RESAMPLING METHODS
IN NONPARAMETRIC SEISMIC HAZARD ESTIMATION

Beata ORLECKA-SIKORA

Department of Geophysics, Faculty of Geology
Geophysics and Environmental Protection
University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: orlecka@geolog.geol.agh.edu.pl

Abstract

Proper estimation of the cumulative distribution function (CDF) of magnitude is essential for the probabilistic seismic hazard analysis. Due to a possibility of the multi-component structure of frequency-magnitude relation, the recently proposed nonparametric approach to the seismic source size characterization is often in use. Because of its model-free character, the nonparametric approach cannot, however, provide confidence interval estimates of CDF. As a remedy to this problem we propose a use of resampling, namely the bootstrap and jackknife techniques for constructing the confidence intervals of magnitude CDF on the basis of one-sample data. Studies of Monte Carlo generated magnitude data of given population distributions show that the bias corrected and accelerated method (BC$_a$ method) is a satisfactory solution of the considered problem. Even more accurate CDF confidence interval estimates can be obtained by the proposed modification of the BC$_a$ procedure. The procedure uses the second order bootstrap samples and is named the iterated BC$_a$ method.

Key words: seismic hazard, cumulative distribution function magnitude.

1. INTRODUCTION

Seismic hazard analysis is a standard tool used to estimate the expected maximum level of ground motion intensity which is generated by seismic events. There are two
possible approaches to estimating the seismic hazard: the deterministic and the probabilistic. The Probabilistic Seismic Hazard Analysis (PSHA) was originally proposed by Cornell (1968). One of its steps is the probabilistic characterization of the seismic source. The most often used parameter to represent the source is the probability that the specified value of seismic event magnitude $M_p$ is exceeded in $D$ time units

$$R(M_p, D) = 1 - \exp\left(-\lambda \cdot D \left[1 - F(M_p)\right]\right),$$

where $\lambda$ is the mean activity rate, $F(\cdot)$ is the Cumulative Distribution Function (CDF) of magnitude. The quantity $R$ is often named the seismic hazard. It arises from the above definition that in order to evaluate the magnitude exceedance probability, $R(M_p, D)$, one must know the magnitude distribution for the studied region, $F(m)$.

Due to the fact that observed frequency-magnitude relations for seismic data can exhibit a multi-component structure in both, natural earthquake and mining induced seismic data (Lasocki, 2001; 2002), Kijko et al. (2001) proposed a nonparametric, model-free estimator of magnitude distribution

$$\hat{F}_N(m|m_{\text{max}}) = \begin{cases} 0 & \text{for } m < m_{\text{min}} \\ \frac{\sum_{i=1}^n \left[ \Phi\left( \frac{m - m_i}{\alpha_i \cdot h} \right) - \Phi\left( \frac{m_{\text{min}} - m_i}{\alpha_i \cdot h} \right) \right]}{\sum_{i=1}^n \Phi\left( \frac{m_{\text{max}} - m_i}{\alpha_i \cdot h} \right) - \Phi\left( \frac{m_{\text{min}} - m_i}{\alpha_i \cdot h} \right)} & \text{for } m_{\text{min}} \leq m \leq m_{\text{max}} \\ 1 & \text{for } m > m_{\text{max}} \end{cases}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\xi^2/2) d\xi$ is the standard Gaussian CDF, $m_{\text{min}}$ is the level of catalogue completeness, $m_{\text{max}}$ is the upper limit of magnitude range, $\alpha_i$ are the local bandwidth factors and $h$ is a positive smoothing factor. The studies showed that the nonparametric estimates, in their mean values, provide good results regardless how complicated the magnitude distribution is (Kijko et al., 2001).

In practical studies, one data sample is analysed and thus no mean estimate but one-sample estimate is obtained. In such cases the estimated magnitude distribution can differ from its actual value and in result this approach can lead to an over- or under-estimation of seismic hazard of unknown percentage. In order to avoid these problems, the point estimation should be replaced by the interval estimation of expected value of magnitude distribution.

The interval estimation would be possible if the CDF estimator distribution were known, which is not the case in the nonparametric approach. The estimator distribution can be approximately reconstructed from many samples. This solution is used in
studies based on simulations but cannot be proposed for practical hazard investigations.

We propose in the present work to approach the problem of magnitude CDF interval estimation using resampling methods, that is, bootstrap and jackknife. These methods enable us to get many samples from one actual sample, so that all of them have asymptotically the same statistical properties as the original sample (Hall, 1992; Efron and Tibshirani, 1998).

The first studies performed on simulated data showed that resampling methods can improve nonparametric seismic hazard estimation (Orlecka and Lasocki, 2002). In the present paper, the use of resampling methods for estimating confidence intervals for the cumulative distribution function of magnitude and seismic hazard (eq. 1) is investigated. An interval estimation algorithm appropriate for nonparametric problems was presented by Efron (1987). Although, in general, this algorithm turned out to be a good tool for the studied specific problem, in order to obtain a better accuracy of constructed confidence intervals for low-size samples, a special treatment is required. To deal with such samples, a modification of Efron’s algorithm is proposed. The modified technique provides satisfactory results regardless of the sample size and actual shape of magnitude distribution.

2. BOOTSTRAP AND JACKKNIFE RESAMPLING

The objective of resampling is to replicate a data sample in such a way that the replicas are not identical with the original but preserve its probabilistic properties. The bootstrap and jackknife are two different techniques of resampling. A bootstrap sample \( x^* = (x_1^*, x_2^*, ..., x_n^*) \) is obtained by randomly sampling \( n \) times, with replacement, from the original data points \( x_1, x_2, ..., x_n \). Let \( \theta \) be a parameter and \( \widehat{\theta} = s(x_1, x_2, ..., x_n) \) be its sample estimate. Each bootstrap sample corresponds to a bootstrap replication \( \widehat{\theta}^* \) of parameter \( \theta \), where

\[
\widehat{\theta}^* = s(x^*)
\]

(Efron and Tibshirani, 1998).

The jackknife differs from bootstrap in sample generation manner. Given a data set \( x = (x_1, x_2, ..., x_n) \), the \( i \)-th jackknife sample \( x_{(i)\text{jack}} \) is defined to be \( x \) with the \( i \)-th data point removed (Efron and Tibshirani, 1998),

\[
x_{(i)\text{jack}} = (x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n)
\]

for \( i = 1, 2, ..., n \). The \( i \)-th jackknife replication \( \hat{\theta}_{(i)\text{jack}} \) of parameter \( \theta \) is \( s(\cdot) \) evaluated for \( x_{(i)\text{jack}} \), say

\[
\hat{\theta}_{(i)\text{jack}} = s(x_{(i)\text{jack}})
\]
Among some other methods for bootstrap confidence interval construction, Efron (1987) proposed the bias corrected and accelerated method (BCa method) especially for nonparametric problems. Let from now on $\theta$ denote a value of the actual magnitude CDF for a given magnitude value. According to BCa method, the $(1-2\alpha)$ probable confidence interval of $\theta$ is given by

$$\left(\hat{\theta}^*_{\alpha_1}; \hat{\theta}^*_{\alpha_2}\right),$$

(6)

where $\hat{\theta}^*_{\alpha_1}$ and $\hat{\theta}^*_{\alpha_2}$ are bootstrap estimated percentiles of $\theta$ and

$$\alpha_1 = \Phi\left(\frac{\hat{z}_0 + \hat{z}_a + z_{\alpha}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha})}\right), \quad \alpha_2 = \Phi\left(\frac{\hat{z}_0 + \hat{z}_a + z_{1-\alpha}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha})}\right).$$

(7)

Here $\Phi(\cdot)$ is the standard normal cumulative distribution function, and $z_{\alpha}$ and $z_{1-\alpha}$ are the percentiles of the standard normal distribution. In order to evaluate $\hat{\theta}^*_{\alpha_1}, \hat{\theta}^*_{\alpha_2}$ and the value of bias-correction, $\hat{z}_0$, a number (say $k$) of bootstrap estimates of $\theta$, $\hat{\theta}^{*\cdot(1)}, \ldots, \hat{\theta}^{*\cdot(n)}$ is drawn. The estimates $\{\hat{\theta}\}$ are sorted in ascending order leading to a series $\hat{\theta}^{*\cdot(1)}, \ldots, \hat{\theta}^{*\cdot(n)}$ such that $\hat{\theta}^{*\cdot(i)} \leq \hat{\theta}^{*\cdot(i+1)}$ for every $i$. The percentile estimate, $\hat{\theta}^*_{\alpha}$, is the $k\alpha$-th element of the series $\hat{\theta}^{*\cdot(1)}, \ldots, \hat{\theta}^{*\cdot(n)}$. $\hat{z}_0$ is obtained directly from the proportion of bootstrap replications that are less than the parameter estimate $\hat{\theta}$ obtained from the original sample data

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\text{number of bootstrap replications} \hat{\theta}^{*\cdot(i)} < \hat{\theta}}{k}\right),$$

(8)

where $\Phi^{-1}(\cdot)$ indicates the inverse function of standard normal CDF. The acceleration constant $\hat{a}$ can be evaluated in different ways, for example,

$$\hat{a} = \left[\frac{1}{n} \sum_{i=1}^{n} (\hat{\theta}^{*\cdot(i)} - \hat{\theta}^{*\cdot(\text{jack})})^2\right]^{1/2},$$

(9)

where $\hat{\theta}^{*\cdot(\text{jack})}$ denotes $i$-th jackknife replication of parameter $\theta$ and $\hat{\theta}^{*\cdot(\text{jack})} = \frac{1}{n} \sum_{\text{jack}} \hat{\theta}^{*\cdot(\text{jack})}$ and $n$ is the original magnitude sample size.

Studies of possibilities to use the BCa method in the seismic hazard interval estimation were carried on using Monte Carlo generated data of a given population distribution. We investigated the performance of the BCa procedure in relation to the sample size and complexity of underlying magnitude distribution. Three models of magnitude distribution were considered (see Table 1). The first model is exponential and
represents a magnitude population which follows the Gutenberg-Richter relation. The second model is a mixture of truncated exponential and normal distributions. Such a model represents an abnormally increased occurrence probability of events with magnitudes close to the expected value of the normal distribution, e.g., the characteristic earthquake model (e.g., Schwartz and Coppersmith, 1984; Romanowicz and Rundle, 1993). The third model is composed of two truncated exponential distributions. It describes populations following a kinked log-linear magnitude-frequency relations (e.g., Knopoff, 2000).

For every model, samples of 400, 200, 100 and 50 random magnitudes were generated. For every sample, a set of 1000 bootstrap samples was drawn. Each of the bootstrap samples was used to estimate magnitude CDF according to formula (2). Finally, the BCₐ confidence intervals for CDF were calculated. The obtained confidence interval estimates of magnitude CDF were compared with the actual CDF values of the models. The results for Model I and 400 and 100 magnitude samples are presented in Figs. 1 and 2, respectively.

The obtained results show that the BCₐ algorithm in general provides reasonable confidence interval estimates for nonparametric CDF. The intervals contain actual CDF values for a wide range of tested magnitudes. The intervals become wider when the sample size decreases.

<table>
<thead>
<tr>
<th>Model</th>
<th>Probability density function</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$f(m) = 2.07 \exp[-2.07(m - 4.3)]$</td>
<td>$b = 0.9$ (or $\beta = 2.07$) $m_{\text{min}} = 4.3$</td>
</tr>
<tr>
<td>II</td>
<td>$f(m</td>
<td>m_{\text{max}}) = 0.9 \frac{2.07 \exp[-2.07(m - 5.5)]}{1 - \exp[-2.07(8.7 - 5.5)]} + 0.1 \frac{1}{\sqrt{2 \cdot 0.3 \pi}} \exp \left[-0.5(m - 7.0)^2\right] (0.3)^2$</td>
</tr>
<tr>
<td>III</td>
<td>$f(m</td>
<td>m_{\text{max}}) = 2.46 \lambda \exp[-2.46(m - 5.5)]$ for $5.5 \leq m \leq 7.4$ $3.68 \mu \exp[-2.46(m - 5.5)]$ for $m &gt; 7.4$ $\lambda = \left{1 - \left(1 - \frac{2.46}{3.68}\right) \exp[-2.46(7.4 - 5.5)]\right}^{-1}$ $\mu = \frac{2.46}{3.68} \exp[-3.68(7.4 - 5.5)]$ $3.68 \exp[-2.46(7.4 - 5.5)]$</td>
</tr>
</tbody>
</table>
The results for the samples drawn from the complex distributions are similar. Figs. 3, 4 and 5 present the results obtained from a 400 event sample drawn from the distribution Model II and from 200 and 50 event sample drawn from the distribution Model III, respectively.

The presented examples confirm that the BC$_3$ method is effective in estimating the confidence intervals of magnitude CDF regardless of the complexity of the magnitude distribution.

It is, however, visible in the presented examples that when the sample size is reduced, the magnitude range for which the intervals are estimated becomes narrower.
Fig. 3. BCₐ estimates of 95% confidence interval of magnitude CDF. The data sample of 400 events was drawn from the distribution Model II.

Fig. 4. BCₐ estimates of 95% confidence interval of magnitude CDF. The data sample of 200 events was drawn from the distribution Model III.

Fig. 5. BCₐ estimates of 95% confidence interval of magnitude CDF. The data sample of 50 events was drawn from the distribution Model III.
The example of this effect can be noted in Figs. 4 and 5, which present results for the distribution Model III (complex model). For the 200 element sample, the confidence intervals seem to be properly estimated from 5.5 to 7.5 magnitude value (Fig. 4). For the 50 element sample such a range spreads only from 5.7 to 6.5. Moreover, in every case the upper limit of confidence interval is usually closer to the actual CDF value than the lower limit. This suggests that the $BC_n$ confidence intervals for CDF are shifted downwards systematically with respect to the actual confidence intervals.

Fig. 6. Theoretically calculated 95% confidence intervals of magnitude CDF and its $BC_n$ estimates. The data sample of 100 events was drawn from the exponential model of magnitude distribution (Model I). The part for high magnitude values is magnified.
In order to compare estimates obtained by the BCₐ algorithm with actual confidence intervals, the theoretical confidence intervals were calculated for the exponential model of magnitude CDF (distribution Model I). The comparison showed in Fig. 6 confirms the suggestion on the systematic shift of the BCₐ confidence intervals in relation to the theoretical confidence intervals. The effect is the most significant for small magnitudes. In the most interested range of high magnitude values, the BCₐ confidence intervals become narrower but tend to locate above the theoretical confidence intervals. Eventually, the actual CDF values appear outside the estimated intervals.

4. ITERATED BOOTSTRAP APPLIED TO BCₐ ALGORITHM AND SOME RESULTS OF THIS PROCEDURE

In order to improve the accuracy of magnitude CDF bootstrap confidence intervals obtained by BCₐ algorithm, we propose to use iterated bootstrap to estimate the bias-correction parameter  \( z_\alpha \). Namely, for each of 1000 bootstrap samples a set of 200 second order bootstrap samples is drawn and on the basis of them the estimate of parameter  \( z_\alpha \) is calculated. In this way 1000  \( z_\alpha \) values are obtained. Their mean value  \( \bar{z}_\alpha \) is used to calculate  \( \alpha_1 \) and  \( \alpha_2 \) values.

The magnitude CDF confidence intervals obtained by the BCₐ procedure and by the iterated BCₐ algorithm from a 100 element data sample drawn from the exponential distribution are compared with the theoretical confidence intervals in Fig. 7. One can notice that the iterated BCₐ procedure provides better results than standard BCₐ method: the confidence intervals obtained in the first way are more accurate.

![Fig. 7. Theoretically calculated 95% confidence intervals of magnitude CDF and their BCₐ, and iterated BCₐ estimates. The data sample of 100 events was drawn from the exponential model of magnitude distribution.](image-url)
Fig. 8. Standard and iterated BCa estimates of 95% confidence intervals of magnitude CDF. The data sample of 100 events was drawn from the Gutenberg-Richter relation based distribution (distribution Model I).

Fig. 9. Standard and iterated BCa estimates of 95% confidence intervals of magnitude CDF. The data sample of 50 events was drawn from the distribution Model III.

The iterated BCa confidence intervals were also calculated for every model presented above, for data samples consisting of 400, 200, 100 and 50 random magnitudes. Figures 8 and 9 present some results for a 100 event sample drawn from the distribution Model I and for a 50 element sample drawn from the distribution Model III, respectively.

The presented examples show that the iterated BCa confidence intervals of magnitude CDF are more accurate than the confidence intervals calculated with the standard BCa algorithm. Limits of intervals obtained from the iterated procedure are similar
to such limits from the standard procedure, but in many cases the former are smoother than the latter.

An important advantage of the iterated BC$_a$ method is that in many cases this method enables to estimate the confidence intervals for wider magnitude range than the standard BC$_a$ approach. The example presented in Fig. 10 shows that the iterated BC$_a$ confidence intervals are acceptable to magnitude 6.75, while those provided by the standard BC$_a$ algorithm are acceptable to magnitude 6.5.

![Fig. 10. Standard and iterated BC$_a$ estimates of 95% confidence intervals of magnitude CDF. The data sample of 100 events was drawn from the distribution Model III.](image)

The magnitude CDF confidence intervals can be easily transformed into the confidence intervals of the seismic hazard. Such intervals for the seismic hazard are shown in Figs. 11 and 12.

5. CONCLUSIONS

The results obtained on simulated data showed that resampling methods can improve significantly the seismic hazard analysis. Resampling is the only way that enables to calculate the interval estimates of magnitude CDF and seismic hazard on the basis of one data set.

The proposed bootstrap approach to the seismic hazard confidence interval assessment is independent of the underlying magnitude distribution. The use of the non-parametric estimator of magnitude CDF allows to construct confidence intervals re-
Theoretically calculated seismic hazard of $M_p = 9.0$ and 95% confidence interval estimates of seismic hazard. The data sample of 400 events was drawn from the Gutenberg-Richter relation based distribution (the distribution Model I).

Fig. 11. Theoretically calculated seismic hazard of $M_p = 9.0$, $\lambda = 9.0$ and 95% confidence interval estimates of seismic hazard. The data sample of 400 events was drawn from the Gutenberg-Richter relation based distribution (the distribution Model I).

Fig. 12. Theoretically calculated seismic hazard of $M_p = 7.3$, $\lambda = 9.0$ and 95% confidence interval estimates of seismic hazard. The data sample of 400 events was drawn from the distribution Model III.

gardless of whether the actual magnitude distribution follows the Gutenberg-Richter relation or is complex.

All the examples show that it is possible to construct the CDF confidence intervals even if the data sample comprises a small number of elements.

Further studies are required to evaluate the performance of this approach as a function of the number of bootstrap samples.
Acknowledgments. This work was financed by the Polish State Committee for Scientific Research under contract No. 6 P04D 01820 during the period 2001-2003.

References

Received 16 June 2003
Accepted 1 July 2003