Università
Ca'Foscari
Venezia

**Dipartimento
di Scienze Ambientali
Informatica e Statistica**

Technical Report Series

Rapporto di Ricerca DAIS-2013-1

Gennaio 2013

A. Albarea

Statistical analysis of historical earthquake catalogues: California vs. Italy

Statistical analysis of historical earthquake catalogues: California vs. Italy*

Andrea Albarea
Department of Environmental Science, Informatics and Statistics
Ca' Foscari University of Venice

January 16, 2013

1 Introduction

Data depth is a statistical method that allows to order the points of a given space according to centrality with respect to an assumed probability distribution. The idea of using data depth to study the spatial distribution of earthquake epicenters was put forward by Small (Small, 1990). More specifically, his intuition was inspired by an illustration of data depth of directional data. He argued that a possible application was to study the spatial distribution of earthquake epicenters on the Earth' surface. California and Italy are, historically, regions with an important seismic activity. The aim of this work is to provide a map of seismic risk for both countries using the relevant earthquake catalogues. Statistical methods include the centrality ordering (Liu et al, 1999) of data depth and kernel density estimation (Silverman, 1986). We start with a preliminary descriptive analysis of the catalogues, then we study the spatial distribution of epicenters and finally we consider the joint distribution of the geographical coordinates and the magnitude of the shocks to obtain a comprehensive investigation of the data. The fault structures of California and Italy are very different and our results allow a detailed illustration of the two situations. From the methodological point of view, the result the present study provide a first comparison of data depth and kernel density estimation as data smoothers.

2 Californian earthquake catalogue 1769 - 2000

The catalogue considered for this paper, called CDMG 2000, is a list of Californian earthquake during the period 1769 – 2000. Basically, it is an update of the CDMG 1996 catalogue (Petersen et al, 1996). The updating was done in two steps. The first

*Research project *Statistical analysis of historical earthquake catalogues*, supervisor M. Romanazzi. Financial support by DAIS, research funds of M. Romanazzi and PRIN 2008 research project *Approximate likelihood methods for high-dimensional dependencies*, coordinator P. Vidoni

step was to extend the CDMG 1996 to 2000, that for convenience we call NSCM 2000 . This was made possible thanks to the studies and the data bases produced by Oppenheimer D. of U.S. Geological Survey (USGS) and Kagan Y., Jackson D. and Rong Y. of University of California (UCLA) at Los Angeles. The second step was to merge NSCM 2000 with the CDMG map sheet 49 for events with magnitude equal to, or greater than, 5.5 (Toppozada et al, 2002). This added about 50 pre-1932 events to the NSCM 2000. [1] It must be noted that in the resulting catalogue slightly different magnitude scales are used but, at any rate, all magnitudes greater than 6.0 are moment magnitudes. For more details on earthquake magnitude scales in use, a convenient reference is (Woessner et al, 2010).

The threshold magnitude for CDMG 2000 starts from values equal to, or greater than, 4.0. The catalogue extends for about 100 km beyond the state borders to include events in neighboring states or offshore that could have caused damage in California. For the sake of comparability with Italian catalogue (see Section 4), events of magnitude not smaller than 4.5 were extracted from the catalogue resulting in 2206 events over the threshold during the period 1769 – 2000.

Variables used in the analysis are:

- TIME: occurrence time of each shock,
- LATITUDE, LONGITUDE: geographical coordinates of the epicenter of earthquake,
- INTENSITY: magnitude (recall that only shocks with magnitude not less than 4.5 are considered).

The very famous San Andreas fault is a tectonic structure which extends for about 1300 km across California, at the boundary between the North American plate and the Pacific plate. It was discovered in 1895 by Andrew Lawson, Professor of Geology at the University of Berkeley, who took this name from a small lake, Lagoon of San Andreas, located on a valley formed by the fault on the south of San Francisco.

The fault can be divided in three segments:

- southern segment or Mojave segment: it starts near Salton Sea and continues northwest to meet San Bernardino Mountains,
- central segment: the fault continues in a northwestern direction from Parkfield up to Hollister,
- northern segment: it starts at Hollister, crosses San Francisco peninsula, then follows California coast up to Cape Mendocino.

The most significant earthquake of the catalogue occurred in 1857, when a fracture of 350 km in central and southern California, from Parkfield to Cajon Pass, provoked Fort Tejon earthquake (moment magnitude: 7.9), which caused only 1 death. Another

[1] CDMG 2000 and map sheet 49 are available at http://www.conservation.ca.gov/cgs/rghm/quakes/Pages/Index.aspx and they are maintained by California Department of Conservation's (DOC) and by California Geological Survey (CGS).

devastating shock happened in 1906, when a fracture of 430 km in northern California, from San Juan Bautista to Eureka, provoked San Francisco Earthquake (moment magnitude: 7.8) which killed 3000 people and caused a lot of damage both for the earthquake and the fire broken out in San Francisco.

A further confirmation of the fact that along San Andreas Fault occurred great earthquakes comes from the catalogue. Indeed in just over two centuries there were over 25 shocks of moment magnitude 7.0 or greater.

3 Descriptive analysis of Californian catalogue

In this section the main features of Californian catalogue are illustrated with the aid of descriptive statistics. The histogram in Figure 1 confirms the typical decreasing shape of magnitude distribution, the mean and the median magnitude being 5.01 and 4.80, respectively. The difference between the minimum value (magnitude: 4.500) and the first quartile (magnitude: 4.502) is very small, implying that 25% of the observations fall just on the right of the minimum value. On the other hand, the 95% percentile of magnitude is 6.3, that is, 5% of the earthquakes occurred in the reference period have a magnitude equal to or greater than 6.3.

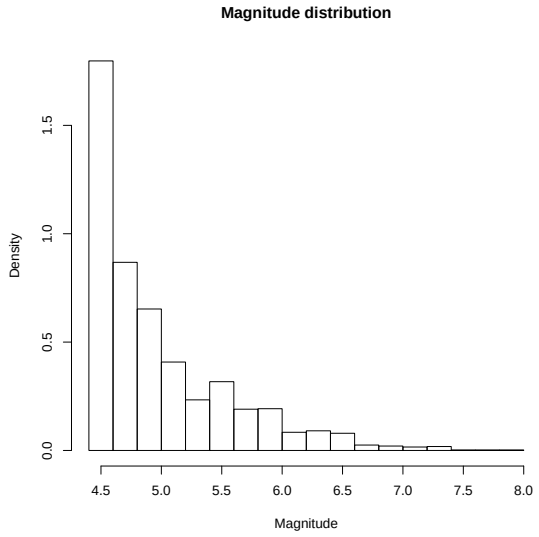


Figure 1: Californian catalogue. Histogram of magnitude.

The histogram in Figure 2 shows the distribution of earthquakes over time. The *increasing* shape must be interpreted with care because it is surely influenced by a more precise recording of events even at lowest magnitude levels determined by the improvement of technology in the reference period. However, the peak in the seismic activity in the period $(1940 - 1960]$ could be a real feature of the underlying process,

because at that time automatic recording of shocks was present throughout California.

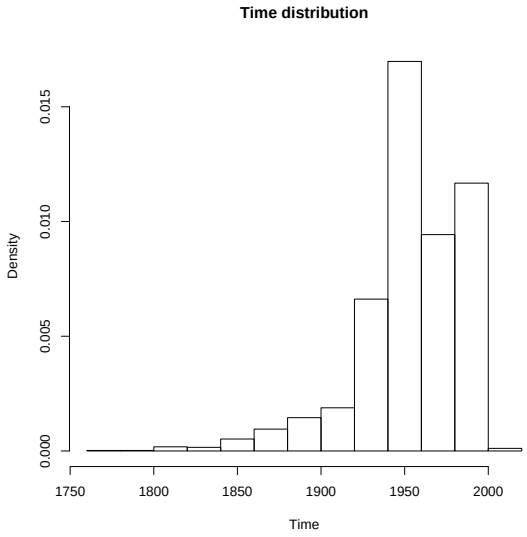


Figure 2: Californian catalogue. Histogram of time.

Previous considerations are confirmed by Table 1 that reports the percentage of events recorded in each half-century. Indeed 64.55% of the total number of earthquakes occurred in the last 50 years.

Time	%
[1769-1850]	0.91
(1850-1900]	5.71
(1900-1950]	28.83
(1950-2000]	64.55

Table 1: Distribution of Californian earthquakes 1769 – 2000 ($M_w \geq 4.5$).

The top panel of Figure 3 shows the time course of shock magnitude, while the bottom panel considers just shocks with magnitude equal to or greater than 6.3, the quantile of the order 95%. The second plot suggests some gaps between major events. Moreover, no important shocks are recorded before 1800.

Figure 4 illustrates the main features of the conditional distributions of magnitude in each half-century. Red lines correspond to events with a magnitude above the threshold of 6.3.

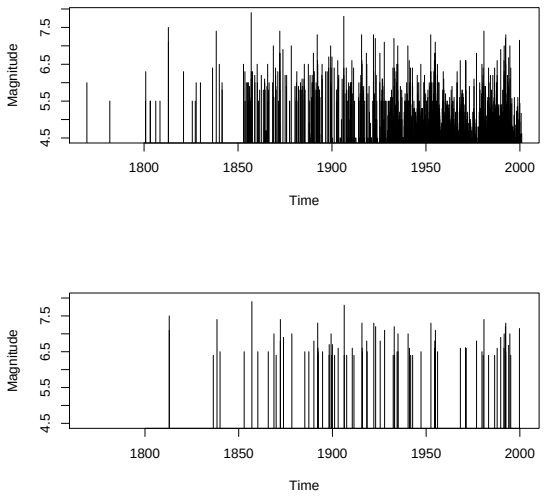


Figure 3: Californian catalogue. Time series of earthquake magnitude. Top: all events; bottom: events with magnitude greater than 95% quantile.

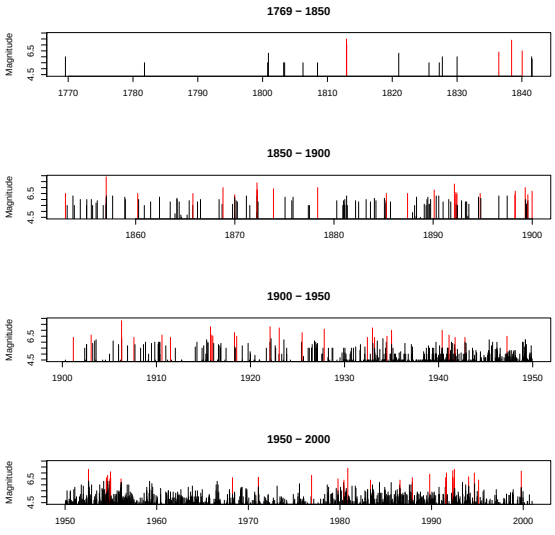


Figure 4: Californian catalogue. Time series of earthquake magnitude by time.

Figure 5 shows the overall distribution of shock magnitude and the conditional distributions in each half-century. The first two and the last two distributions (cor-

responding to periods 1769 – 1850, 1850 – 1900 and periods 1900 – 1950, 1950 – 2000, respectively) seem to have a similar behavior, hypothesis partially confirmed by the Kolmogorov-Smirnov test whose p-values are reported in Table 3. Once again, it is difficult to interpret these findings because of the confounding effect of technology improvement. Summary statistics of shock magnitude and results of Kolmogorov-Smirnov test of equality of time conditional distributions are reported in Tables 2 and 3, respectively.

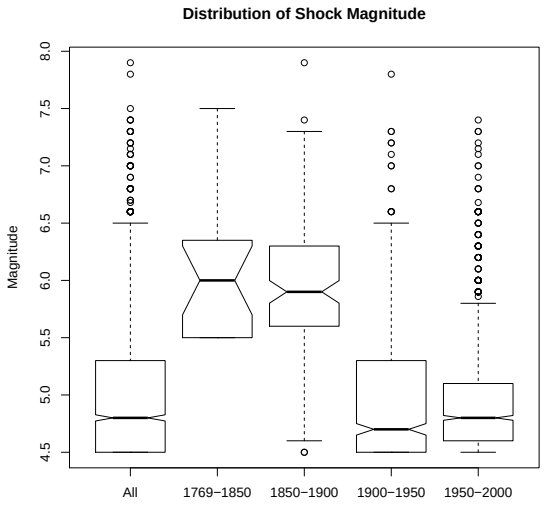


Figure 5: Californian catalogue. Boxplots of magnitude by time.

	Magnitude					
Time	Min	1.st Quartile	Median	Mean	3.rd Quartile	Max
[1769-1850]	5.5	5.5	6.0	6.1	6.3	7.5
(1850-1900]	4.5	5.6	5.9	5.9	6.3	7.9
(1900-1950]	4.5	4.5	4.7	5.0	5.3	7.8
(1950-2000]	4.5	4.6	4.8	4.9	5.1	7.4
[1769-2000]	4.5	4.5	4.8	5.0	5.3	7.9

Table 2: Californian catalogue. Summary statistics of earthquake magnitude by period of time.

Time				
	[1769-1850]	(1850-1900]	(1900-1950]	(1950-2000]
[1769-1850]		0.591	$1.668e-10$	$1.389e-13$
(1850-1900]			$2.2e-16$	$2.2e-16$
(1900-1950]				$2.2e-16$
(1950-2000]				

Table 3: Californian catalogue. P-value of Kolmogorov-Smirnov test of equality of time conditional distributions of magnitude.

In seismology waiting (or inter event) time is defined to be the time elapsed between consecutive earthquakes in a given region. Figure 6 represents the waiting-time distribution (in years) for all the earthquakes in the catalogue and for different periods of time (the ranges are the same as in previous analyses). Again, the decreasing trend of location statistics (i. e., medians) of time conditional distributional is surely influenced by technology improvement. Corresponding summary statistics are reported in Table 4. Together with events very close in time, there are also very long inter-event gaps, of 10 or more years, mostly in the period 1769 - 1850 when many earthquakes with low magnitude were not recorded. Table 5 reports p-values of Kolmogorov-Smirnov test of inter-event between period of time.

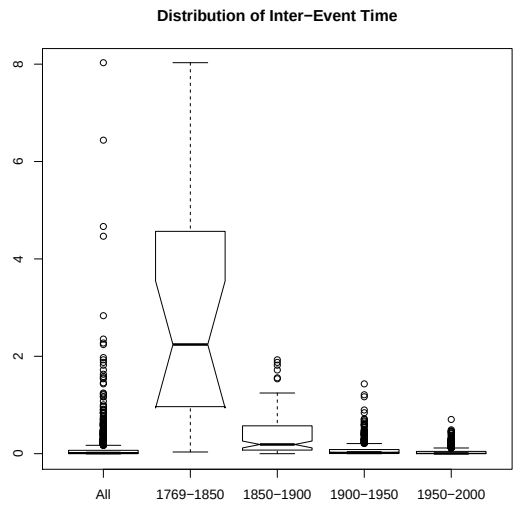


Figure 6: Californian catalogue. Boxplots of waiting time.

	Magnitude					
Time	Min	1.st Quartile	Median	Mean	3.rd Quartile	Max
[1769-1850]	0.035	0.966	2.242	3.789	4.566	19.000
(1850-1900]	0	0.074	0.189	0.377	0.572	1.928
(1900-1950]	0	0.003	0.027	0.078	0.085	1.434
(1950-2000]	0	0	0.006	0.036	0.047	0.703
[1769-2000]	0	0	0.014	0.105	0.069	19.000

Table 4: Californian catalogue. Summary statistics of waiting time of time- conditional distributions.

Time				
	[1769-1850]	(1850-1900]	(1900-1955]	(1955-2000]
[1769-1850]		$3.187e-07$	$9.545e-10$	$4.778e-11$
(1850-1900]			$2.2e-16$	$2.2e-16$
(1900-1950]				$3.952e-14$
(1950-2000]				

Table 5: Californian catalogue. P-value of Kolmogorov-Smirnov test of equality of waiting-time distributions.

Tables 6 and 7 report the conditional distributions of time given magnitude and the conditional distributions of magnitude given time, respectively. The first table confirms that most earthquakes are concentrated in the last century (1900 - 2000), while in the latter the proportion of major earthquakes decreases with time.

	Magnitude		
Time	[4.5-5.5] %	(5.5-6.5] %	(6.5-8.0] %
[1769-1850]	0.44(8)	2.97(9)	5.36(3)
(1850-1900]	1.46(27)	28.38(86)	23.21(13)
(1900-1950]	28.91(534)	29.05(88)	25.00(14)
(1950-2000]	69.19(1278)	39.60(120)	46.43(26)

Table 6: Californian catalogue. Conditional distribution of time by magnitude. Numbers in parentheses are absolute frequencies.

	Magnitude		
Time	[4.5-5.5] %	(5.5-6.5] %	(6.5-8.0] %
[1769-1850]	40.00(8)	45.00(9)	15.00(3)
(1850-1900]	21.43(27)	68.25(86)	10.32(13)
(1900-1950]	83.96(534)	13.84(88)	2.20(14)
(1950-2000]	89.75(1278)	8.42(120)	1.83(26)

Table 7: Californian catalogue. Conditional distribution of magnitude by time. Layout as in Table 6 .

Figure 7 offers an overall description of space-time distribution of earthquake epicenters occurred in California during the period covered by the catalogue. Most of the shocks occurred in the southern part of California. Moreover, many great earthquakes appear to be arranged along San Andreas fault. However, the northern part of California suggests a different behavior than the rest of the state, that is, a comparatively greater concentration of major shocks.

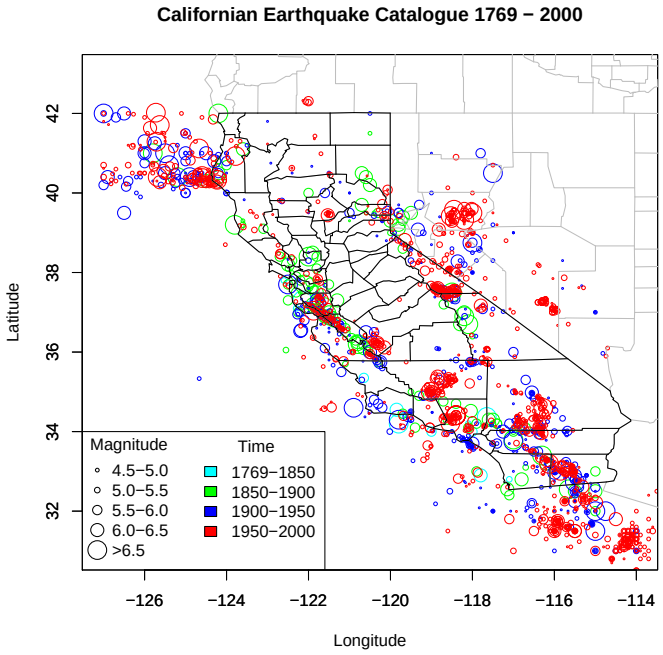


Figure 7: Californian catalogue. Bubble map of epicenters with bubble colors corresponding to time and bubble radii proportional to magnitude.

4 Descriptive analysis of Italian catalogue

A detailed account of the seismic Italian catalogue considered in this work was presented in Albarea (2012). The covered period is from 1600 to 2003 [2] and the total number of recorded events is 1469. The variables are the same as the Californian catalogue with the addition of a categorical variable describing the tectonic zone where each earthquake occurred. However, this information is not considered here. Like in the Californian case, only shocks with magnitude not less than 4.5 are included in the catalogue. Despite the fact that the period in question is much broader than California, the number of events is much lower. This could suggest that California has a greater seismic activity than Italy.

The histogram in Figure 8 shows that most Italian earthquakes are not so strong, indeed the mean and median magnitude are 5.023 and 4.830, respectively. During the reference period there were few earthquakes with great intensity and most of them occurred in the South of Italy. For example, magnitude quantile of the order 95% is equal to 5.85, i. e., just 5% of earthquakes have a magnitude higher than 5.85.

It is instructive to make a comparison between Italian and Californian catalogue. The first quartile of magnitude is 4.81 for Italy while for California it is 4.502. The discrepancy could be partially explained by the difference of devices used in recording earthquakes. Indeed instruments able to record every jolt, even those with very low intensity, started to be used in Italy much later than in California. The quantile of the order 95% is 5.85 in Italy and the corresponding statistic in the Californian catalogue is 6.3.

Another confirmation of what said above comes from Figure 9 that presents the plots of the cumulative distribution function of magnitude for both Californian and Italian catalogue. The decision to consider only earthquakes occurred in the period 1900 onwards is due to the fact that in this period of time there should be a more accurate registration of the seismic events, thanks to technology improvement. This subset will be used again below to compare the two catalogues. It could be noted that the two graphs in the figure are similar in that most of the observations for both catalogues are attributable to the last century. Italian cumulative distribution function is lower than its Californian analogue from the lower threshold 4.5 up to about 5 and then it is higher. This behavior is explainable with a lower dispersion of Italian magnitude than Californian magnitude.

[2] Data set kindly provided by Basili, R. and Rotondi, R.

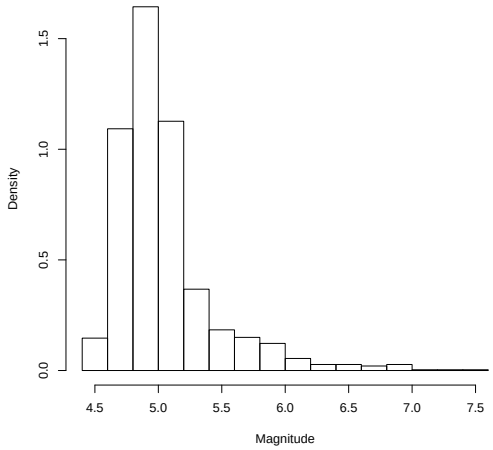


Figure 8: Italian catalogue. Magnitude distribution.

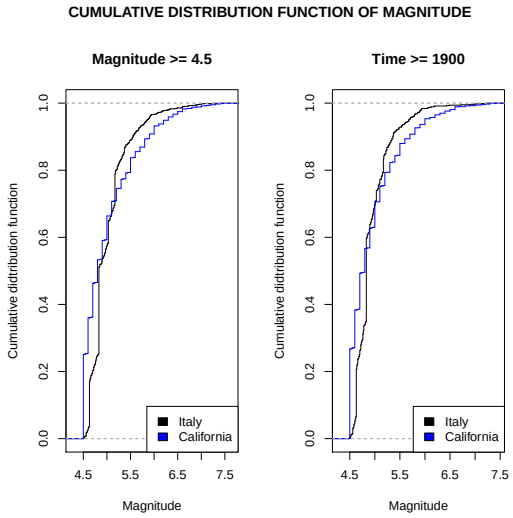


Figure 9: Cumulative distribution function of magnitude. Left: all events; right: events occurred in 1900 or later.

The time series of recorded shocks and their magnitude is shown in the top panel of Figure 11. The lower panel of Figure 11 shows earthquakes with magnitude greater than 6.3. This threshold was chosen because it is the 95-th percentile of the magnitude of Californian earthquakes and allows a comparison of the two catalogues for major events. Empirical evidence suggests California to have a much higher seismic activity

than Italy, in particular for great shocks. This finding is corroborated by the number of earthquakes with intensity greater than 6.3 recorded during the period 1850 - 2000.

Before that time, it seems that Italy has had a greater seismic activity than California, at least in terms of number of recorded events. However, this can merely depend on the fact that, differently from California, in Italy there is abundance of historical documents accounting for past events. A (public) data base of historical earthquakes starting from 217 B. C. is available at http://emidius.mi.ingv.it/DBMI04 (Boschi et al, 1997) [3] This confirms that a safer comparison of California and Italy is only possible after 1900, when devices for precise recording of earthquakes began to be widely available.

Figure 10: Inscription related to earthquake occurred in 442 A.D. found in the Colosseum.

[3] An example of such documents is the following quotation from Paulus Diaconus, *Historia Romana, Liber XIII, 16.* "Sub his fere diebus (442 A. D.) tam terribili terraemotu Roma concussa est ut plurimae aedes eius et aedificia corruerint". This event is confirmed by an inscription found in the Colosseum (see Figure 10) to celebrate the restoration of the monument by Flavius Synesius Gennadius Paulus and Rufius Caecina Felix Lampadius (CIL VI, 32089).

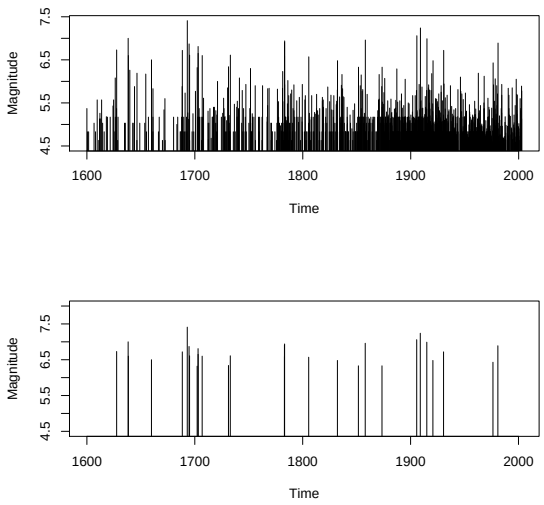


Figure 11: Italian catalogue. Time series of earthquake magnitude. Top: all events; bottom: events with magnitude greater than 6.3.

Figure 12 shows the time series Italian earthquakes, by century. Events with a magnitude greater than, or equal to, the threshold of 6.3 are highlighted. Results confirm what said above, i. e., the seismic activity of Italy appears to produce a minor number of devastating shocks with respect to California.

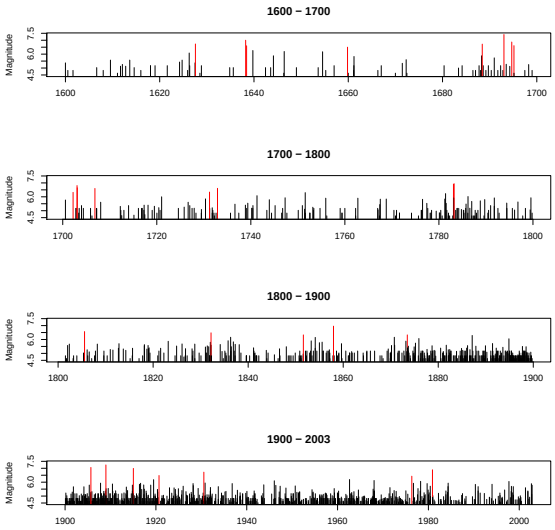


Figure 12: Italian catalogue. Time series of earthquake magnitude by century.

Figure 13 compares the seismic activity of the last century in Italy and California. During this period in the California occurred 2060 earthquakes of which 62 events with a magnitude greater than 6.3, while in Italy occurred 812 shocks with just 7 earthquakes over the magnitude threshold of 6.3.

Figure 14 describes the distribution of shock magnitude by century and for the entire period. As already noted, the *decreasing* trend of location statistics, i. e., median magnitude, with respect to time must be considered an artifact of technology improvements allowing a more accurate recording of small shocks.

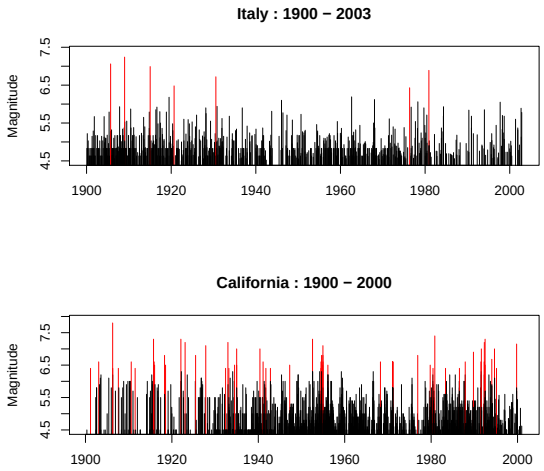


Figure 13: Time series of earthquake magnitude in the last century: California vs Italy.

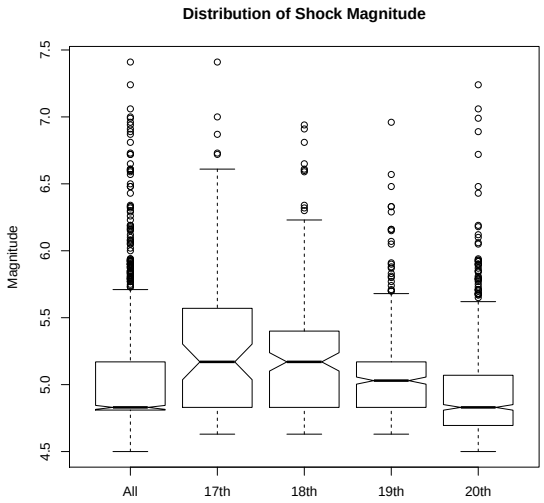


Figure 14: Italian catalogue. Boxplots of magnitude by time.

Time distribution of events in Italian catalogue is shown in Figure 15 (see also Table 8). The overall shape is similar to Californian catalogue (see Figure 2), with a peak in the density of events corresponding to the fist half of 20th century. Once again,

this finding must be considered with caution because of the underestimation of earlier seismic activity caused by the lack of technology.

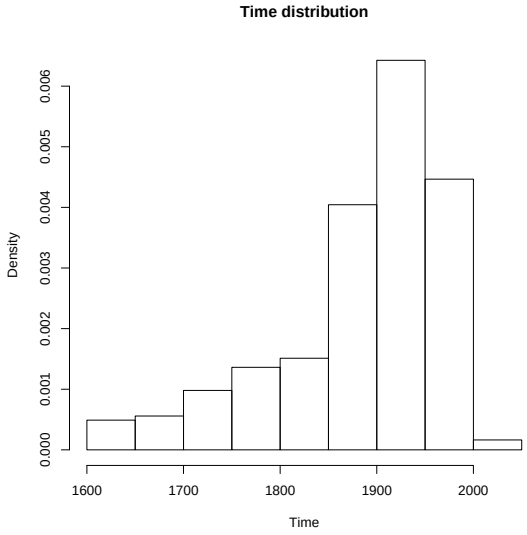


Figure 15: Italian catalogue. Histogram of shock times.

Time	%
[1600-1700]	5.24
(1700-1800]	11.71
(1800-1900]	27.77
(1900-2003]	55.28

Table 8: Distribution of Italian earthquakes 1600 – 2003 ($M_w \geq 4.5$) by century.

Figure 16 reports the distribution of inter-event times, by century and for the entire period. A strong decreasing trend of all location statistics is evident, caused by technology improvements in the recording of minor shocks. In the last century, when presumably almost all earthquakes are recorded, the time elapsed between two consecutive shocks becomes very small; the median waiting time is 0.083, about 30 days. In the 17th and 18th centuries the inter-event time is significantly higher because the catalogue includes major earthquakes only. A similar behavior is apparent from Figure 6 for the Californian catalogue. However, in the earlier times, when no specific technology was available, Italy inter-event time appears to be *lower* than California, whereas in the last century the relation is reversed, e. g., the median inter-event time in California is 0.006, about 2 days. This is because San Andreas fault has a seismic activity much more continuous with respect to Italian faults.

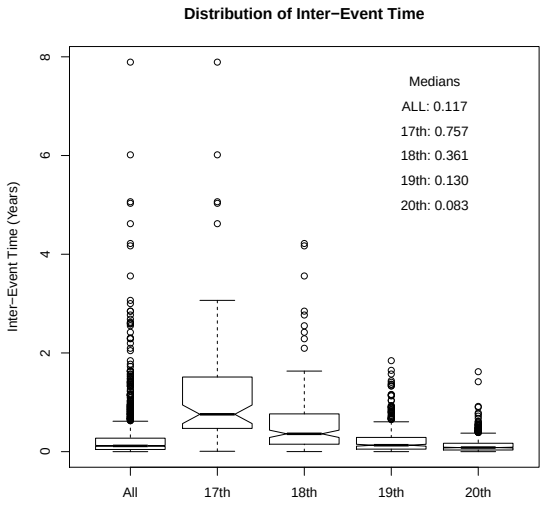


Figure 16: Italian catalogue. Boxplots of waiting time.

Figure 17 compares the distribution of inter-event in the last century for the reference catalogues. The plot suggests the waiting time between consecutive earthquakes in California to be lower than Italy. This result can be explained by Figure 13, where the number of shocks in California is clearly higher than in Italy implying the inter-event time to be lower.

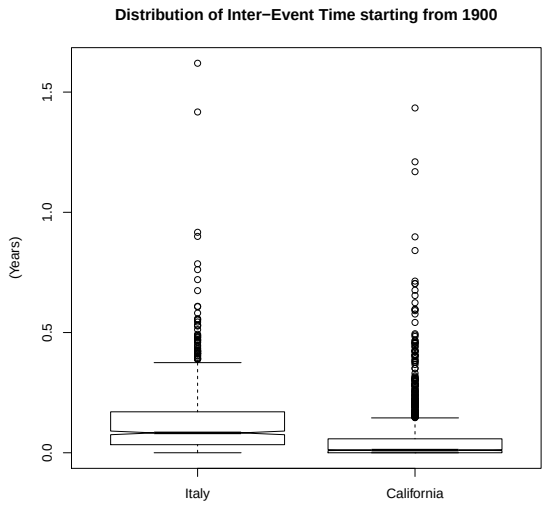


Figure 17: Inter-event time in the last century: California vs Italy.

Figure 18 gives an overall picture of Italian catalogue, with a representation of epicenter locations and the magnitudes and times of the corresponding shocks. Clearly, even though areas with a greater concentration of phenomena are apparent, earthquakes occur all around Italy, with the notable exception of Sardinia. The comparison with Californian map in Figure 7 suggests Italy to have a high seismic activity producing mainly events with lower magnitudes than California.

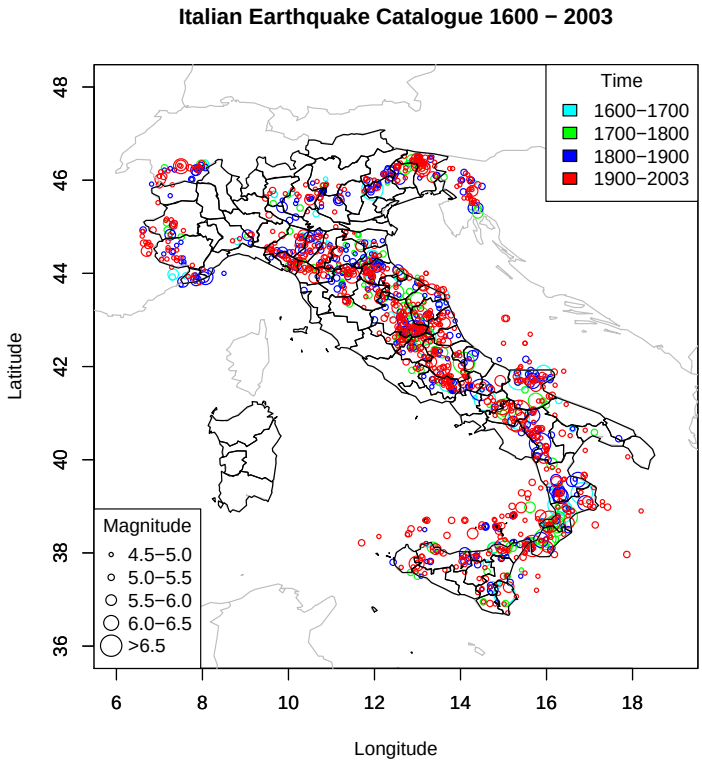


Figure 18: Italian catalogue. Bubble map of epicenters with bubble colors corresponding to time and bubble radii proportional to magnitude.

5 Statistical Methods

In the previous sections we analyzed the main features of the catalogues with simple descriptive analysis. Now we focus the attention on the spatial distribution of epicenters and describe two competing statistical methods able to act as *(spatial) smoothing techniques*. The first is *data depth* and the second is *significant feature extraction*. Their rationale is very different because data depth essentially provides a ranking in centrality of multivariate data with respect to a distribution or a data set, while the extraction of significant features is basically an application of kernel density estimation. The concept of depth is relatively new and the related techniques and applications are still under development. On the contrary, kernel density estimation is a well-established statistical method witnessed by an extensive literature produced especially in the last thirty years. Both methods are described below in some detail so as to help the reader to follow the discussion of the analysis of Italian and Californian catalogues in Section 6.

5.1 Data Depth

Data Depth (DD) is a statistical method to rank the points of a space according to centrality with respect to an assumed probability distribution or a data set. Usually, reference space is Euclidean space $\mathbb{R}^p$, i.e., Data Depth is multivariate in nature, univariate case is obtained as a particular case. Through the application of a ranking function (depth function) each point of the space receives a non negative score describing its degree of centrality (depth value). Depth implicitly defines a center as the point with maximum depth value, and decreases along any direction from the center.

Definition 1 (Depth Function)**.** A depth function $d(.)$ measures the centrality of a point with respect to a probability distribution F or a data set. It is a function from the Euclidean space $\mathbb{R}^p$ into the set of non negative real numbers, that is $d : x \in \mathbb{R}^p \longrightarrow d(x;F) \geq 0$.

A depth function should satisfy the following properties:

1. affine invariance: $d(x;F) = d(Ax+b;F_{A,b})$;
2. monotonicity relative to deepest point: depth decreases along the rays from the center;
3. F centro-symmetric about $c \Longrightarrow d(x;F) \leq d(c;F)$ for all x;
4. if $\| x \| \to \infty \Longrightarrow d(x;F) \to 0$;

Commonly used depth functions include Mahalanobis' and other depth functions (Liu et al, 1999) also with extensions to directional and functional data. Geometrical depth functions are nonparametric and use simple geometrical structures (e.g., simplices, halfspaces) to capture information about reference distribution or data set. The main difference with respect to Mahalanobis' depth is that they are data adaptive. In fact, the contours of simplicial and halfspace depth tend to follow the structure of data, while the contours of Mahalanobis' depth are always ellipsoids, independently of the structure of data. A possible trouble for halfspace and simplicial depth can occur in high dimension because of information sparsity.
Two examples of depth-based functionals are:

- location functional: maximizer of the depth function, that is the deepest point of the space with respect to the distribution;
- dispersion functional: Lebesgue integral of the depth function.

Depth-based functionals are devised in the general multivariate situation and they have the advantage of simple geometrical interpretations.

Definition 2 (Depth-based location)**.** For a depth function $d(.)$, the location parameter (or multivariate median) of the distribution is

$$\theta(F) = argmax_x d(x;F)$$

and the corresponding sample statistic is

$$\theta(\hat{F_n}) = argmax_x d(x;\hat{F_n}).$$

Depth-based medians are important because they provide nonparametric and remarkably robust multivariate estimators of location. A general dispersion parameter is the (Lebesgue) integral of $d(x;F)$

$$\gamma_F = \int_{R^p} d(x;F)dx. \tag{1}$$

Here we concentrate on simplicial depth $d_S(\cdot;F)$, defined to be the probability coverage of random simplices. A random simplex $S_{p+1}^{(F)} \equiv S_{p+1}^{(F)}(X_1,\cdots,X_{p+1})$ is the convex hull of $p+1$ random observations $X_i, i=1,\cdots,p+1$ from the probability distribution F.

Definition 3 (Simplicial depth)**.** Let F be a probability distribution and let $\mathscr{S}_{p+1}^{(F)}$ be the class of random simplices of $\mathbb{R}^p$ from F. For $x \in \mathbb{R}^p$

$$d_S(x;F) = P_F(S_{p+1}^{(F)} \in \mathscr{S}_{p+1}^{(F)} : x \in S_{p+1}^{(F)}). \tag{2}$$

At the beginnings of data depth it was almost a postulate that depth ranks could single out just one center of a distribution, corresponding to the maximizer of the ranks, whatever the shape of the distribution, unimodal or multimodal. A new concept of depth, called local depth, shows that some generalized depth functions can indeed account for multimodal data, having multiple centers. Local depth measures centrality conditional on a bounded neighbourhood of each point of the space. These generalizations are called local depth functions because they just consider the behavior of the probability distribution in a nearby region of the point under consideration, instead of the entire space.

A local version $ld_S(\cdot;F,\tau)$ of simplicial depth is obtained by constraining the size of simplices not to exceed a given size $\tau > 0$ (Agostinelli and Romanazzi, 2011). Suitable measures $t(S_{p+1}^{(F)})$ of size are diameter or volume.

Definition 4 (Local simplicial depth)**.** Let the notation be as in Definition 3. For a given $\tau > 0$,

$$ld_S(x;F,\tau) = P_F(S_{p+1}^{(F)} \in \mathscr{S}_{p+1}^{(F)} : x \in S_{p+1}^{(F)}) \cap t(S_{p+1}^{(F)}) \leq \tau). \tag{3}$$

Remark 5. In the univariate case, it is easily shown (Agostinelli and Romanazzi, 2011) that, when $\tau \to \infty$, $ld_S(x;F,\tau)$ converges to $d_S(x;F)$, $x \in \mathbb{R}^p$. Hence, in this case, local depth can be considered a family of (generalized) depth functions, indexed by τ, and the family includes global depth.

The τ parameter dictates the width of the neighbourhood around each point of the space. It is somewhat similar to the bandwidth, or window size, in kernel density estimation (Rosenblatt, 1956). Since it is constant in the whole of the space, local depth ranks (like depth rank) can be used to order data points according to centrality.

5.2 Kernel Density Estimation

During the last 30 years, density estimation aroused a wide interest in Statistics (Sheater, 2004). Important contributions are (Silverman, 1986), (Bowman and Azzalini, 1997), (Simonoff, 1996) and (Wand and Jones, 1995). The present account considers only density estimation based on kernel method.

Kernel density estimation (KDE) is an important method of data smoothing allowing inferences to be drawn from a finite sample about the parent population. It is intended to produce a nonparametric estimate of the (unknown) probability density function of a random variable or a random vector.

Definition 6 (Kernel density estimator)**.** Suppose to have a random sample of p-variate data $X_1, ..., X_n$ with an unknown continuous distribution with probability density function (pdf) $f(x)$ and cumulative distribution function (cdf) $F(x)$. The kernel density estimator is represented as follows:

$$\hat{f}(x) = \frac{1}{nh^p} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \,, \tag{4}$$

where K is the kernel function and h denotes the bandwidth parameter.

The main assumptions to satisfy for the kernel are the following:

- $\int K(x)dx = 1$,
- $\int xK(x)dx = 0$, which implies $K(-x) = K(x)$,
- $\int x^2 K(x)dx = \sigma_K^2 < \infty$.

There are several examples of kernel functions, e. g., Epanechnikov, Biweight, Gaussian, but in the following only the Gaussian kernel is considered. Its expression is:

$$K_{Gau}(x) \equiv K(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2}), \tag{5}$$

i. e., it is coincident with the standard normal density function.

A major problem in KDE is the choice of the bandwidth h. It is clear that a large bandwidth tends to oversmooth the population density with the risk of masking real structures, while a small bandwidth could show many artifacts depending on sample noise with no relation to population features. Hence, finding the value of h that minimizes the error between the estimated density and the true density is a more influential aspect than choosing the kernel function.

The mean squared error (MSE) is the standard measure of accuracy of the density estimator. We note that the MSE of $\hat{f}(x)$ is a function of the argument x:

$$\begin{aligned} MSE(\hat{f}(x)) &= E(\hat{f}(x) - f(x))^2 \\ &= Bias(\hat{f}(x))^2 + Var(\hat{f}(x)). \end{aligned}$$

A measure of the global accuracy of $\hat{f}(x)$ is the mean integrated squared error (MISE):

$$
\begin{aligned}
MISE(\hat{f}(x)) &= E\int_{-\infty}^{\infty}(\hat{f}(x)-f(x))^2dx \\
&= \int_{-\infty}^{\infty} Bias^2(\hat{f}(x))dx + \int_{-\infty}^{\infty} Var(\hat{f}(x))dx.
\end{aligned}
$$

The recent literature on KDE has discussed at length the problem of the optimal bandwidth choice. More details on (multivariate) bandwidth selection problem are given by Duong and Hazelton (2003) and Duong and Hazelton (2005) and references therein.

The present implementation of KDE is based on the R package *feature*, available on CRAN at `http://cran.r-project.org`. Feature significance is a technique for deciding whether some specific features - such as local extrema - are statistically significant. The methodology implemented in the program (Duong et al, 2008) is essentially a framework for feature exploration in multidimensional data which combines (differentiation of) KDE and hypothesis tests for modal regions. With one- and two-dimensional data, reported features are local extrema, ridges, valleys and steep gradients. For three- and higher-dimensional data, features are essentially captured by significant modal regions or local maxima. The R package *feature* is illustrated by Chaudhuri and Marron (1999) for 1-dimensional data, Godtliebsen et al (2002) for 2-dimensional data and Duong et al (2008) for 3- and 4-dimensional data. The program gives several options to display and compute kernel density estimates, statistically significant gradients and curvature regions. Significant gradient and/or curvature regions often correspond to important properties of the distribution.

6 Results

6.1 Spatial Analysis

Empirical simplicial depth is defined to be the proportion of simplices including a given point x, with the shape of simplices depending on the dimension. For example, when studying the spatial distribution of earthquakes, simplices are triangles with vertices corresponding to epicenters. For the Italian catalogue there are $\binom{1469}{3}$ such triangles and $\binom{2206}{3}$ for the Californian catalogue. In the present application, empirical simplicial depth is the proportion of sample simplices determined by earthquake epicenters including a given location. This value can be interpreted as a measures of seismic risk. In fact, the simplicial depth value of a station under consideration is exactly the proportion of times it belongs to triangles formed by the epicenters of earthquakes. Of course, it is more reasonable to consider local depth functions instead of global ones. For the sake of comparability (local and global) depth values are normalized as follows:

$$
d^* = (d - d_{min})/(d_{max} - d_{min}). \tag{6}
$$

The choice of τ parameter is very influential for local depth because it measures the size of the neighbourhoods in the analysis. Consider again the spatial analysis. If we use the diameter to measure simplex size, τ is the maximum distance (degrees) between

the epicenters corresponding to simplex vertices. If we use the volume, τ is the area of the simplex. Both diameter and volume present advantages and disadvantages. In particular:

- Diameter. The advantage of diameter is that, for any given simplex and a given value of τ, a circular neighbourhood with radius $\rho(\tau)$, depending on τ, can always be found which contains the simplex. Therefore, any point enclosed by the simplex has a maximum distance $\rho(\tau)$ from the vertices of the simplex. This adds a nice geometrical interpretation to local depth. In general, the empirical distribution of simplex diameters shows a wider range than simplex volumes. A good choice of *tau* as a particular quantile of sample simplex diameters requires care. A value too high would produce too wide neighbourhoods and a value too low would emphasize chance effects masking the signal. A side effect is a greater computational effort.

- Volume. As for diameter, an appropriate value is a low-order percentile of the empirical distribution of simplex volumes. In the bivariate case, comparability with diameter can be restored by taking the square root of the corresponding value of the percentile chosen. A possible problem with volume arises from almost degenerate simplices, i. e. , simplices with almost null volume and positive, possibly high, diameter. Sometimes these simplices are at the origin of depth artifacts, that is, regions of the space with artificially high depth values, not corresponding to density peaks.

6.2 Spatial distribution of epicenters of Californian catalogue

We study the spatial distribution of epicenters for Californian catalogue using (local) depth values of a grid built on the square with sides (latitude x longitude, degrees) $(32,43)) \times (-127,-114)$. The grid step is 0.1 degrees for both latitude and longitude for a total of 15851 points.

Figure 19 shows the results of depth analysis (local simplicial depth, diameter version, $\tau = 2.57$ degrees, corresponding to 5% quantile of simplex diameters) and bivariate normal kernel smoothing with a bandwidth of 0.84 degrees for both coordinates. Note that depth τ and kernel bandwidth are different and this can be considered normal because the scales of representation are depth in the first case, density in the second. We are mainly interested in finding areas with higher depth/density.

In the top panel of Figure 19 Californian map is colored with different levels of gray, darker shades corresponding to higher depth values or risk. In the bottom panel we use a scale from white (low value) to red (high value) to display density values; statistically significant regions are delimited by blue lines. The results appear remarkably similar. Both methods locate the area of maximum risk in southern California, near the border with Mexico and in the counties of San Bernardino and Los Angeles. Central California has two risky areas, the first near the ocean (San Francisco zone), the second near Nevada border, but the risk level is markedly lower than in southern California of the county. Finally, in northern California both methods detected a risky area near the coast and off-shore, where San Andreas fault ends. The emerging pattern is a concentration of epicenters mainly around San Andreas fault and concentration

decreases from south to north. These findings are in agreement with the results of exploratory analysis displayed in Figure 7. It will be interesting a comparison with the full analysis, considering both locations of epicenters and magnitudes.

Figure 20 is similar to Figure 19, except that the area of simplices is used instead of the diameter. Here a bivariate normal kernel with a bandwidth of 0.30 degrees for both coordinates is used and for depth analysis τ is also equal to 0.30 degrees (squared root of 5% quantile of simplex areas). However, recall that the interpretation of the bandwidth is somewhat different in the two situations. The top panel of Figure 20 suggests volume to produce a smoother risk map than diameter, closely tracing San Andreas fault track. Again, darkest zone, i. e., highest risk, is located in the south of California. Another two risky areas are found near San Francisco and in the north. However, overall the map of risk given by diameter appears more realistic, because some structures found by volume version could be artifacts with no correspondence to reality. It is also apparent that in this case too bivariate kernel and volume version of depth produce similar results.

6.3 Spatial distribution of epicenters of Italian catalogue

To study the spatial distribution of epicenters in Italy the same criteria as in the previous section are used. A grid is built on the square enclosing Italy with sides between latitudes 36 and 48 and longitudes 6 and 19, with a total number of points equal to Californian grid (15851).

Figure 21 presents in the top panel the results of depth analysis, diameter version, and in the bottom panel the results of bivariate normal kernel smoothing. Kernel bandwidth for both latitude and longitude of 0.84 degrees. In depth analysis the value of τ is again the 5% quantile of simplex diameters. Corresponding to this choice, it turns out that the maximum distance between epicenters corresponding to simplex vertices is 2.20 degrees. According to depth the more risky zone is located in the center, precisely in the tectonic area called Central Northern Apennines West. Also, Central Northern Apennines Alps is a second region with a remarkably high value of depth. If we compare depth with kernel, we can conclude that the more risky areas are the same for both methods. The only difference is that depth does not detect a remarkable risk in the northwest of the country. Moreover, kernel smoothing computes risky also the Adriatic sea, even though at a lower level. This result could just be an artifact because that zone is not famous for its earthquake history.

The top panel of Figure 22 displays the Italian map of seismic risk given by depth, volume version (τ is the squared root of 5% quantile of simplex areas, 0.33 degrees), and the bottom panel displays the map given by kernel smoothing (bandwidth = 0.33 degrees for both coordinates). Volume and diameter versions have a similar behavior excepth that volume tends to create artifacts in the form of thin segments connecting seismic zones far apart. These artifacts are visible in the north of Adriatic sea and in the south of Tyrrhenian sea. However, both versions locate the main risky zone in the center of Italy. Again as in Figure 21, depth analysis and kernel smoothing show a similar trend with the exception of northwest of Italy, where the former does not seem to capture the information, instead, given by bivariate kernel.

6.4 Joint analysis of geographical coordinates and magnitude

In the present section we consider the joint distribution of epicenter locations (described by their geographical coordinates) and magnitudes to obtain a more complete analysis of seismic risk.

Depth analysis uses a grid which is the cartesian product of the geographical grid of the previous section and a number of intervals of magnitude range. Results are shown as in the spatial analysis, with the only difference that, as the space dimension is 3, it is necessary to use one map for each magnitude interval. As a rule, gray scale is conditional on each magnitude class, implying that the same color (gray) level in different plots may not correspond to the same depth value. Therefore, for the sake of clarity, a legend is added to each plot reporting the value corresponding to each gray level. Here we concentrate on results obtained from diameter version. Results from volume analysis are reported in the Appendix.

On the other hand, kernel analysis is performed *conditional* on each magnitude class and then it is really a sequence of separate bivariate analyses. In any case, we are

able to compare the results produced by the two methods.

6.5 Californian catalogue

The top panels of Figures 27, 23, 28 and 24 display the depth-based (diameter version) risk map of California for different values of magnitude. For the first three classes of magnitude the results remains fairly stable: the south of the country appears to be more at risk, with an extension to the center of California. For the last magnitude class (Figure 24), previous results are confirmed with the addition of a coastal and off-shore area in northern California near Cape Mendocino, where in fact strong earthquakes occurred in the past. From the gray bar, it can be noted that depth value, i. e., seismic risk, is much lower than previous magnitude classes.

The bottom panels of Figures 27, 23, 28 and 24 display the results of kernel analysis for the different magnitude classes. The bandwidth is the same as in the spatial analysis. For lower magnitude values, southern California is confirmed to be the more risky area, but differently from depth analysis some risk is also reported in northern California. For strong earthquakes the differences appear even more important. Indeed in kernel analysis some coastal areas in the center and in the north, near Cape Mendocino, have a density value higher than in the south of California.

6.6 Italian catalogue

As discussed in Albarea (2012) (note, however, that the 3 – dimensional grid is different), lower magnitude events are mainly concentrated in northern and Central Italy, whereas when magnitude increases the more risky areas are concentrated in the south of the country. This conclusion is now confirmed by the maps reported in the top panels of Figures 33, 25, 34 and 26. The bottom panels of the same figures illustrate the results of kernel analysis for different magnitude classes. The same bandwidth is used as in the spatial analysis. Overall, kernel and depth produce similar results. As noted in spatial analysis, the main difference is in the northwest of the country, where kernel detects a risky zone in the lower magnitude class.

7 Final remarks

The first part of the paper is devoted to a comparison of the main features of Californian and Italian catalogues. The comparison can be divided into two periods of time, before the advent of advanced technological devices able to (automatically) record the magnitude of an earthquake and after. In the first period earthquakes of great intensity that caused damage to structures or people are mainly reported. This information is likely to be recorded by historical documents and Italy with respect to California has several documents relating to past events. The use of technology to measure seismic activity has been introduced in California a couple of decades before Italy and this partly explains why the seismic activity in California appears much more intense than Italy not only in terms of strong earthquakes but also in terms of number of events.

In the second part of the paper we compare statistical depth (local version using simplex diameter as bandwidth) with kernel density estimation as data smoothers. We start considering just the spatial distribution of epicenters for both catalogues and then we add the magnitude. The comparison between depth and kernel suggests an overall similarity with the exception of spatial-magnitude analysis of California. For great earthquakes kernel density estimation detects the more risky zone in the Pacific ocean near Cape Mendocino, in the north of California, while for data depth it is located in the south of the country, particularly around San Bernardino and near Mexico border.

Appendix A: California

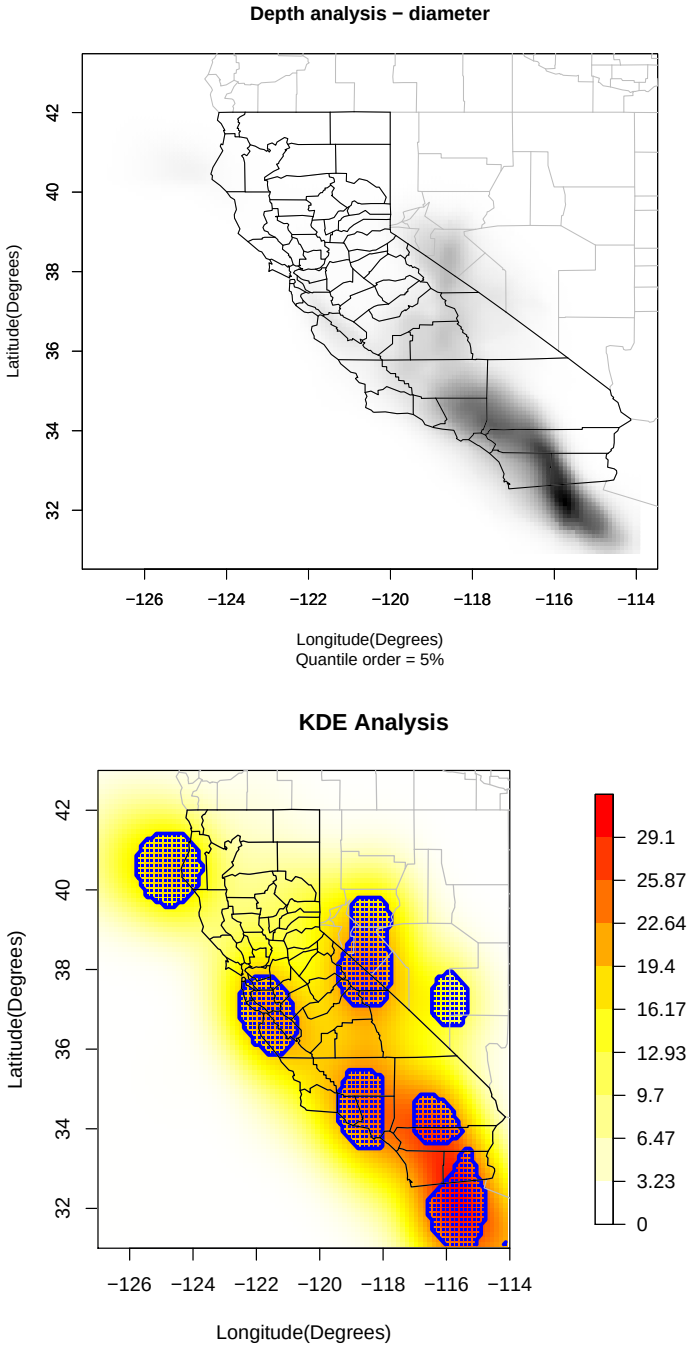


Figure 19: Californian map of seismic risk. Top: depth analysis. Bottom: kernel smoothing, bandwidth: 0.84 (density values are multiplied by 1000).

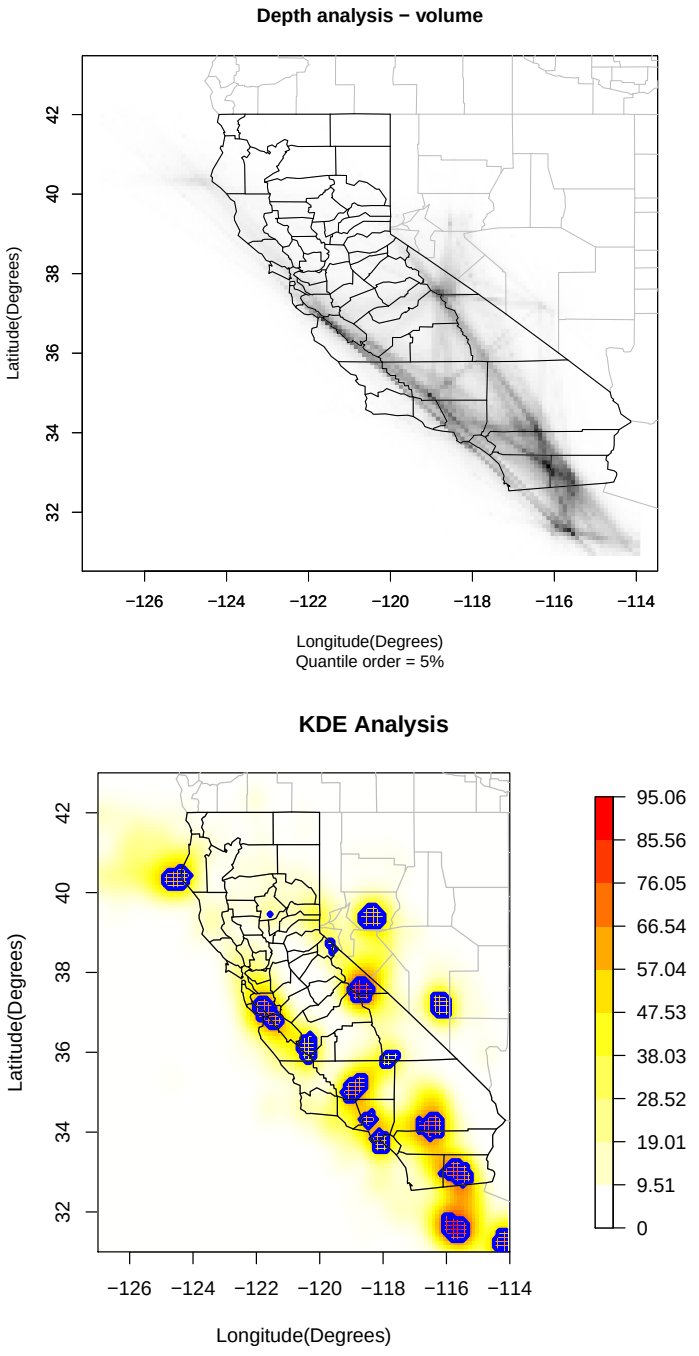


Figure 20: Californian map of seismic risk. Top: depth analysis. Bottom: kernel smoothing, bandwidth: 0.30 (density values are multiplied by 1000).

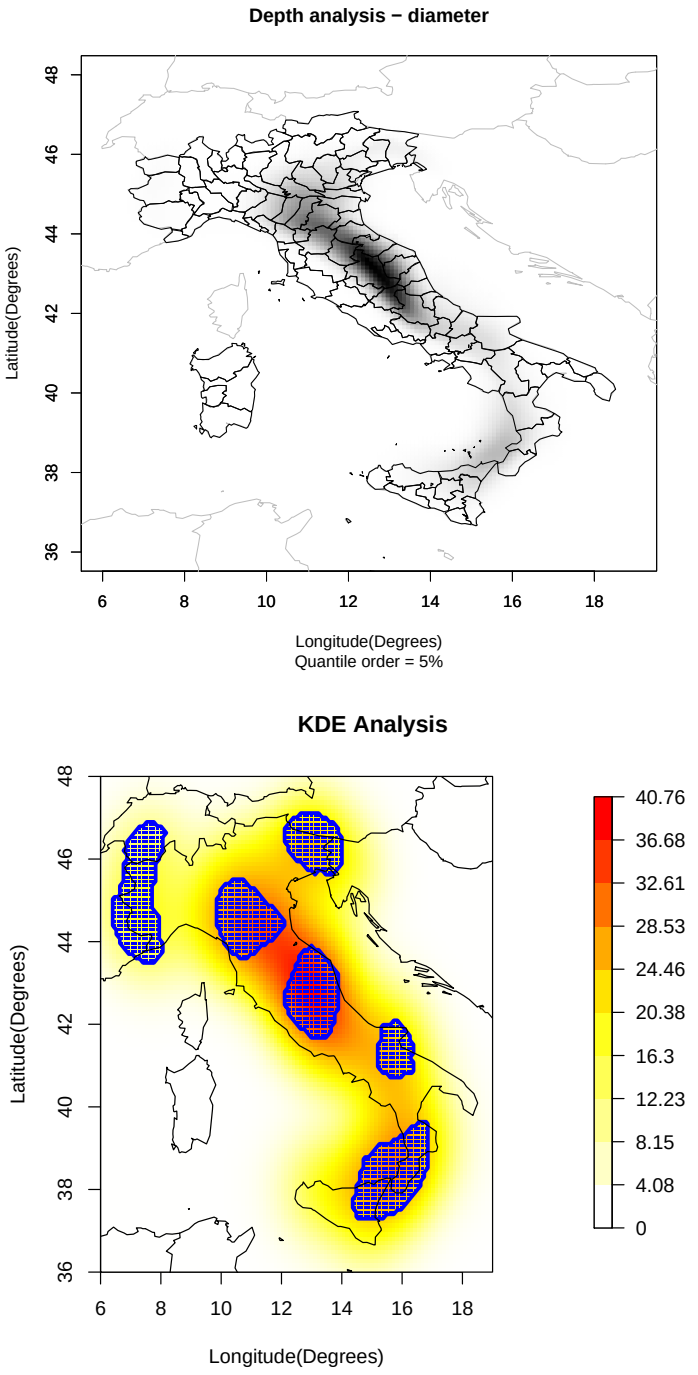


Figure 21: Italian map of seismic risk. Top: depth analysis. Bottom: kernel smoothing, bandwidth: 0.84 (density values are multiplied by 1000).

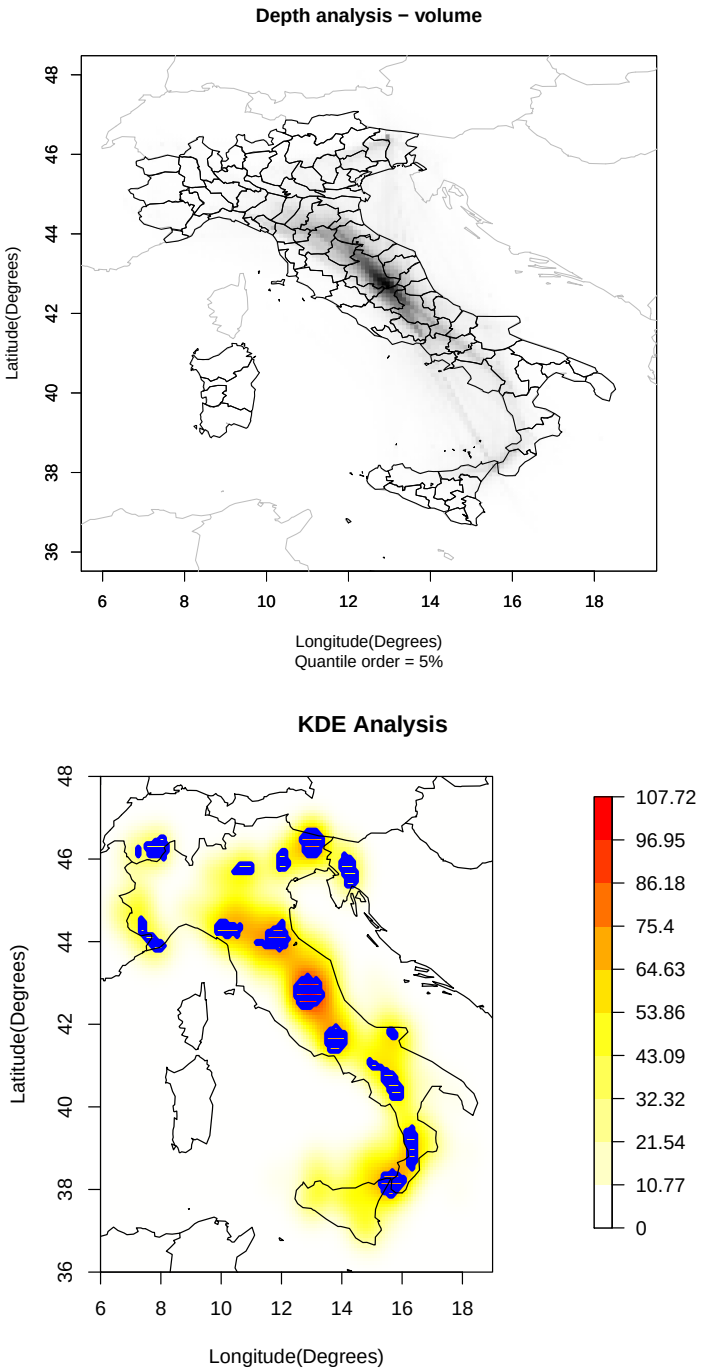
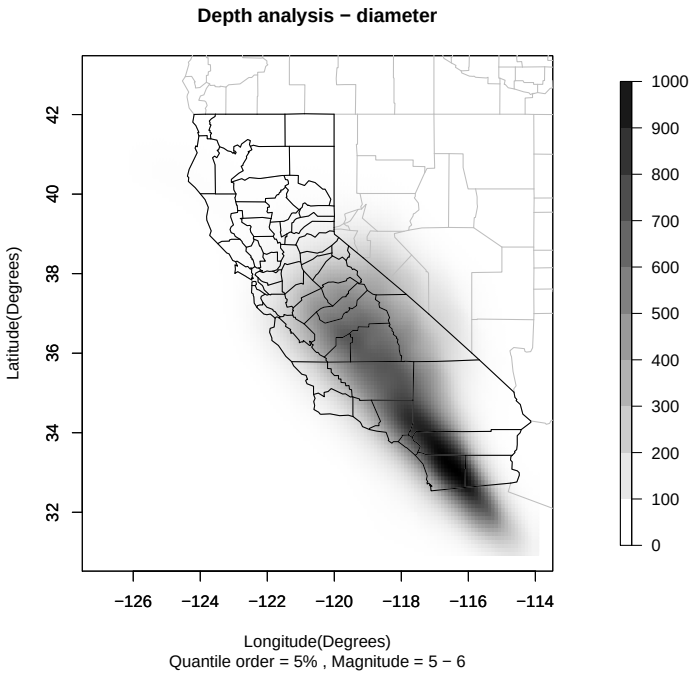


Figure 22: Italian map of seismic risk. Top: depth analysis. Bottom: kernel smoothing, bandwidth: 0.33 (density values are multiplied by 1000).



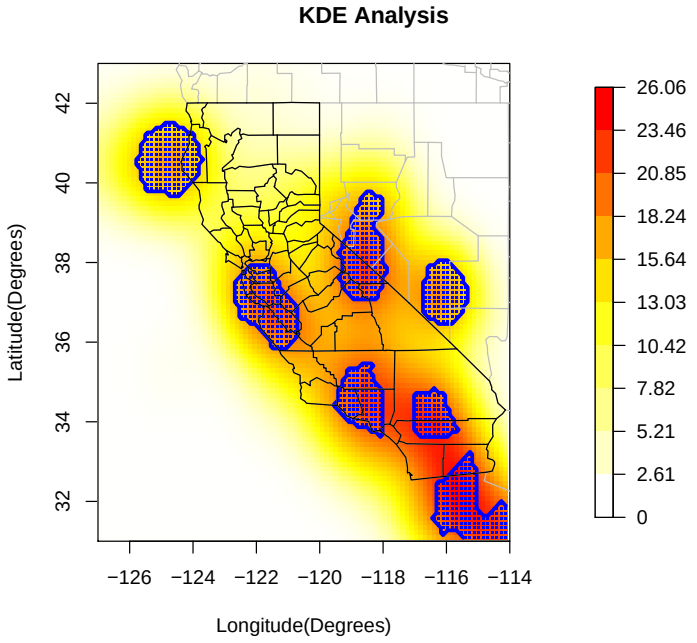


Figure 23: Californian map of seismic risk; magnitude: 5 – 6. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.84 (density values are multiplied by 1000).

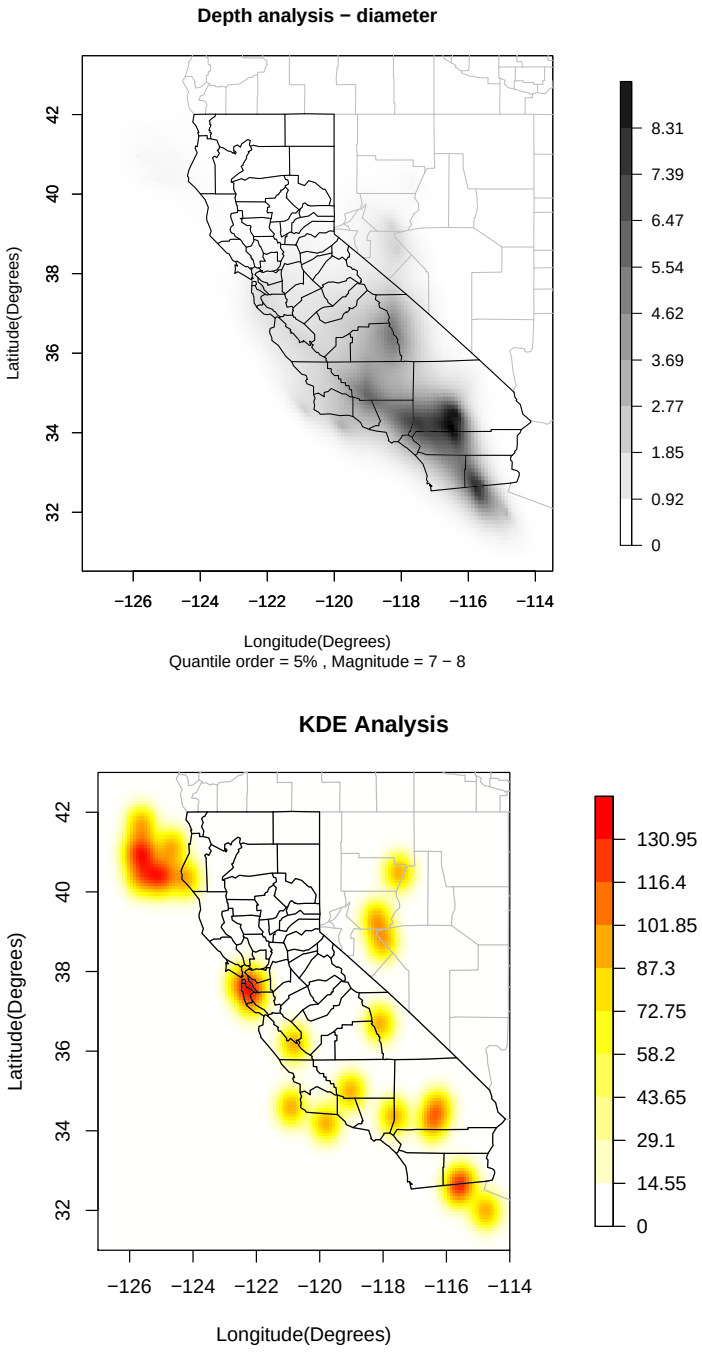
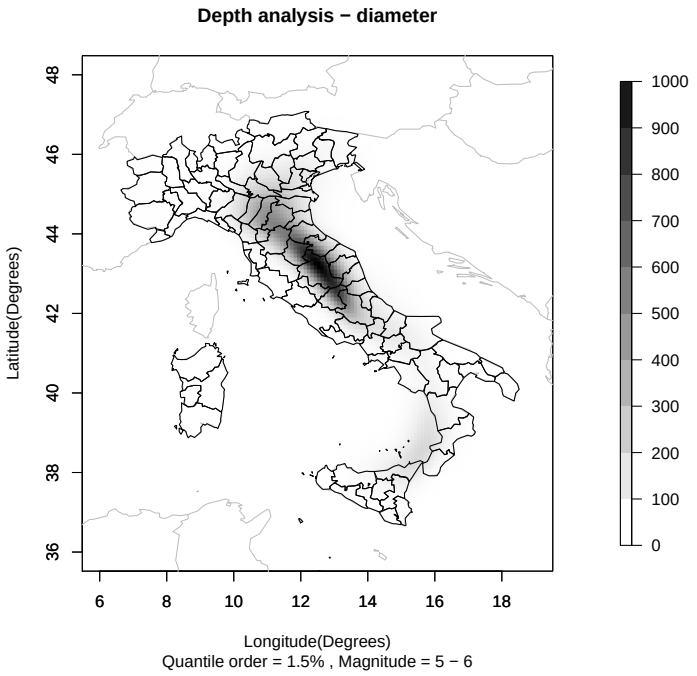


Figure 24: Californian map of seismic risk; magnitude: 7 – 8. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.84 (density values are multiplied by 1000).



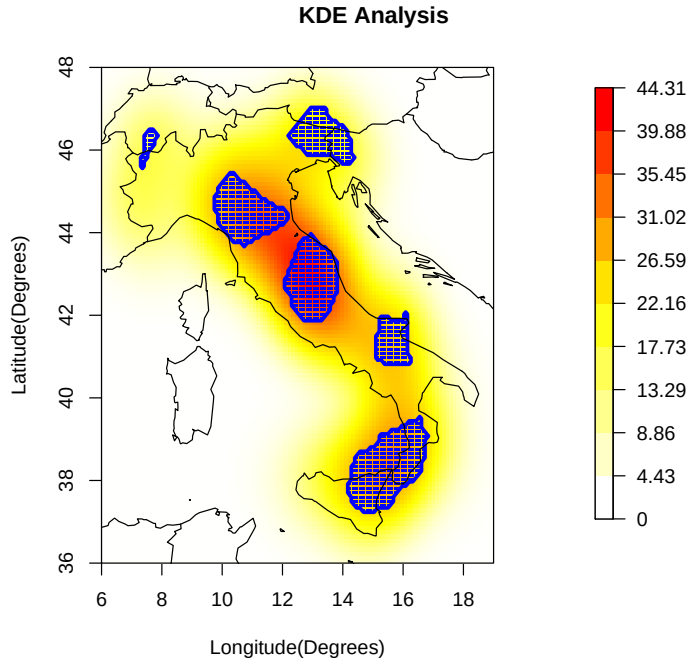


Figure 25: Italian map of seismic risk; magnitude: 5 – 6. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.84 (density values are multiplied by 1000).

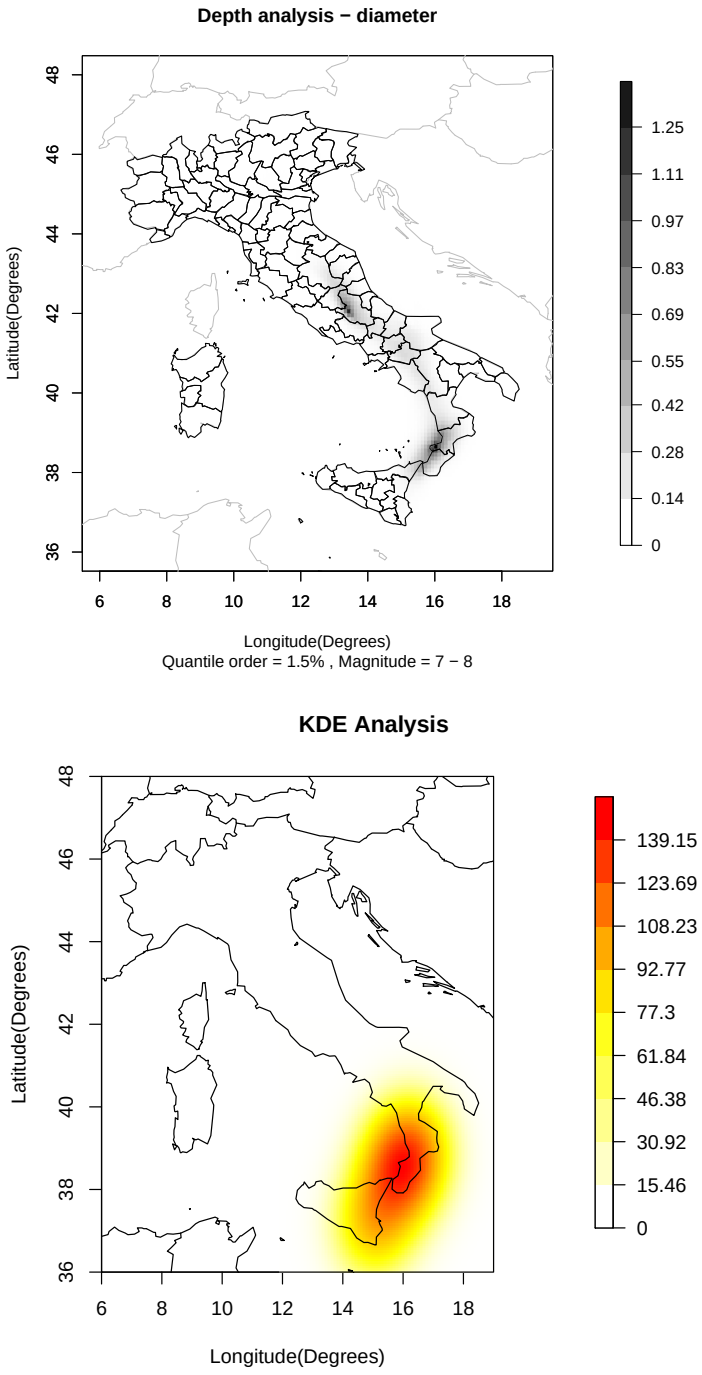


Figure 26: Italian map of seismic risk; magnitude: 7 – 8. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.84 (density values are multiplied by 1000).

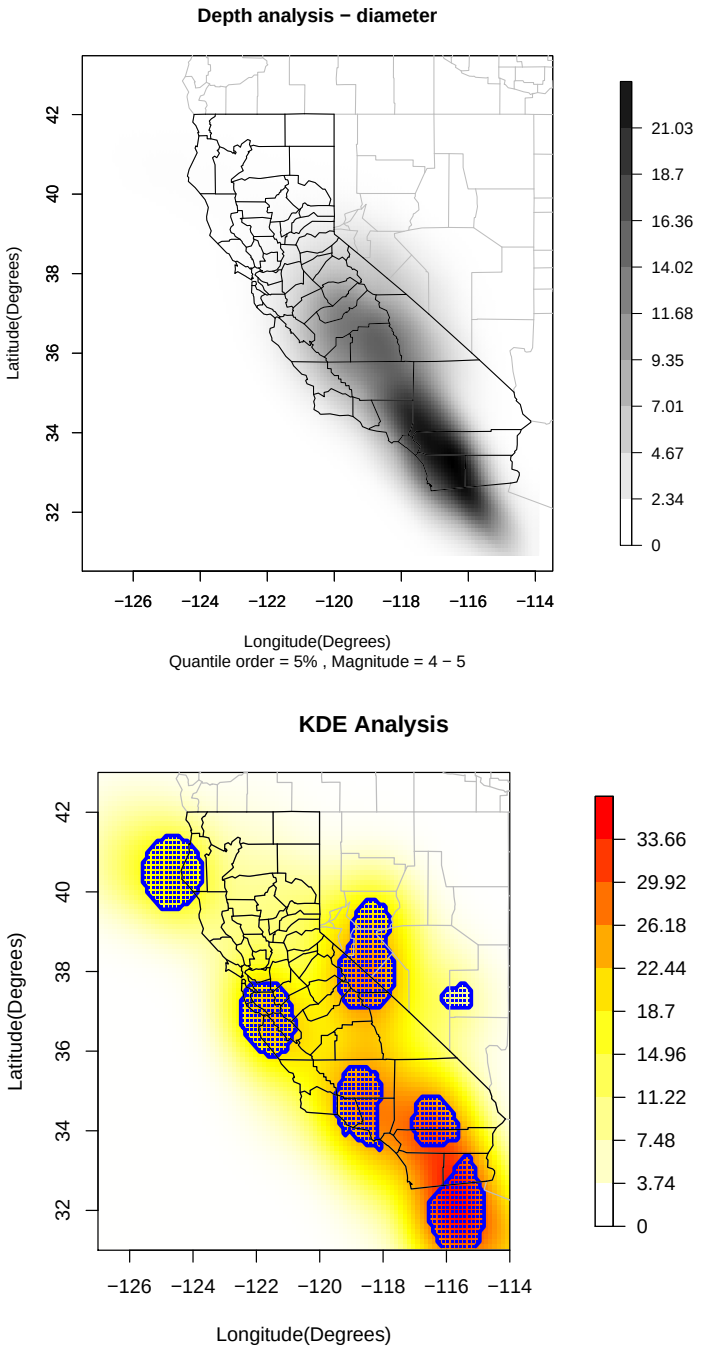


Figure 27: Californian map of seismic risk; magnitude: 4 – 5. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.84 (density values are multiplied by 1000).

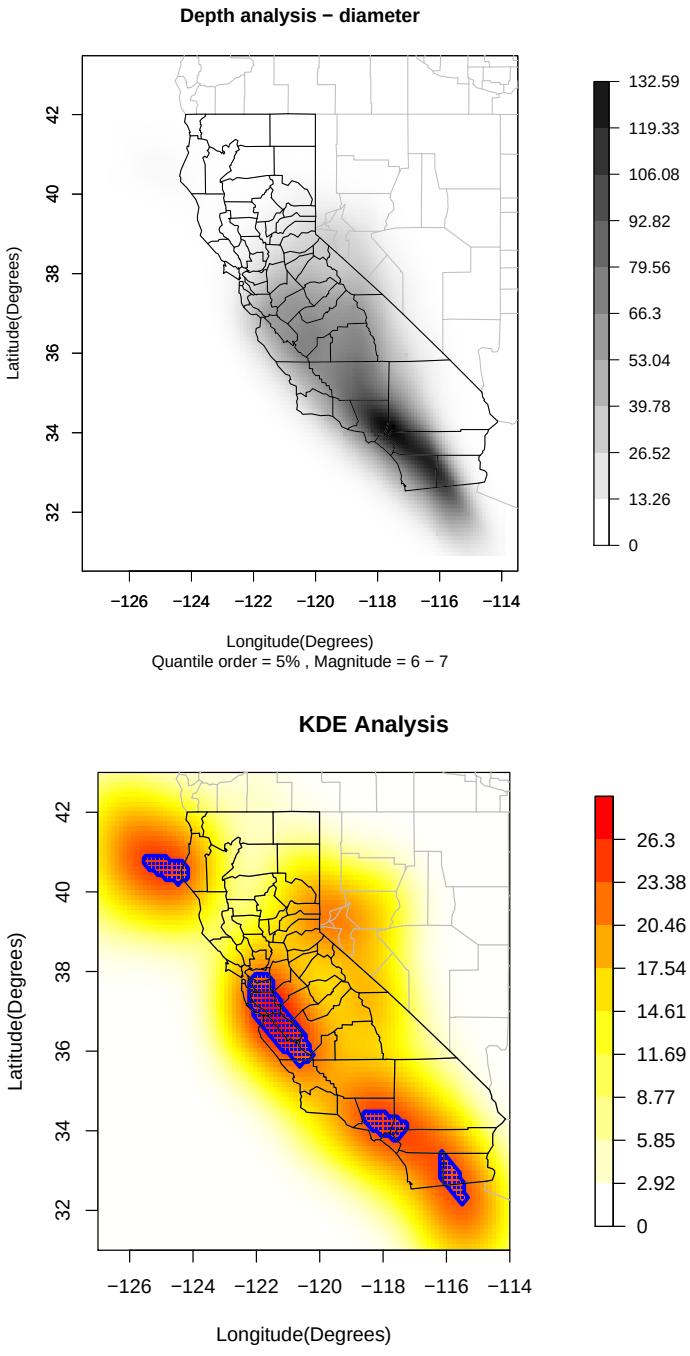


Figure 28: Californian map of seismic risk; magnitude: 6 – 7. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.84 (density values are multiplied by 1000).

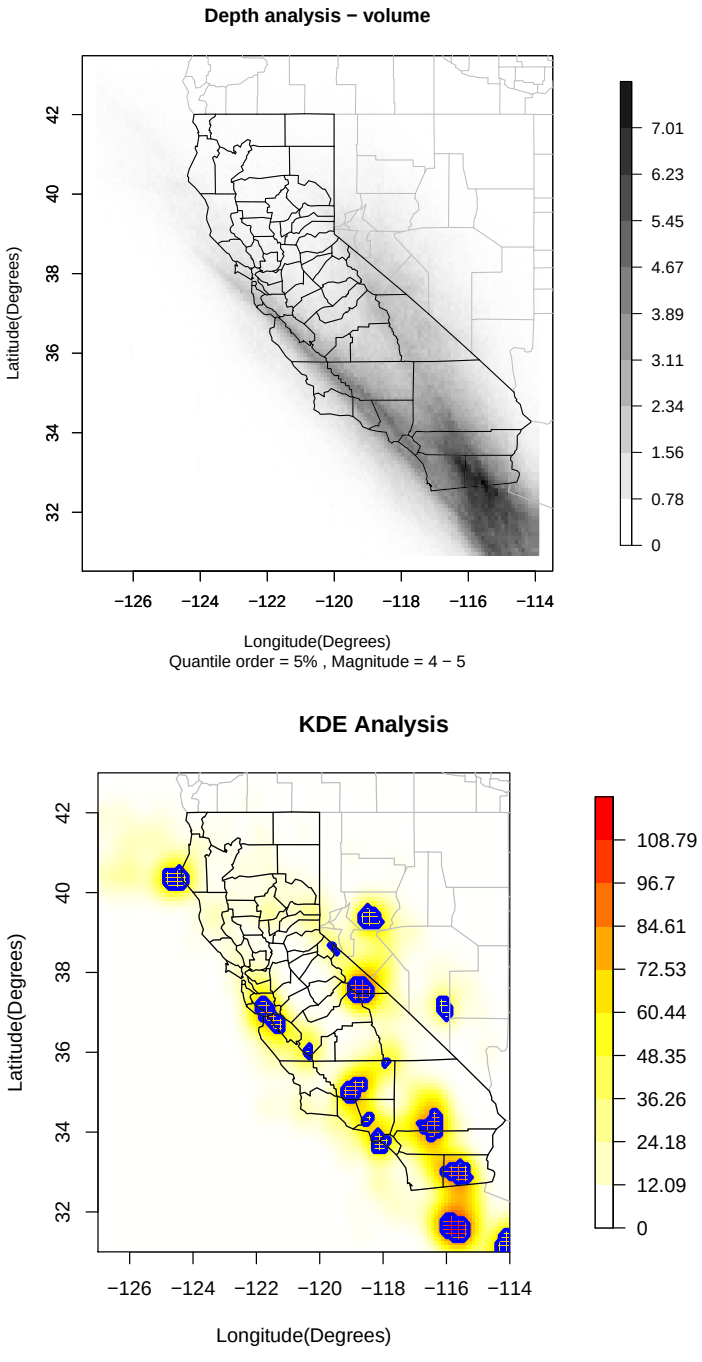
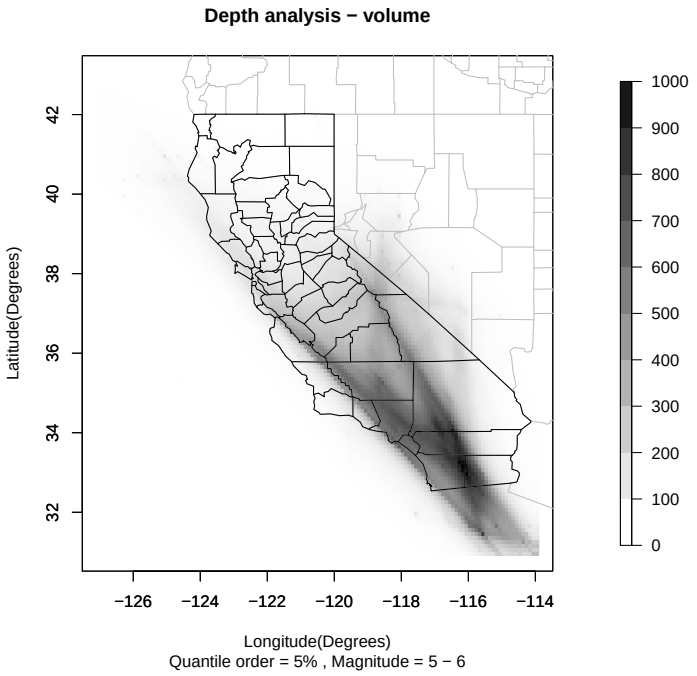


Figure 29: Californian map of seismic risk; magnitude: 4 – 5. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.3 (density values are multiplied by 1000).



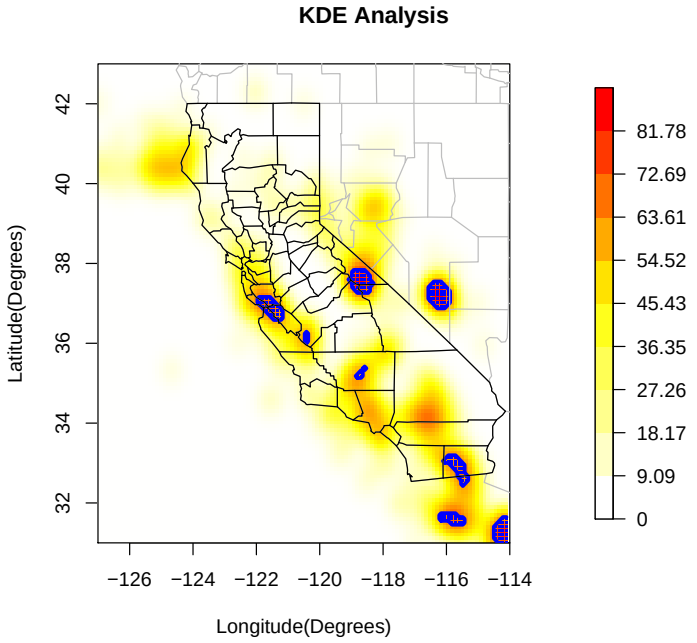


Figure 30: Californian map of seismic risk; magnitude: 5 – 6. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.3 (density values are multiplied by 1000).

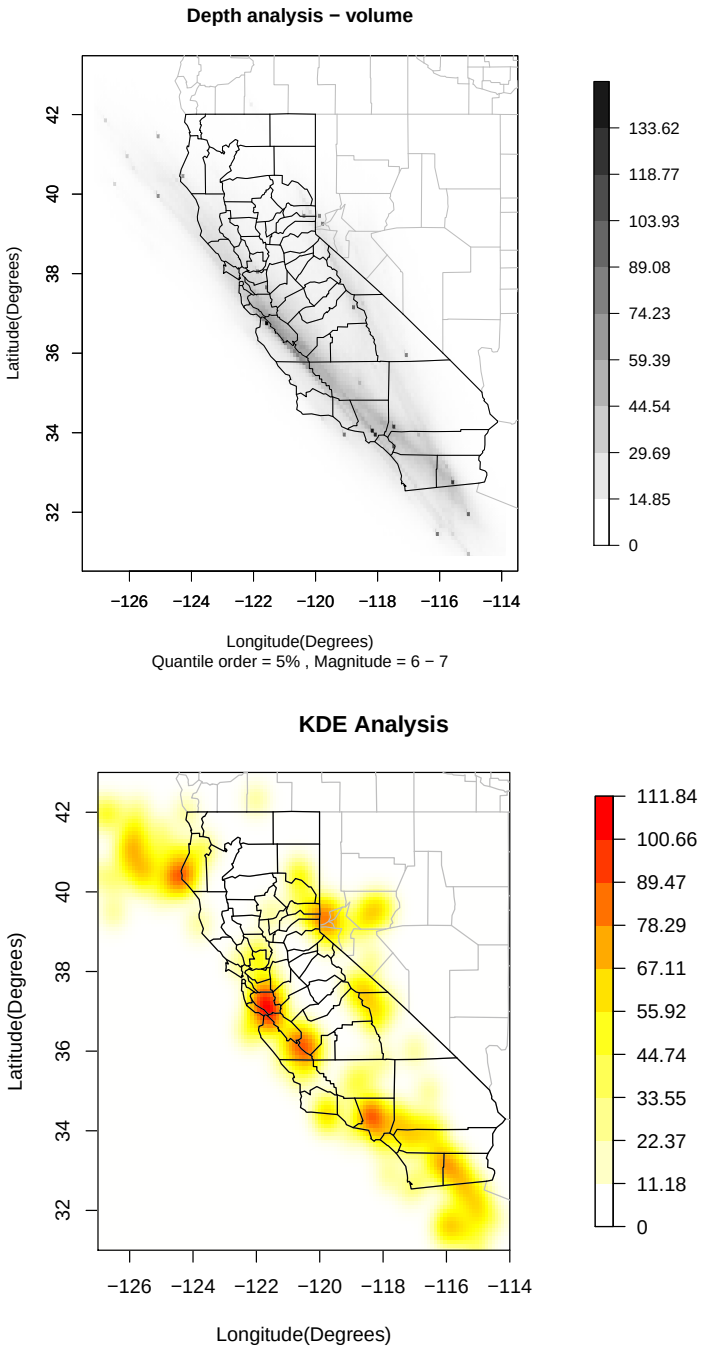


Figure 31: Californian map of seismic risk; magnitude: 6 – 7. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.3 (density values are multiplied by 1000).

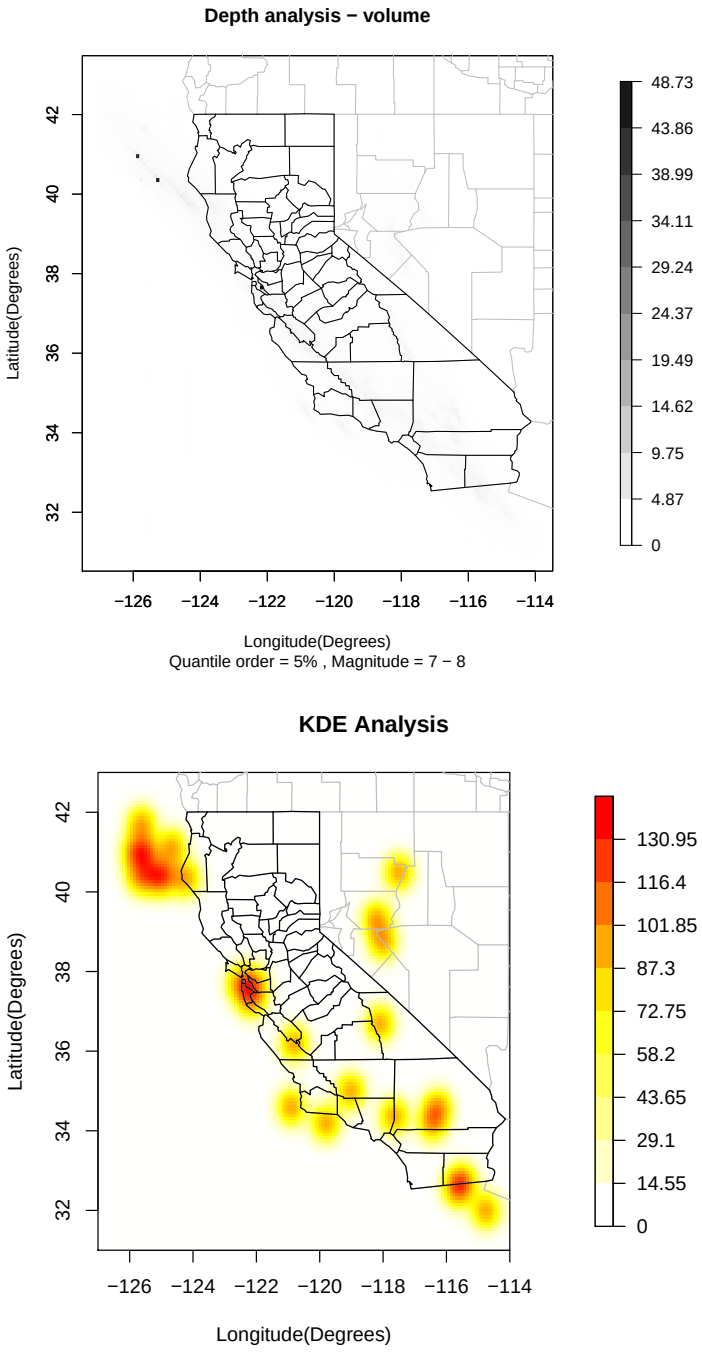


Figure 32: Californian map of seismic risk; magnitude: 7 – 8. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing, bandwidth: 0.3 (density values are multiplied by 1000).

Appendix B: Italy

References

Agostinelli C, Romanazzi M (2011) Local depth. Journal of Statistical Planning and Inference 141:817–830

Albarea A (2012) Statistical analysis of italian earthquake catalogue 1600 - 2003. Tech. rep., Department of Statistics, Ca' Foscari University, Venezia

Boschi E, Guidoboni E, Ferrari G, Valensise G, Gasperini P (1997) Catalogo dei forti terremoti in Italia dal 461 a.C. al 1980, vol. 2. ING-SGA, Bologna

Bowman A, Azzalini A (1997) Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations. Oxford Univ. Press.

Chaudhuri P, Marron J (1999) Sizer for exploration of structures in curves. Journal of the American Statistical Association 94:807–823

Duong T, Hazelton M (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics 15:17–30

Duong T, Hazelton M (2005) Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics 32:485–506

Duong T, Cowling A, Koch I, Wand M (2008) Feature significance for multivariate kernel density estimation. Computational Statistics and Data Analysis 52:4225–4242

Godtliebsen F, Marron J, Chaudhuri P (2002) Significance in scale space for bivariate density estimation. Journal of Computational and Graphical Statistics 11:1–22

Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference. The Annals of Statistics 27:783–858

Petersen M, Bryant W, Cramer C, Cao T, Reichle M, Lienkaemper J, McCrory P, Schwartz D (1996) Probabilistic seismic hazard assessment for the state of california. California Division of Mines and Geology Open-File Report 96-08, US Geological Survey Open-File Report 96-706

Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. Ann Math Statist,Number 3 27:832–837

Sheater SJ (2004) Density estimation. Statistical Science 19:588–597

Silverman B (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London

Simonoff J (1996) Smoothing Methods in Statistics. Springer, New York

Small CG (1990) A survey of multidimensional medians. International Statistical Review 58:263–277

Toppozada T, Branum D, Reichle M, CL H (2002) San Andreas fault zone, California: M $\geq$ 5.5 earthquake history. Bulletin of the Seismological Society of America 92:2555–2601

Wand MP, Jones MC (1995) Kernel Smoothing. Chapman & Hall, London

Woessner J, Hardebeck J, Haukkson E (2010) What is an instrumental seismicity catalog. Community Online Resource for Statistical Seismicity Analysis

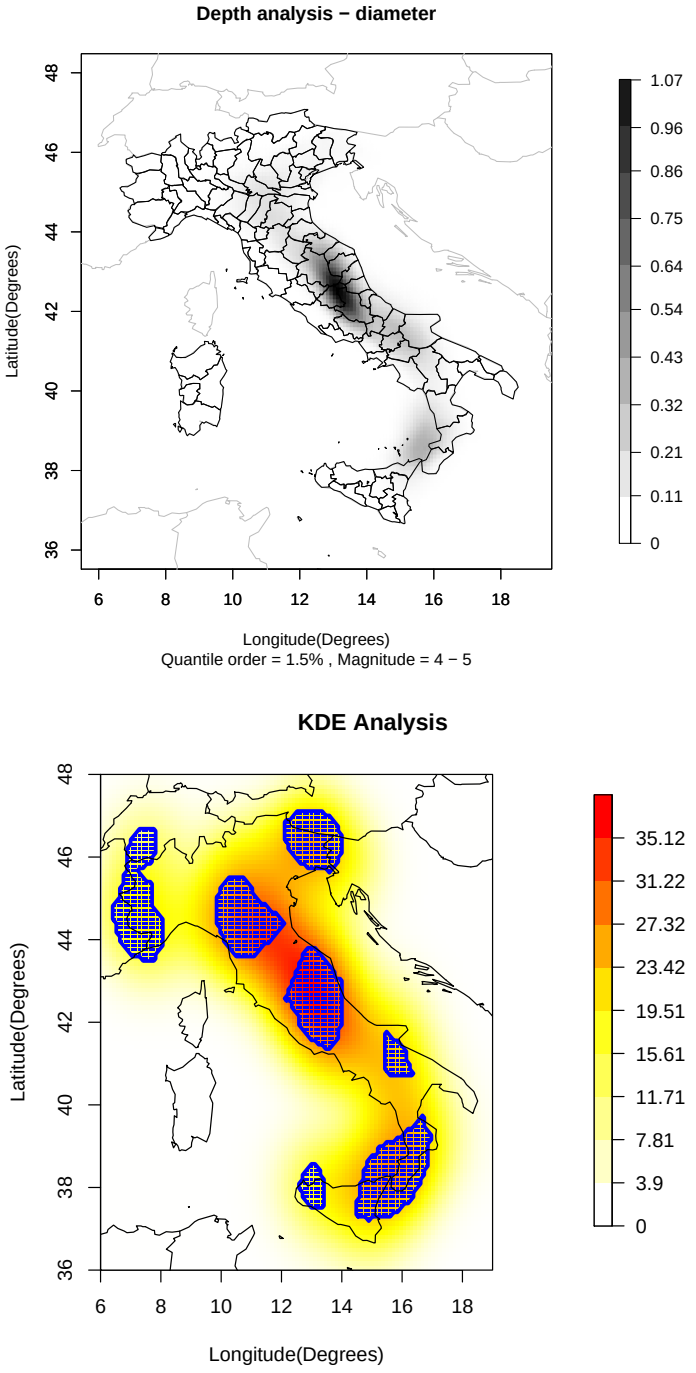


Figure 33: Italian map of seismic risk; magnitude: 4 – 5. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing bandwidth = 0.84, (density values are multiplied by 1000).

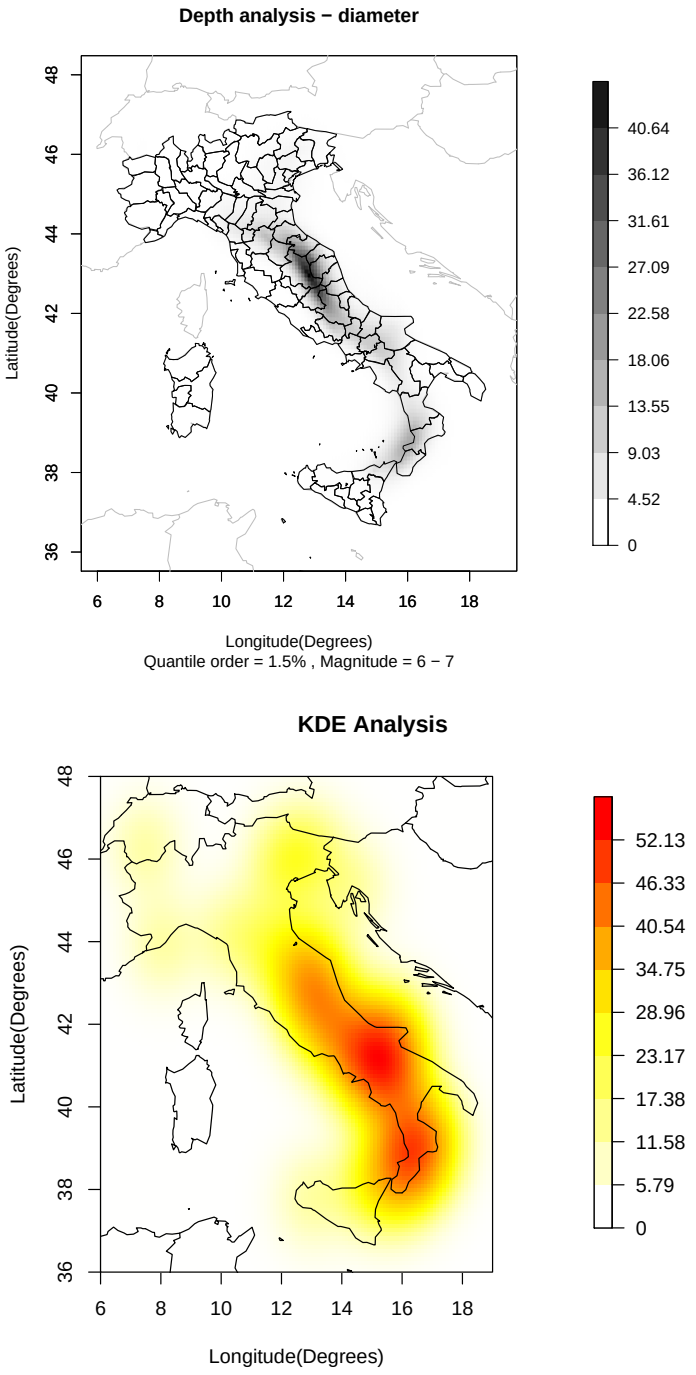


Figure 34: Italian map of seismic risk; magnitude: 6 – 7. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing bandwidth = 0.84, (density values are multiplied by 1000).

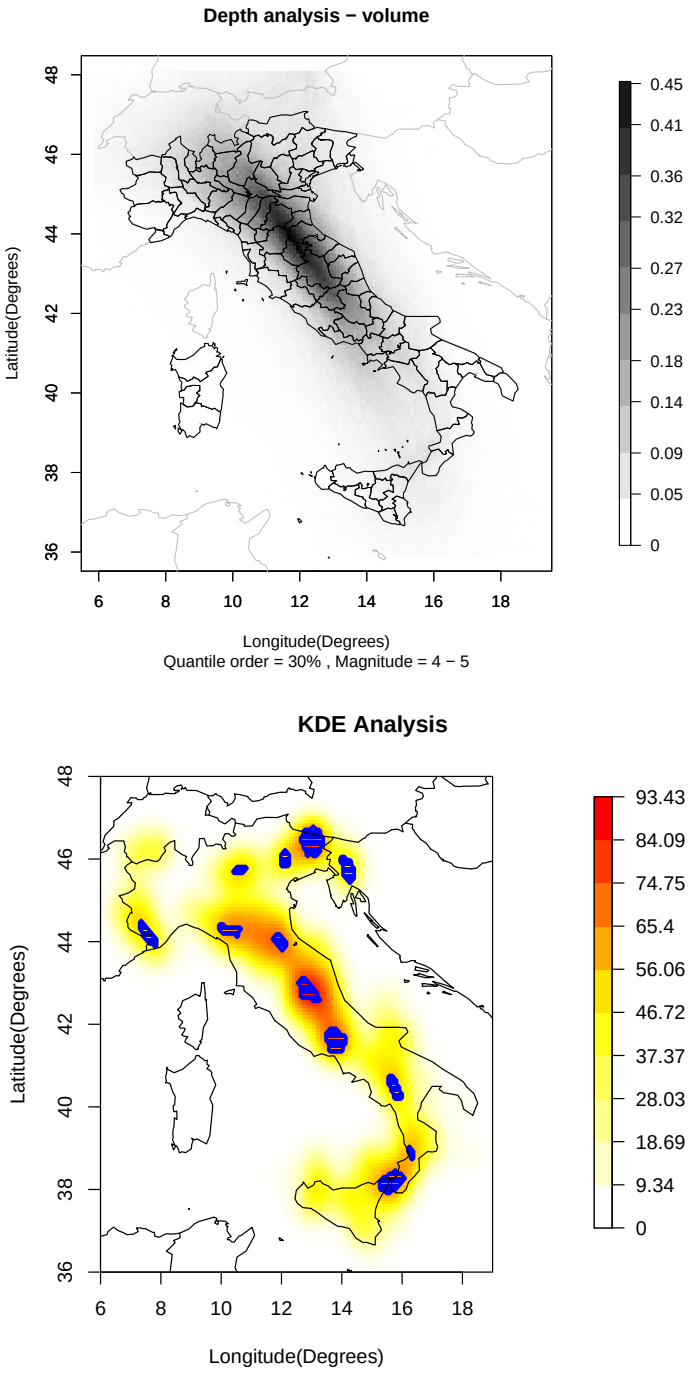


Figure 35: Italian map of seismic risk; magnitude: 4 – 5. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing bandwidth = 0.33, (density values are multiplied by 1000).

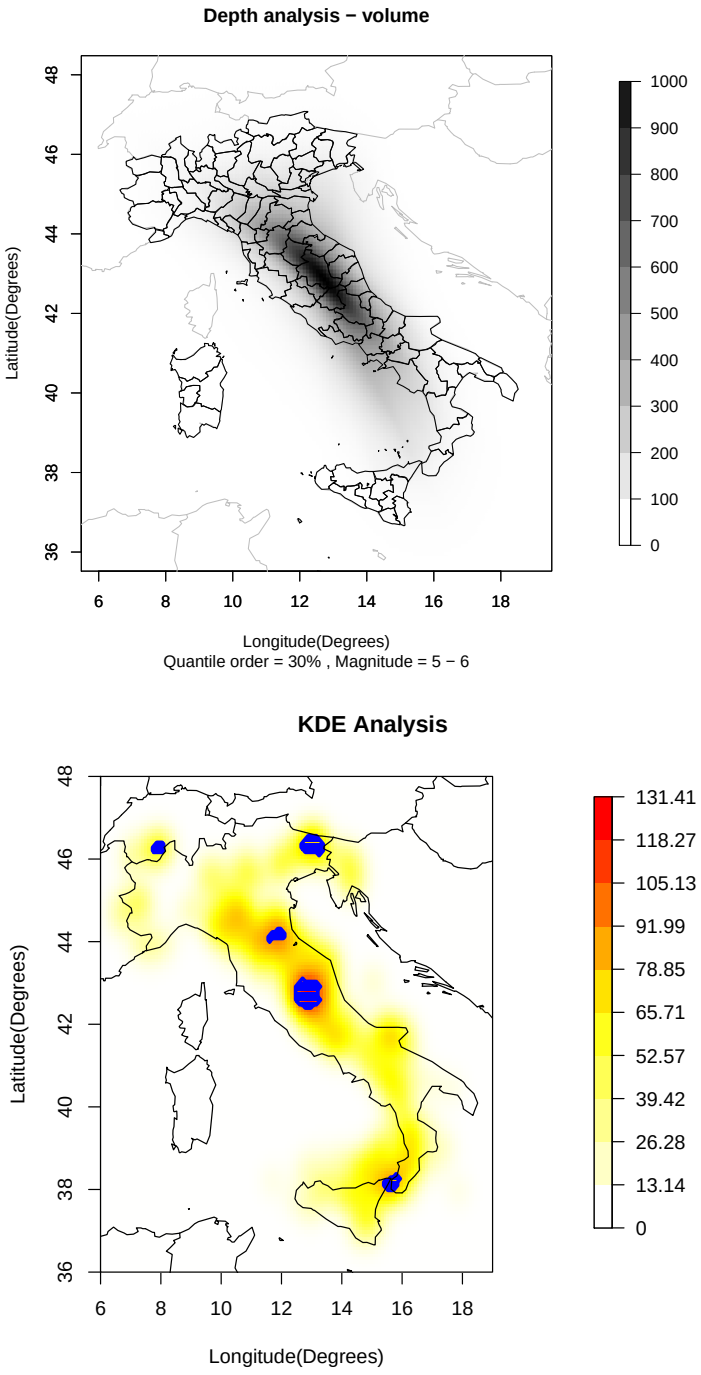


Figure 36: Italian map of seismic risk; magnitude: 5 – 6. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing bandwidth = 0.33, (density values are multiplied by 1000).

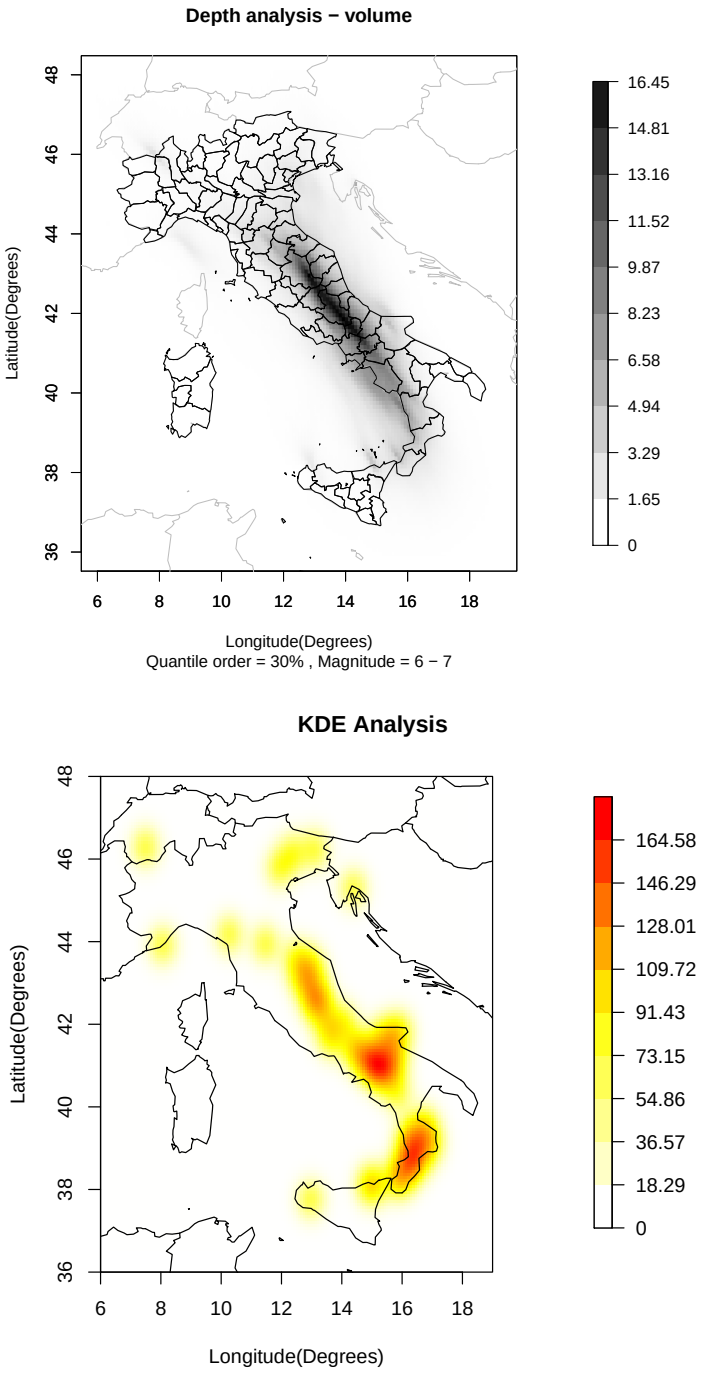


Figure 37: Italian map of seismic risk; magnitude: 6 – 7. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing bandwidth = 0.33, (density values are multiplied by 1000).

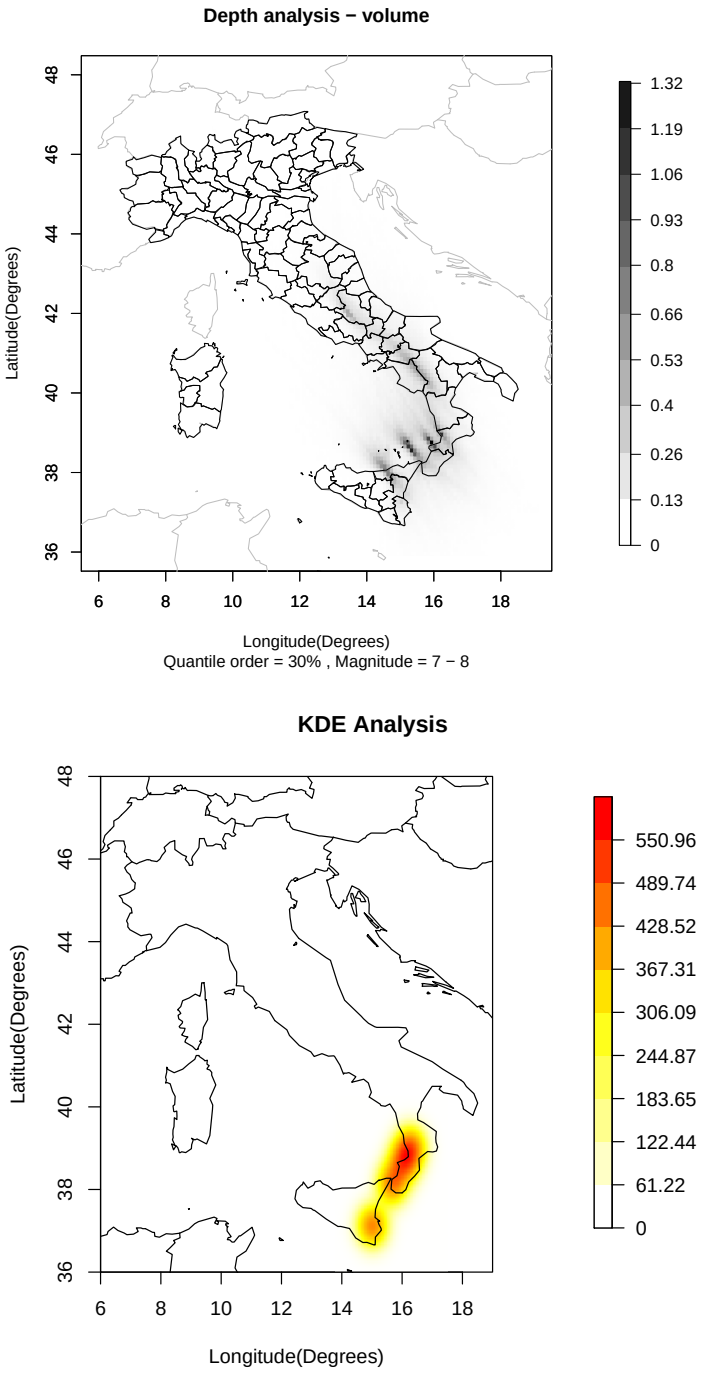


Figure 38: Italian map of seismic risk; magnitude: 7 – 8. Top: depth analysis (values are multiplied by 1000). Bottom: kernel smoothing bandwidth = 0.33, (density values are multiplied by 1000).