing $\{1973, \ldots, 2008\} \setminus \{1997, 2005\}$. In Fig. 5, we plot $\widehat{J_{\text{inhom}}}(t)$ against t with the unknown μ and μ_0 in (2) replaced by their estimates $\hat{\mu}_I(x, y)$ and $\hat{\mu}_0$, the minimal estimated intensity in W. Also shown are the envelopes formed by the largest and smallest values of the estimated inhomogeneous J-function observed among 19 independent samples from an inhomogeneous Poisson process with the same intensity function. We conclude that X_I is more clustered than its Poisson counterpart up to $t \approx 1$ degrees. For the impact of the estimators used in plug-in ratio estimators such as (2), see Stoyan and Stoyan (2000).

For the pooled data considered in Fig. 5, we were forced to use the same data to estimate the intensity and J_{inhom} functions. For individual years, this can be avoided. As an illustration, consider the last three years. To estimate the intensity function, we exclude the year of interest as well as 1997 and 2005, then calculate the kernel estimator using an isotropic Gaussian kernel with standard deviation 0.5 taking into account also earthquakes happening within a distance of about one degree away from the Pakistan border and normalize the result so as to place unit mass on W. The intensity function for year i is obtained by scaling by the number of earthquakes in year i. Therefore, the intensity function estimator is mass preserving. The resulting $\widehat{J_{\text{inhom}}}(t)$ are plotted against t in Fig. 6. In 2006, the empirical J_{inhom}-function lies below the envelopes, indicating clustering above that due to the inhomogeneity;
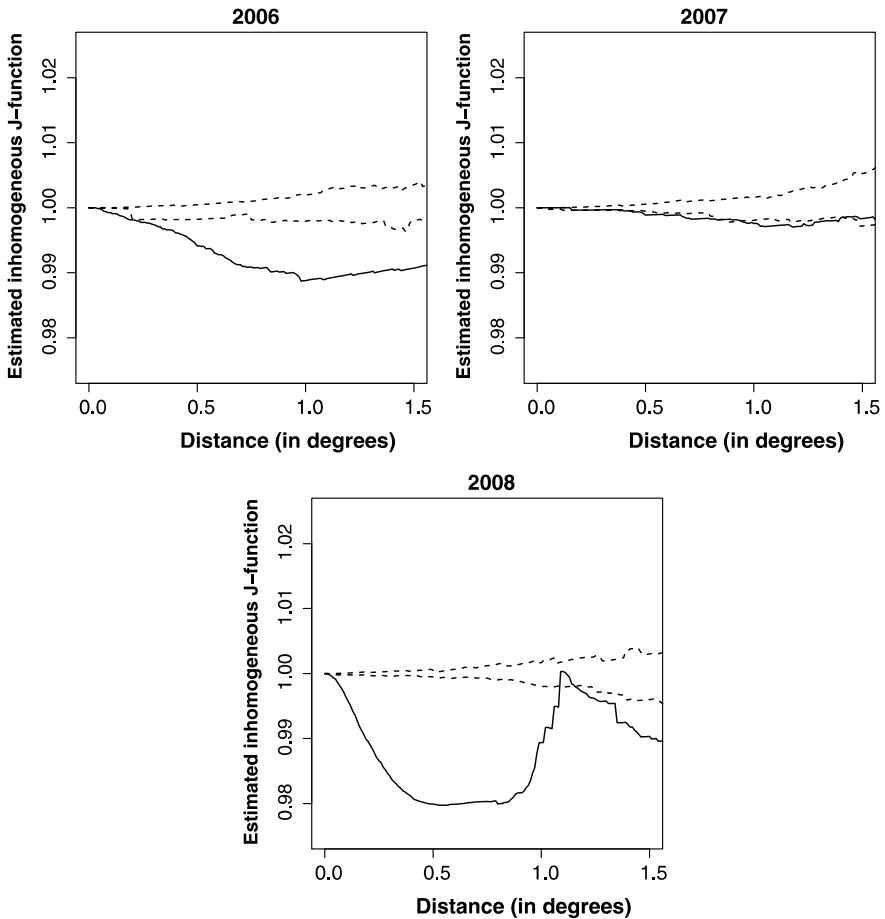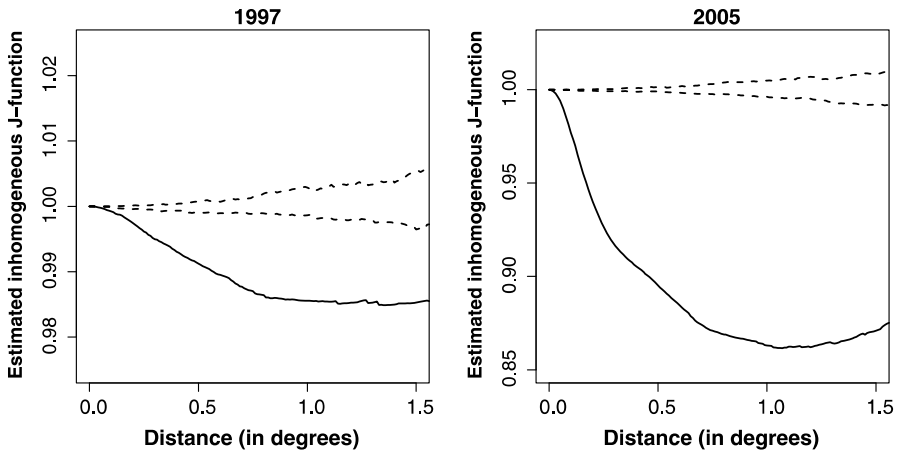
Fig. 6 $\widehat{J_{\mathrm{inhom}}}(t)$ (y-axis) plotted as *solid lines* against t (x-axis, in degrees) for the patterns of earthquakes in 2006 (*leftmost frame*), 2007 (*middle frame*) and 2008 (*rightmost frame*) with upper and lower envelopes (*broken lines*) based on 19 independent realisations of an inhomogeneous Poisson process

the pattern in 2008 also exhibits strong attraction between the points, especially for small t. In 2007, there is significant but milder clustering at intermediate range.

For the years in which an earthquake of magnitude of at least seven occurred, that is, for 1997 and 2005, the estimated inhomogeneous J-function strongly suggests clustering after accounting for the effect of spatial variability, cf. Fig. 7.

6 Aftershocks in the Major Earthquake Years

From the two previous sections, we conclude that there is no evidence for a temporal trend, but that there appears to be clustering in the data that cannot be attributed to geological inhomogeneity. In order to be able to formulate a cluster model, information about the spread of aftershocks around a main event is needed. We therefore focus

Fig. 7 $\widehat{J_{\mathrm{inhom}}}(t)$ (*y*-axis) plotted as *solid lines* against *t* (*x*-axis, in degrees) for the patterns of earthquakes in 1997 (*leftmost frame*) and 2005 (*rightmost frame*) with upper and lower envelopes (*broken lines*) based on 19 independent realisations of an inhomogeneous Poisson process

on the two years (1997 and 2005) in which earthquakes of magnitude larger than 7 occurred and for which we have access to more detailed data than for other years. The scientific literature distinguishes a general earthquake intensity pattern from a pattern of aftershocks. These usually form a concentrated cluster during a relatively short time interval after a major earthquake event. The 2005 Kashmir earthquake was studied by Anwar et al. (2012). They found that the seismicity in northern Pakistan decreased sharply following the shock on October 8. In particular, the number of earthquakes (of magnitude at least 4) that occurred more than a month after the first shock was found to be negligible compared to the number in the first month. We therefore concentrate on the 176 events (out of a total of 203 earthquakes during the whole of 2005) in our data that happen between October 8, 2005, and November 7, 2005. A histogram of the time (counted in days from October 8th) of these earthquakes is given in Fig. 8. The figure confirms the picture painted above: a large number of aftershocks occurred during the first couple of days following the main shock, with a steep decline in numbers afterward.

Figure 9 depicts the locations of aftershocks, which can be well described by two clusters: one corresponding to the main earthquake, the other to the next largest earthquake with a magnitude equal to 6.4 occurring approximately seven hours later. Pretending that the displacements in the two clusters follow a two-dimensional Gaussian distribution with mean zero and equal covariance matrices $\sigma^2 I_2$ for some $\sigma^2 > 0$, we may apply Fisher's linear discriminant function to assign the aftershocks to either of the clusters. Here, I_2 denotes the two by two identity matrix. The result is presented in Fig. 9.

From the picture, it is clear that the variances of the displacements are not identical, but nevertheless a visually good separation is achieved. The estimated within group variances are given in Table 1.

Anwar et al. (2012) did not consider the year 1997. In contrast to the pattern of aftershocks in 2005, in 1997 there is no clear decrease in aftershock occurrence fol-

Fig. 8 Histogram of times (in days from October 8, 2005)
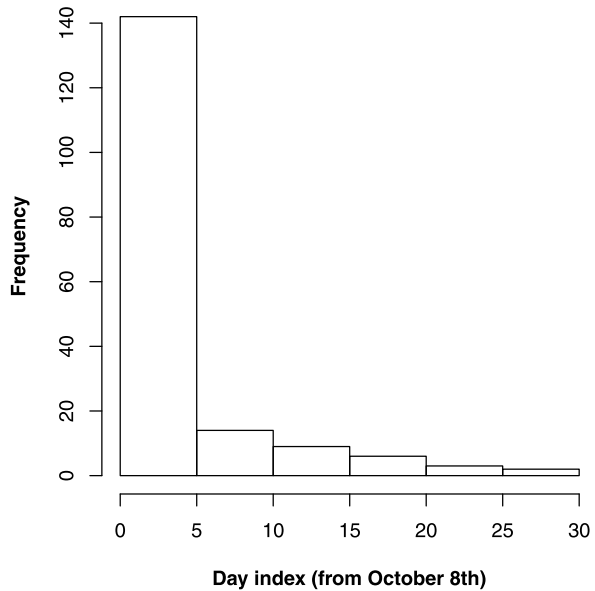
Fig. 9 Earthquake locations occurring in the period October 8–November 7, 2005. The locations of the two largest shocks are identified by a *crossed box*; their clusters are colour coded with *blue* for the main shock and *red* for the shock of second largest magnitude
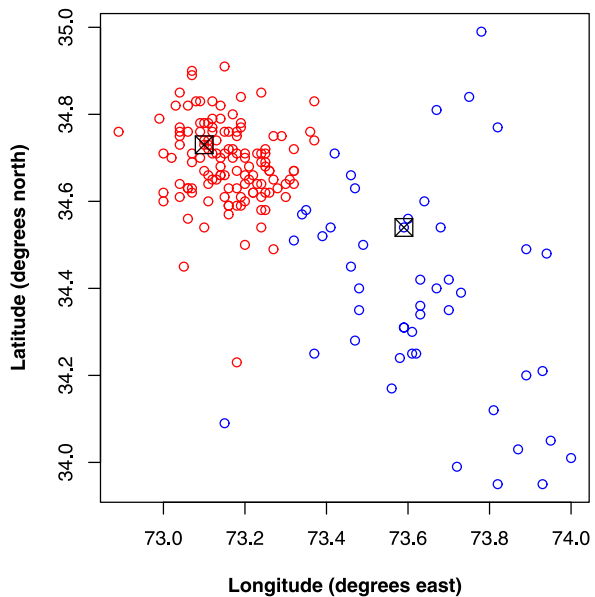
lowing the main shock and the pattern is more diffuse, a type of behaviour typical of a swarm. We therefore extracted the cluster by hand and estimated the variance of the displacements from the epicentre in longitude and latitude (see Table 1). Instead of looking at displacements with respect to the epicentre, we could have computed the sample variance, which would have led to slightly smaller values but would be difficult to integrate with the values obtained for 2005. The pooled within groups sample

Table 1 Estimated within group variances in longitude ($\hat{\sigma}_x^2$) and latitude ($\hat{\sigma}_y^2$) for aftershock clusters in the years 1997 and 2005

	$\hat{\sigma}_x^2$	$\hat{\sigma}_y^2$
1997	0.0817	0.0969
2005	0.0394	0.0813
2005	0.0117	0.0105

variance is 0.038 (standard deviation 0.19). It may be conjectured that the spread of aftershocks is related to the magnitude of the earthquakes (cf. the difference in aftershock patterns between the two largest shocks in 2005) but we do not have enough data to support this conjecture.

7 Model

We are now in a position to formulate a multivariate marked point process model for the aggregated Pakistan data described in Sect. 2. Taking into account that aftershocks occur over relatively short periods of time only and the low point counts in most years, we set $Z = (Z_1, \ldots, Z_{36})$ where the Z_i are independent but not identically distributed marked point processes with locations in W and marks in $[4.5, \infty)$ that represent the magnitudes. Thus, Z_i represents the earthquakes in year $1972 + i$. For the marks, we assume random labelling according to a shifted exponential distribution $f_M(m) = \beta \exp[-\beta(m - 4.5)]$, $m \geq 4.5$, cf. Sect. 2.

We model the marked point processes Z_i, $i = 1, \ldots, 36$, as Poisson cluster processes: each 'parent' generates a Poisson number of offspring with a mean number $A(m)$ that depends on the magnitude m of the parent. The offspring locations are independent and normally distributed with probability density $f_N(\cdot - (x, y))$ centred at the parent location (x, y) and having covariance matrix $\sigma^2 I_2$ where I_2 is the 2×2 identity matrix (cf. Sect. 6). Note that a parent in the Pakistan territory W may generate offspring across the border, and that some earthquakes recorded in Pakistan may arise from a parent in a different country. We therefore assume that the parent locations in the i-th year form a point process on the set $W_b \supseteq W$ consisting of W and a buffer zone large enough to make the probability of a parent in $\mathbb{R}^2 \setminus W_b$ generating offspring in W negligible, and suppose it is Poisson with locally finite intensity measure $\alpha_i \lambda(x, y)$, $i = 1, \ldots, 36$, $(x, y) \in W_b$. For identifiability reasons, we normalize the process so that λ is a probability density on W, i.e., $\int_W \lambda(x, y)\, dx\, dy = 1$.

We base inference on the first order moment measure. By Proposition 6.3.III in Daley and Vere-Jones (2003), the intensity function of X_i, the observed earthquake locations in $1972 + i$ can be written as

$$\mu_i(x, y) = \alpha_i \int_{4.5}^{\infty} A(m) f_M(m)\, dm \int_{W_b} f_N\big((x, y) - (u, v)\big) \lambda(u, v)\, du\, dv. \quad (3)$$

The joint intensity function of locations and marks in the i-th year at $((x, y), k)$ is simply $\mu_i(x, y) f_M(k)$. Equation (3) should be seen in the light of Sect. 5: for I

Fig. 10 Map of earthquake
hazard per square degree
latitude–longitude

Hazard map
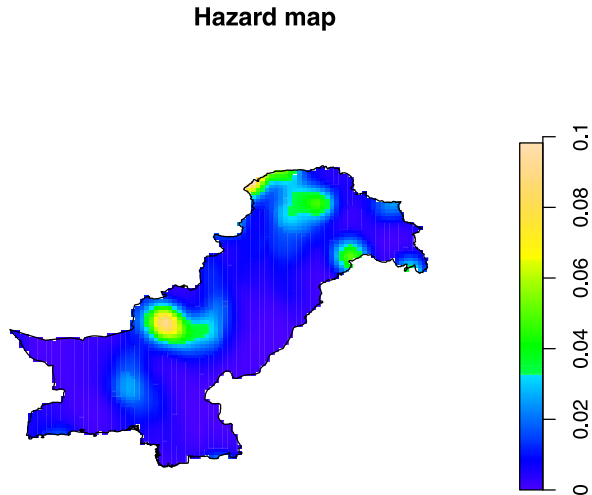
representing $\{1973, \ldots, 2008\} \setminus \{1997, 2005\}$,

$$\mu_I(x, y) = \left(\sum_{i \in I} \alpha_i\right) \int_{4.5}^{\infty} A(m) f_M(m) \, dm \int_{W_b} f_N\big((x, y) - (u, v)\big) \lambda(u, v) \, du \, dv.$$

Therefore, μ_I is proportional to a convolution of the normal density f_N with λ. We refrain from formulating a model for $A(m)$ due to the paucity of data on cluster sizes resulting from a parent event of magnitude m.

It remains to estimate λ. To do so, we work in the Fourier domain. The Fourier transform of μ_I is proportional to the product of those of f_N and λ. Plugging in the estimators for μ_I (Fig. 4) and σ, the standard deviation of the normal distribution (Sect. 6), transforming back and normalizing yields an estimator of λ. The resulting hazard map $\hat{\lambda}$ is shown in Fig. 10. The hot spots reflect the geology, with high values of $\hat{\lambda}$ along the convergence zones (cf. Sect. 2). The region where the 1997 earthquake occurred is clearly visible, as are the earthquake prone areas in the north.

8 Hazard Map Assessment

To validate the hazard map (Fig. 10), ideally we would like to proceed as in Sect. 5. However, in order to sample from the cluster process model of Sect. 7, we need to know how many events originate from a parent of given magnitude. Since this is not feasible based on the data at our disposal, we take the leave-one-out approach that enables us to assess whether any particular year has an undue influence on the hazard. Thus, for every normal year (labelled i) in which no earthquake of magnitude larger than 7 occurs, we estimate the intensity function $\mu_{I \setminus \{i\}}$ and de-convolve to obtain another estimate $\widehat{\lambda_{(-i)}}$ of the hazard map λ (cf. Sect. 7). Here, as before I contains the indices representing all normal years, that is, those except 1997 and 2005. Should we find a large difference for a particular year, this would indicate

Fig. 11 Integrated squared
difference for hazard (*circles*)
and normalized intensity
(*crosses*) using the
leave-one-out principle applied
to the years 1973–2008
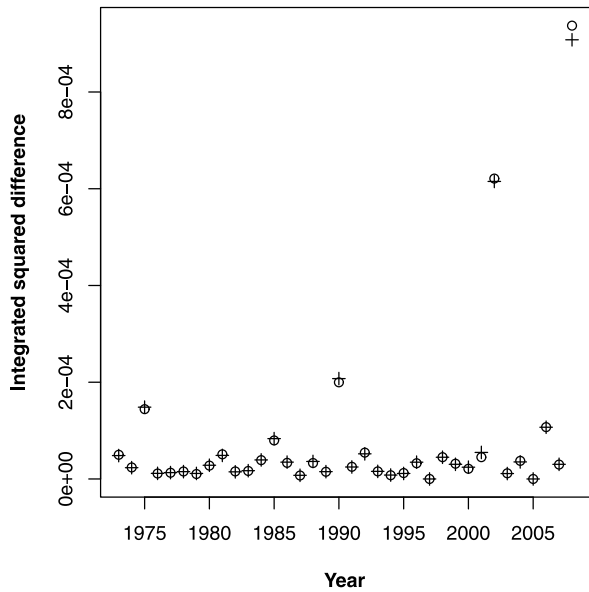
that we had mistakenly labelled that year as normal. Such validation procedures are common in modelling studies (Hall 1982; Efron 1982). For exact interpolators such as kriging, care is taken that individual points are left out. In estimating earthquake intensities, the effects of leaving them out individually would lead only to minor changes in value that supposedly are of no relevance in the final conclusion and would slow down the computations considerably. A graph of

$$\int_W \left(\widehat{\lambda_{(-i)}(x, y)} - \widehat{\lambda(x, y)} \right)^2 dx\, dy,$$

where $\hat{\lambda}$ is the hazard map estimated using all years except the two years in which a major earthquake occurred, plotted against i is given in Fig. 11. The same figure also contains the integrated squared difference between the normalized intensity function $\tilde{\mu}$ (cf. Sect. 5) estimated using all normal years and using those in $I \setminus \{i\}$, $i \in I$, only.

It can be seen from Fig. 11 that the effect of leaving a year out is small. For most years, the integrated squared difference is smaller than 0.0002. Since Pakistan has a surface area of 92 square degrees, this amounts to 2.2×10^{-6} per square degree latitude-longitude. For comparison, the mean value of $\hat{\lambda}$ is 0.01. The outliers are the years 2002 and 2008 in which small clusters of shallow earthquakes occur, but the values are still quite small. We conclude that there is no need to delete any particular year from our analysis.

9 Discussion

In this manuscript, we addressed the occurrence of earthquakes at the national scale. We focused on several questions that are relevant within this context. Earthquakes,

like most natural disasters, are not forced to have their effects in a single country, and we have included disasters that occurred across the different borders in our considerations, as much as these might have an effect in the country. With the advent of a more inter-regional approach, however, a similar analysis might be done where attention focuses, for example, on a group of countries around one particular fault. Adversely, a similar analysis might be applicable as well to a region within a country that is particularly vulnerable.

An interesting issue that we discovered in passing when carrying out the analyses was that the major shock in 2005 could be modelled in a more convincing way when we applied a double model. The major shock was followed by another large shock, and both generated aftershocks in an almost perfect way. The second shock and all its aftershocks could, alternatively, have been included as a set of aftershocks of the first event. This raises the issue at which stage one should distinguish a shock from an aftershock. Although these terms are intuitively clear, we were not successful in finding a sharp and unambiguous definition in the literature.

Relevant information such as population density could be taken into account, but that was not available to us when carrying out the analysis. Similarly, the transition of shock waves through the earth crust could serve to support the model that we applied. This would also require additional data, in particular referring to the geological complexities. We felt that this was outside the scope of the current study, and may be of little relevance as well when the emphasis would be on the support after the occurrence of a disaster.

The approach described in the paper could be modified by including such additional variation. There is also an interesting issue of data quality related to this study, i.e., the restriction to earthquakes of a magnitude larger than 4.5. It is reported in the literature that earthquakes of a magnitude less than 4.5 are less well estimated. When studying the distribution of all earthquakes of size 4.0 and larger, we found that the histogram of the occurrences was deviating from the presumed exponential model. Therefore, this theoretical model served as a tool to identify reliability of the observed earthquakes and thus confirmed that earthquakes of a lower magnitude were less reliably recorded. A similar issue was raised when considering the depths of occurrence of the earthquakes, where particular depths were occurring at a large frequency. In fact the depths of 10 km occurred in nearly 30% of all earthquakes, the depth of 33 km in over 40%. For this reason, we did not include the depth of the earthquakes as a reliable parameter: Further studies have to show how depth related information can be included into the analysis.

10 Conclusions

This paper presents a range of tools to analyze the occurrence pattern of earthquakes at the national scale. The paper leads to several conclusions. First, a test has been developed that allows one to check whether a pattern of point-like events (such as earthquakes) shows homogeneity through a range of years. This test has been based upon the KPSS test statistic and for our data showed no indication of a spatial trend over the study period. Second, the inhomogeneous J-function has been applied for

the first time to quantify spatial interaction in the apparently inhomogeneous pattern of earthquakes within the borders of a single country. Even when accounting for non-homogeneity, additional clustering was found to be present. Third, a relatively standard analysis of the aftershocks that occurred during two years of severe earthquakes in the country was carried out. We found that in one year a mixture of two major shocks fitted the data better than a model in which all aftershocks arose from a single earthquake. Finally, a multivariate marked point process model was formulated in which the components were Poisson cluster processes with normally distributed offspring. A Fourier analysis of the spatial intensity of aggregated earthquakes was carried out to obtain a hazard map. We found hot spots in the northern and midwestern parts of the country. In all, the modelling of earthquakes at the national level of Pakistan turned out to be an exciting and relevant analysis of a spatio-temporal point pattern. We are convinced that much additional modelling can and should be done, and that a proper interpretation of the derived results can lead to a careful management of cities and infrastructure that may safe lives during future disasters.

References

Adamopoulos L (1976) Cluster models for earthquakes: regional comparisons. Math Geol 8(4):463–475

Adelfio G, Chiodi M (2010) Diagnostics for nonparametric estimation in space-time seismic processes. J Environ Stat 1(2):1–13

Adelfio G, Schoenberg FP (2009) Point process diagnostics based on weighted second order statistics and their asymptotic properties. Ann Inst Stat Math 61(4):929–948

Anwar S, Stein A, van Genderen JL (2012) Implementation of the marked Strauss point process model to the epicenters of earthquake aftershocks. In: Shi W, Goodchild M, Lees B, Leung Y (eds) Advances in geo-spatial information science. Taylor & Francis, London, pp 125–140. ISBN 978-0-415-62093-2

Baddeley A, Turner R (2005) Spatstat: an R package for analyzing spatial point patterns. J Stat Softw 12(6):1–42

Baddeley AJ, Møller J, Waagepetersen R (2000) Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. Stat Neerl 54(3):329–350

Besag J, Diggle PJ (1977) Simple Monte Carlo tests for spatial pattern. Appl Stat 26(3):327–333

Daley DJ, Vere-Jones D (2003) An introduction to the theory of point processes. Volume I: Elementary theory and methods, 2nd edn. Springer, New York

Daley DJ, Vere-Jones D (2008) An introduction to the theory of point processes. Volume II: General theory and structure, 2nd edn. Springer, New York

De Jong RM (2000) A strong consistency proof for heteroskedasticity and autocorrelation consistent covariance matrix estimators. Econom Theory 16(2):262–268

Diggle PJ (1979) On parameter estimation and goodness-of-fit testing for spatial point patterns. Biometrics 35(1):87–101

Diggle PJ (1985) A kernel method for smoothing point process data. Appl Stat 34(2):138–147

Efron B (1982) The jackknife, the bootstrap, and other resampling plans. SIAM, Philadelphia

Gabriel E, Diggle PJ (2009) Second-order analysis of inhomogeneous spatio-temporal point process data. Stat Neerl 63(1):43–51

Gelfand AE, Diggle PJ, Fuentes M, Guttorp P (2010) Handbook of spatial statistics. CRC, Boca Raton

Hall P (1982) Cross-validation in density estimation. Biometrika 69(2):383–390

Herrndorf N (1984) A functional central limit theorem for weakly dependent sequences of random variables. Ann Probab 12(1):141–153

Holden L, Sannan S, Bungum H (2003) A stochastic marked point process model for earthquakes. Nat Hazards Earth Syst Sci 2003(3):95–101

Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root. J Econom 54:159–178

Lucio PS, De Brito NLC (2004) Detecting randomness in spatial point patterns: a "Stat-Geometrical" alternative. Math Geol 36(1):79–99

Molnar P, Chen WP (1982) Seismicity and mountain building. In: Hsu KJ (ed) Mountain building processes. Academic Press, London, pp 41–58

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55(3):703–708

Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. J Am Stat Assoc 83(401):9–27

Ogata Y (1998) Space-time point-process models for earthquake occurrences. Ann Inst Stat Math 50(2):379–402

Pei T (2011) A nonparametric index for determining the numbers of events in clusters. Math Geosci 43:345–362

Pei T, Zhua AX, Zhou C, Li B, Qin C (2007) Delineation of support domain of feature in the presence of noise. Comput Geosci 33:952–965

Phillips PCB, Perron P (1988) Testing for a unit root in time series regression. Biometrika 75(2):335–346

Shurygin AM (1993) Statistical analysis and long-term prediction of seismicity for linear zones. Math Geol 25(7):759–772

Stoyan D, Stoyan H (2000) Improving ratio estimators of second order point process characteristics. Scand J Stat 27(4):641–656

Trapletti A, Hornik K (2009) tseries: time series analysis and computational finance. R package version 0.10-22

Van Lieshout MNM (2011) A J-function for inhomogeneous point processes. Stat Neerl 65(2):183–201

Vere-Jones D (1970) Stochastic models for earthquake occurrence. J R Stat Soc B 32(1):1–62

Vere-Jones D, Musmeci F (1992) A space-time clustering model for historical earthquakes. Ann Inst Stat Math 44(1):1–11

Zhuang J, Ogata Y, Vere-Jones D (2002) Stochastic declustering of space-time earthquake occurrences. J Am Stat Assoc 97(458):369–380