



Assessment of the uncertainty in spatial-correlation models for earthquake ground motion due to station layout and derivation method

Erika Schiappapietra¹ · John Douglas¹

Received: 14 May 2021 / Accepted: 5 July 2021
© The Author(s) 2021

Abstract

The evaluation of the aggregate risks to spatially distributed infrastructures and portfolios of buildings requires quantification of the estimated shaking over a region. To characterize the spatial dependency of ground motion intensity measures (e.g. peak ground acceleration), a common geostatistical tool is the semivariogram. Over the past decades, different fitting approaches have been proposed in the geostatistics literature to fit semivariograms and thus characterize the correlation structure. A theoretically optimal approach has not yet been identified, as it depends on the number of observations and configuration layout. In this article, we investigate estimation methods based on the likelihood function, which, in contrast to classical least-squares methods, straightforwardly define the correlation without needing further steps, such as computing the experimental semivariogram. Our outcomes suggest that maximum-likelihood based approaches may outperform least-squares methods. Indeed, the former provides correlation estimates, that do not depend on the bin size, unlike ordinary and weighted least-squares regressions. In addition, maximum-likelihood methods lead to lower percentage errors and dispersion, independently of both the number of stations and their layout as well as of the underlying spatial correlation structure. Finally, we propose some guidelines to account for spatial correlation uncertainty within seismic hazard and risk assessments. The consideration of such dispersion in regional assessments could lead to more realistic estimations of both the ground motion and corresponding losses.

Keywords Spatial correlation · Estimation uncertainty · Ground-motion · Maximum-likelihood · Least-squares regression

✉ Erika Schiappapietra
erika.schiappapietra@strath.ac.uk

¹ Department of Civil and Environmental Engineering, University of Strathclyde, Glasgow G11XJ, UK

1 Introduction

Many authors (e.g. Iervolino 2013; Weatherill et al. 2015; Sokolov and Ismail-Zadeh 2016; Sokolov and Wenzel 2019) have demonstrated the importance of considering regional hazard estimates when evaluating the aggregate risks to spatially-distributed infrastructure and building portfolios. The assessment of the seismic hazard over a geographical region requires the quantification not only of the expected ground shaking at a single location, but also how this shaking could vary over distances of a few kilometres. This variation is captured within spatial-correlation models. Spatial correlations have been increasingly studied over the last 20 years and many researchers have aimed to identify the factors that most affect the spatial dependency of earthquake ground motions. Schiappapietra and Douglas (2020) provide a thorough literature review, shedding light on the dependence of correlation on: (1) the estimation approach and fitting method; (2) earthquake magnitude; (3) structural period; (4) regional and local site-effects; and (5) ground motion prediction equations (GMPEs). Baker and Chen (2020) propose a novel approach to quantify both the uncertainty in the correlation estimation and the underlying correlation variability among different earthquakes. Further insights into the spatial correlation of ground motions are given by studies on numerical ground motion simulations. In this regard, Stafford et al. (2018), Chen and Baker (2019), Huang et al. (2020), Infantino et al. (2021) and Schiappapietra and Smerzini (2021) provide valuable contributions on the factors that cause the spatial dependency of earthquake ground motion to vary from case to case, with particular emphasis on the earthquake rupture process. In general, studies suggest that the spatial correlation structure is period-, regionally- and scenario-dependent.

Spatial correlation models are usually calibrated on a set of multiple events due to the shortage of ground motion observations from each single earthquake. Using data from a single event would often lead to poorly constrained correlation parameters and models that have limited applicability for future earthquakes. Although it is recognized that the correlation varies from event to event, only few studies (e.g. Goda 2011; Heresi and Miranda 2019) have taken into account such event-to-event correlation variability. The consideration of this dispersion in regional probabilistic risk assessment could lead to more realistic estimations of both the ground motion and corresponding losses. Baker and Chen (2020) demonstrated that the true variability in correlation estimates of poorly-recorded events does not significantly differ from that of well-recorded events and that the differences in terms of apparent total variability are exclusively due to the larger estimation uncertainty of poorly-recorded earthquakes.

In this broad framework, the research question we would like to answer is whether it is best to have a local correlation model, even though it is not well constrained, or to implement a global correlation model, characterized by a lower uncertainty but calibrated on worldwide databases? We, therefore, focus our attention on improving the estimation of correlation parameters by using alternative approaches. This study is a continuation of our previous work (Schiappapietra and Douglas 2020) and it aims to provide guidelines for developers and users of spatial-correlation models. To achieve this goal, we use simulations of spatially-correlated ground motion fields which, as opposed to real data, provide a controlled environment where the true model is known.

Section 2 summaries spatial correlation modelling theory and it introduces the approaches for correlation estimation we use throughout this study. Section 3 describes the steps to generate spatially correlated random fields. We propose here two different studies: (1) ground-motion fields simulated on a fine grid, and (2) ground-motion fields simulated

only at recording locations corresponding to those of past earthquakes. Finally, Sects. 4, 5 and 6 discuss the main results and the implications of this work.

2 Spatial correlation modelling

Traditional seismic hazard and risk analysis tools usually employ GMPEs to estimate the earthquake ground motion at a given site. The earthquake ground motion of interest to engineering is often the transient ground shaking that occurs during an earthquake. This ground motion is invariably evaluated in terms of one or more scalar intensity measures (IMs), such as the peak ground acceleration, the peak ground velocity and, occasionally, the peak ground displacement. The ground motion is also often expressed in terms of response spectral acceleration, which represents the maximum response of a single degree-of-freedom system of a given oscillator period and damping subject to the ground motion time-history. GMPEs provide the marginal probability distribution of the IM at a single site as a function of a set of parameters describing the earthquake source, such as the magnitude, the propagation path and local site conditions (e.g. Douglas and Edwards 2016):

$$\log_{10} Y_{ij} = \log_{10} \overline{Y_{ij}}(M, R, S, \theta) + \varepsilon_{ij} + \eta_i \tag{1}$$

where Y_{ij} is the IM of interest at the j^{th} site due to the i^{th} event, whereas $\overline{Y_{ij}}$ is the predicted median function of magnitude (M), distance from the source (R), local-site conditions (S) and other explanatory variables (θ). ε_{ij} and η_i are the within-event and between-event residuals terms, respectively. ε_{ij} represents systematic deviations between observed and median predicted values due to path and local site effects, whereas η_i denotes systematic deviations associated to an event. For this reason, while ε_{ij} is site-dependent, η_i is common for all sites. Both residual terms are assumed to be normally distributed with mean zero and standard deviations φ and τ , respectively. To fully characterise ε , it is necessary to describe how the within-event residuals vary in space, namely to model the spatial dependence of ε_{ij} and ε_{ik} . Baker and Jayaram (2008) demonstrated that spatially distributed within-event residuals are jointly normally distributed. Therefore, their spatial dependence can be completely defined by the covariance matrix, which reflects their correlation structure.

2.1 Spatial variability of within-event residuals

In geostatistics, a common tool to describe the dependence structure of spatial distributed random variables (i.e. the within-event residuals) is the semivariogram, which measures the average dissimilarity of a pair of ε_{ij} and ε_{ik} separated by an inter-site distance h :

$$\hat{\gamma}(h) = \frac{1}{2} \text{Var}[\varepsilon_{ij} - \varepsilon_{ik}] \tag{2}$$

in which Var indicates the variance. The semivariogram is empirically evaluated from observations by pooling all data with a given inter-site spacing h and then using either the robust estimator proposed by Cressie (1985) or the classic method of moments proposed by Matheron (1962). Usually, the individual separation distances between pairs of observations are grouped into bins, so that the semivariances are computed for each pair of sites whose inter-site distance falls in the interval $[h - \Delta, h + \Delta]$. The hypotheses of second-order stationarity and isotropy are generally assumed due to the lack of repeated ground motion observations from the same event at a given site. Therefore, the correlation between

any pairs of sites with equal separation distance is the same, independently of the source-to-site distance and orientation. Under such assumptions, the semivariogram and the correlation are equivalent and the following relation holds (Diggle and Ribeiro 2007; Oliver and Webster 2014):

$$\gamma(h) = \varphi^2 - COV(\varepsilon_{ij}, \varepsilon_{ik}) = \varphi^2 [1 - \rho_\varepsilon(h)] \tag{3}$$

where COV is the covariance matrix and ρ_ε the correlation function. The reader is referred to Schiappapietra and Douglas (2020) for further details.

2.2 Fitting methods for semivariogram models

The experimental semivariogram of Eq. (2) is a discrete function, describing the spatial continuity of the random variable ε . Parametric functions are used to fit the experimental semivariogram to retrieve semivariogram models for any separation distance h . In the literature, a number of admissible models (e.g. spherical, Gaussian and exponential) exist; however, we choose the exponential function to model the correlation structure, as it is the most widely adopted (e.g. Jayaram and Baker 2009; Esposito and Iervolino 2012; Baker and Chen 2020) functional form in engineering seismology. The general form of the exponential function is:

$$\gamma(h) = a \left[1 - \exp\left(-\frac{3h}{b}\right) \right] \tag{4}$$

where a and b are the sill and the practical range of the semivariogram, respectively. The sill represents the variance of the random variable, whereas the practical range is the separation distance at which $\gamma(h)$ equals 95% of the sill value. An illustration of an empirical and fitted semivariogram model is presented in Fig. 1a. Different fitting approaches have been proposed in the geostatistics literature. In general the model coefficients are chosen so that the misfit between observed and predicted values is minimised. Baker and Chen (2020) provide a useful summary of the most common techniques, such as the ordinary (OLS) and

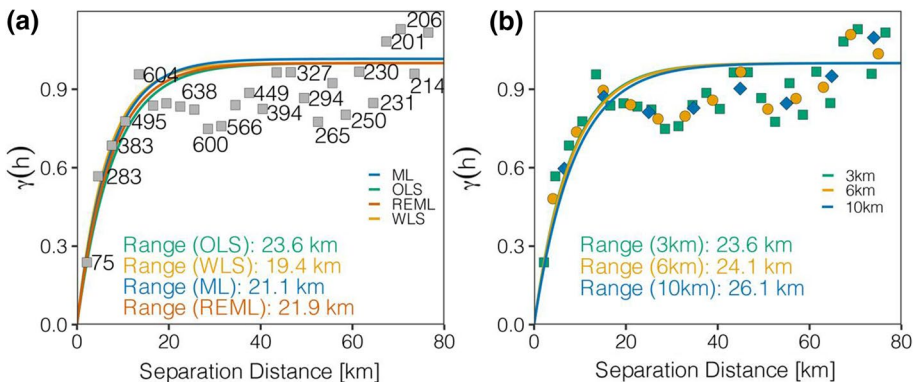


Fig. 1 Empirical and fitted semivariogram models: **a** different techniques to estimate the semivariogram parameters as introduced in this work. The solid line is the exponential fitted model, whereas squares represent the experimental semivariogram. The numbers close to the squares indicate the number of pairs used to compute the semivariances within each bin. **b** Different bin sizes to compute the experimental semivariogram. The exponential models are fitted by using the OLS approach

weighted (WLS) least squares, and they suggest a new weighting function to weight the values from small distance more within the fitting step. A trial-and-error (manual fitting) approach has been chosen by different authors for its versatility in fitting the data. Nevertheless, we discourage performing a visual fit due to its high degree of subjectivity.

In our analysis, we implement the **R** (R Core Team (2019)) package **gstat** (Pebesma 2004) to compute the experimental semivariogram and obtain semivariogram model coefficients by means of the OLS and WLS regression techniques.

2.3 Maximum likelihood estimation

The above-described method of least squares to fit semivariogram models is not direct because it requires the computation of the experimental semivariogram as an intermediate step (Li et al. 2018). Moreover, the final modelling outcomes depend on different assumptions such as the variogram estimators (Oliver and Webster 2014) and weighting functions (Baker and Chen 2020; Schiappapietra and Douglas 2020), and on the introduction of arbitrary parameters such as the bin size. For instance, we plot in Fig. 1b the empirical semivariograms computed by using different bin widths and the corresponding fitted exponential models. Nonetheless, this methodology is the most widely used to determine the dependence structure of spatially distributed random variables and many geostatistical software packages (e.g. **R** packages **geoR** (Ribeiro et al. 2020), **gstat** (Pebesma 2004), **georob** (Papritz 2020) and **MATLAB** functions **variogram** (Schwanghart 2021a), **variogramfit** (Schwanghart 2021b)) allow the user to obtain semivariogram parameter estimates easily (Li et al. 2018). On the other hand, estimation methods based on the likelihood function have increasingly gained influence in geostatistics, particularly in the presence of trends (Oliver and Webster 2014). The parameters of the correlation structure model are directly estimated by maximising a log-likelihood function without needing further steps, such as computing the experimental semivariogram. Despite such an advantage, maximum-likelihood approaches have not been commonly used in engineering seismology. To the authors knowledge, only Ming et al. (2019) employed the maximum-likelihood method to simultaneously estimate the GMPE and correlation function coefficients. One of the main drawbacks of the maximum-likelihood estimation is that it requires the data (e.g. within-event residuals) to be normally distributed. Normality of within-event residuals has been shown to hold, at least within ± 3 standard deviations of the mean (e.g. Strasser et al. 2009).

In our analyses, we take advantages of different techniques, such as the Gaussian maximum likelihood (ML) and the restricted maximum likelihood (REML). In general, the model for a set of geostatistical data $Y_i = Y(x_i)[i = 1, \dots, n]$ at locations x_i is defined as the following:

$$Y(x_i) = D(x_i)^T \beta + B(x_i) + \varepsilon_i \tag{5}$$

where $D(x_i)^T \beta$ is the spatial trend, $B(x_i)$ is the Gaussian random field with zero mean and covariance $R(h, \sigma^2, \alpha)$ and ε_i is an independent distribution error with zero mean and variance τ^2 (nugget effect). σ^2 and α are the sill and range parameters of the covariance function, whereas β is the vector of regression parameters of the spatial trend. For a normally distributed Y , the log-likelihood for the estimation of σ^2 , α and β is defined as:

$$L(\beta, \tau^2, \sigma^2, \alpha) = -0.5 \left\{ n \log(2\pi) + \log \left\{ \left| \sigma^2 R + \tau^2 I \right| \right\} + (y - D\beta)^T (\sigma^2 R + \tau^2 I) (y - D\beta) \right\} \tag{6}$$

Therefore, the model parameters are obtained by maximising the function $L(\beta, \tau^2, \sigma^2, \alpha)$. The reader should refer to Diggle and Ribeiro (2007) and Künsch et al. (2013) for deeper insights into likelihood-based methods. We implement both the **R** packages **geoR** and **georob** for the parameter estimation through maximum-likelihood approaches.

3 Simulations set up

We generate spatially-correlated ground-motion fields (i.e. the within-event residuals at each station location), using a multivariate normal distribution, defined by a zero mean and a covariance function, which reflects the correlation of the within-event residuals. The (unconditional) simulations of Gaussian random fields for given covariance parameters are generated by using the **R** package **geoR**. We opt for an exponential correlation model with correlation length h_0 and we choose different values of h_0 [5, 10, 15, 20, 30, 40 km] to cover the typical estimates reported from ground-motion observations. Such a broad range mainly depends on magnitude, fault mechanism and source effects as well as regional and local-site conditions, even when the same seismic region is considered (e.g. Schiappapietra and Douglas 2020; Infantino et al. 2021; Schiappapietra and Smerzini 2021). A Monte Carlo approach is adopted to generate 1000 simulations of ground-motion residuals for each h_0 to obtain more stable and robust outcomes. Figure 2 illustrates two out of the 1000 simulations generated by imposing correlation lengths of 5 and 15 km, respectively.

For each Monte Carlo simulation (i.e. for each within-event residual distribution), we follow these steps to assess the performance of the different estimation approaches and the influence of different parameters such as the bin size and number of available stations:

- Randomly locate strong-motion recordings stations throughout the region. We select a different number of stations [20, 40, 60, 80, 100] for each h_0 to cover the number of strong-motion stations that earthquakes are usually recorded by.
- Estimate the empirical semivariogram and derive a semivariogram model (Eq. 4) using both least-squares regression and maximum-likelihood techniques.
- Compare the set of range estimates with the imposed initial range h_0 .

It is noted that for each h_0 , the simulated ground-motion fields are the same throughout the analyses that consider different numbers of stations. Therefore, the number of available stations is the only varying parameter. At the same time, the stations layout does not change throughout the analyses for different h_0 . This means that, for equal number of stations, the correlation structure is the only varying parameter among the different h_0 .

The ground-motion fields are generated on a 150 km \times 150 km grid with a 1 km resolution. We believe this grid dimension represents a good trade-off between computational cost and the risk of boundary effects on the results. As shown in Fig. 3, the grid dimension affects the distance cut-off (i.e. the maximum separation distance in the semivariogram computation). In particular, we observe, as expected, a sharp drop of the number of stations at around 50 km when a grid of 100 km is used. By contrast, the number of stations pairs increases for larger separation distances when a grid of either 150 km or 200 km is chosen. Such maximum usable distance has an impact on the range estimates and their variability, particularly when larger values of h_0 are considered. As a matter of fact, we plot in Fig. 4 the median along with the first and third quartiles values of the range estimates as a function of the grid dimension, for two different estimation approaches. The 150 km grid

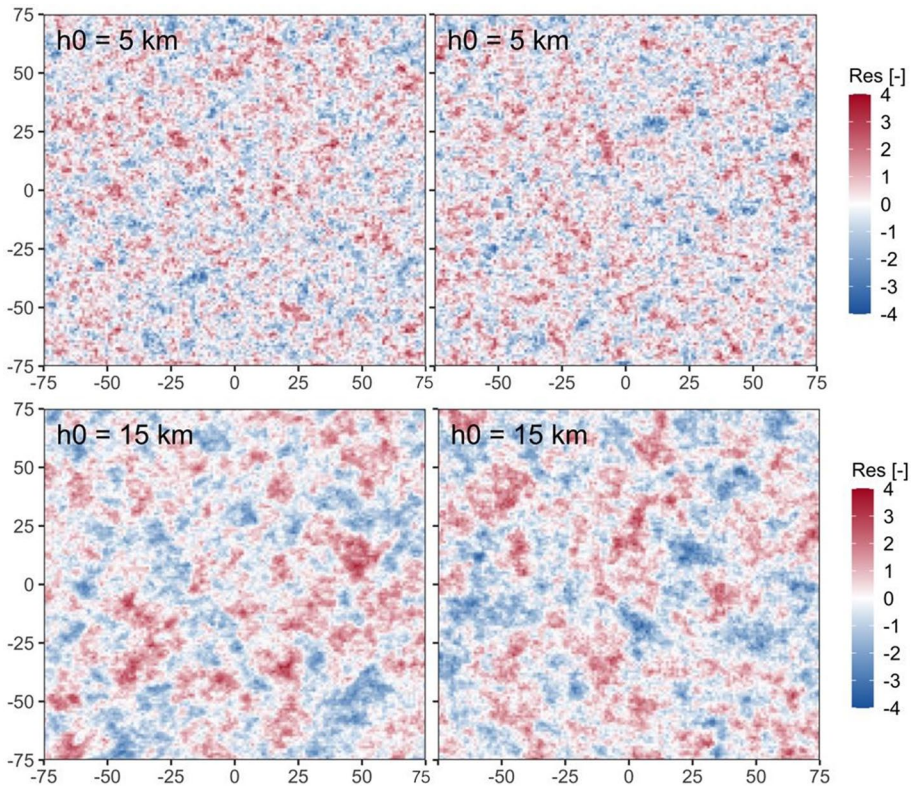


Fig. 2 Examples of Gaussian random fields characterized by a correlation length of 5 km (top row) and 15 km (bottom row)

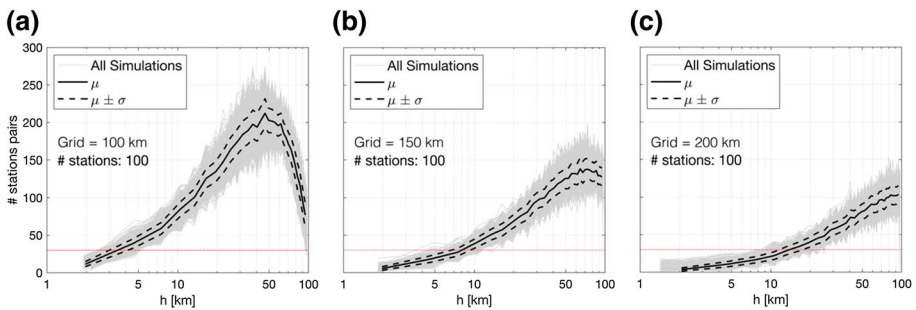
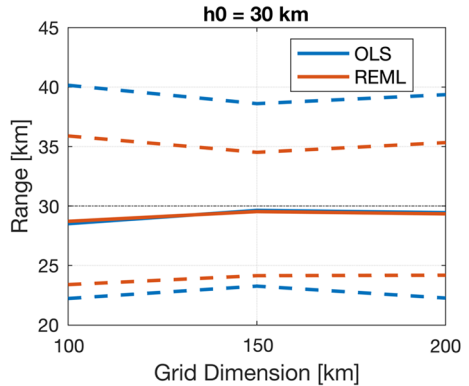


Fig. 3 Number of stations pairs as a function of the separation distance for different grid sizes: **a** 100 km; **b** 150 km; **c** 200 km. The black solid line is the mean value computed over the 1000 simulations, whereas the black dashed lines represent the mean \pm the standard deviation. The red dashed line indicate the minimum number (30) of station pairs required per bin to obtain more robust estimates (see Schiappapietra and Douglas 2020 for further details)

shows the minimum bias with respect to the initial h_0 and the lowest dispersion (as indicated by the first and third quartile) compared to the 100 km grid.

Fig. 4 Range as a function of the grid dimension. Solid lines represent the median values computed over the 1000 simulations, whereas dashed lines indicate the first and third quartile. OLS and REML refer to the two different approaches we use to estimate the range. The black dotted line indicates the initial range value (30 km) imposed in the simulation



4 The effect of bin size

We present here some preliminary analyses performed by varying the bin size from 3 to 10 km. For each h_0 and for an equal number of stations, the bin width is the only varying parameter. This allows the lower range value that is resolvable to be determined and to demonstrate how the bin size and range are interconnected. Figures 5 and 6 show the percentage error and the interquartile range (IQR) computed by:

$$\%error = \frac{\hat{h} - h_0}{h_0} \cdot 100 \tag{7}$$

$$IQR = Q_3 - Q_1 \tag{8}$$

where \hat{h} is the range estimate for each simulated ground motion field and Q_3 and Q_1 are the third and first quartiles, respectively.

Independently of the technique employed to estimate the semivariogram coefficients (in this case, the range), the lower the bin size, the lower the bias and the variability are. Because

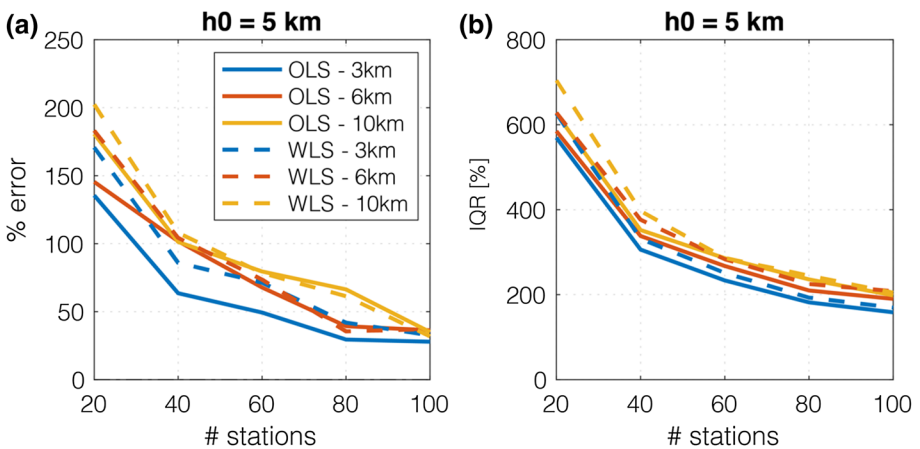


Fig. 5 Bias median (a) and interquartile IQR (b) values as a function of the number of stations for different bin widths and for two different approaches (OLS, WLS). We impose an initial range value of 5 km

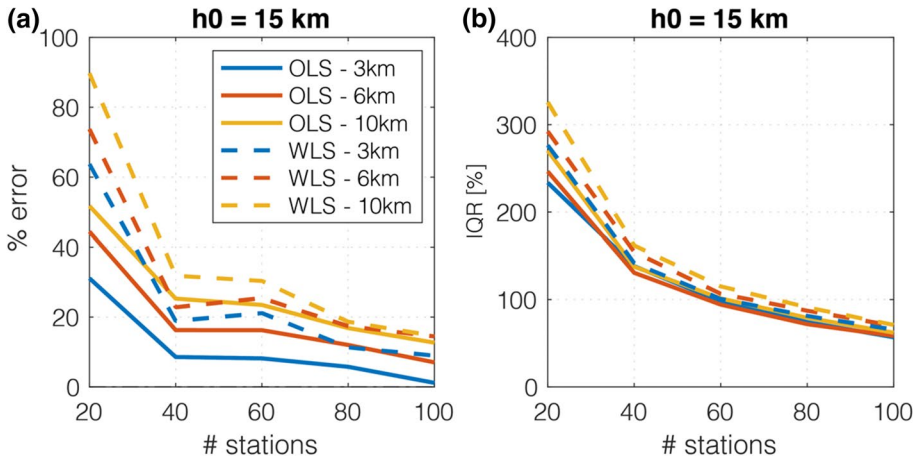


Fig. 6 Bias median (a) and interquartile IQR (b) values as a function of the number of stations for different bin widths and for two different approaches (OLS, WLS). We impose an initial range value of 15 km. Note that the y axis scale is different to that used in Fig. 5

the estimated range value tends to increase with wider bins, we believe that this apparent correlation comes exclusively from the bin size, which thus has a strong impact on the lower range that can be retrieved. Particularly, if the width is too wide, the correlation structure of ground-motion fields correlated over shorter distances may be masked. Consequently, we opt for a bin size of 3 km in the following analyses.

Such a strong dependency on bin size and other choices has motivated us to seek different approaches for the estimation of the correlation coefficients that do not depend on arbitrary parameters like the bin size, the distance cut-off and semivariogram estimator. Hereafter, we present a comparison of the outcomes obtained by means of the different techniques proposed in Sect. 2, namely least-squares regression and the maximum-likelihood approach.

4.1 Least-squares regression versus maximum-likelihood method

We carry out a comparison between range estimates obtained by means of least-squares regression (OLS and WLS) and maximum-likelihood methods (ML and REML). The results of this preliminary analysis are summarised in Fig. 7. It is noted that for each pair of h_0 and number of stations, the bin size is the only varying parameter, so that the comparison is straightforward. Not only do the two maximum-likelihood approaches (ML and REML) provide the lowest bias and variability in the range estimates, but they also return the same outcomes regardless of the bin size. By contrast, OLS and WLS feature increasing median values as the bin becomes wider, as already shown in Figs. 5 and 6. We believe that such results are promising, since ML and REML do not add additional sources of uncertainty related to the choice of the bin width.

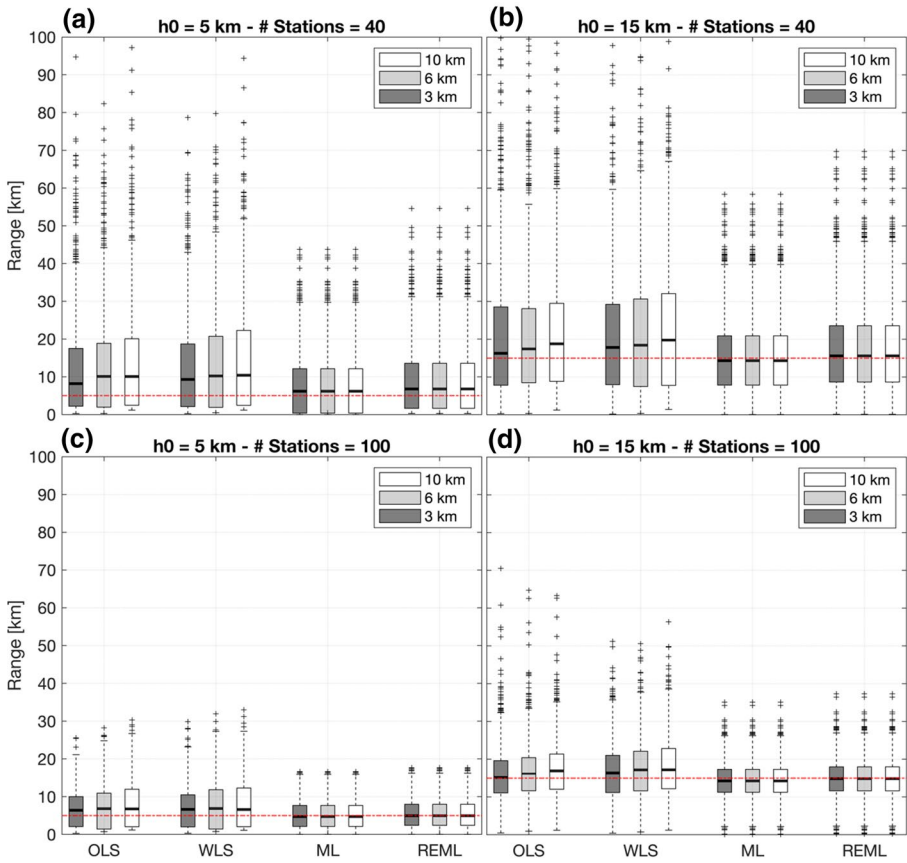


Fig. 7 Boxplots for range estimates obtained by using different bin size, different h_0 (5 and 15 km) and different number of available stations (40 and 100)

5 Dependence on the number of stations

To obtain robust estimates of the correlation structure from ground-motion observations, ideally one would use a large number of closely-spaced data. Although the number of earthquake recordings has dramatically increased over the last decades, seismic stations are not homogeneously distributed, making it difficult to assess the spatial correlation in regions characterized by sparse seismic networks. Baker and Chen (2020) and Schiappapietra and Douglas (2020) already demonstrated that the correlation estimation uncertainty is inversely correlated to the number of available stations. Here, we propose two different studies on the impact of the number of stations that consider both random station locations and station locations based on real networks. The main goal is to illustrate how the maximum-likelihood approaches outperform the least-squares methods especially in terms of estimation uncertainty.

5.1 Randomly simulated locations

We present the results of the simulations performed as described in Sect. 3 for different values of the initial range h_0 [5, 10, 15, 20, 30, 40 km] and different number of stations [20, 40, 60, 80, 100]. Figure 8 (and Figure S1 in the supplementary material) illustrates the median value of the percentage error along with the IQR, which is taken as a measure of the dispersion of the range estimates.

Four main observations can be highlighted. First, independently of h_0 , the maximum-likelihood approaches (ML and REML) generally show a lower dispersion compared to the least-squares methods (OLS and WLS). At the same time, ML and REML provide smaller median % error values, especially when few stations are available. This is a promising result, particularly for those regions characterized by fewer data. Second, median % error values tend to zero as the number of available stations increases. Similarly, the IQR decreases, halving its value as the number of stations rises from 20 to 100. Such outcomes corroborate the findings of Baker and Chen (2020) who demonstrated that the estimation uncertainty is larger for poorly-recorded events and that at least 100 stations are required to provide robust correlation estimates. Third, when a large number of stations is available both least-squares and maximum-likelihood approaches converge towards the same % error and IQR, independently of the h_0 . Fourth, we observe that major differences exist among the proposed estimation techniques for smaller h_0 , compared to the largest ones. We believe that smaller correlations are more difficult to detect as they require a large number of observations at very closely spaced stations. By contrast, earthquake ground motions are often recorded by a limited number of stations separated by many kilometres, with average inter-station distances in the range of 10–20 km, making it easier to measure ground motion correlated over larger distances. As further evidence of such a trend, we plot in Fig. 9 (and Figure S2 in the supplementary material) the % error and the IQR as a function of h_0 for different number of stations considered. It is evident that all the approaches tend to a % error of zero when residuals are correlated over larger distances, independently of the number of stations. Besides, ML and REML provide smaller % error compared to OLS and WLS especially for poorly-recorded events and smaller h_0 . Such behaviour is mainly due to the semivariogram computation, which is required in the least-squares approaches.

Finally, we report in Fig. 10 the boxplots of the range estimates for different numbers of stations and for a given h_0 . Not only do the ML and REML feature narrower boxes (limited by the first and third quartiles, i.e. those used in the computations of IQR) compared to the OLS and WLS boxes, but they are also characterized by smaller whiskers and more confined outliers. The latter are defined as the estimate values that are larger than 1.5 times the IQR. Similar outcomes are obtained also for all other h_0 . For the sake of brevity, we do not show all the figures here.

We recall that for an equal number of stations, the station layout is the same for different h_0 . Therefore, the variability among the different approaches only lies with the different correlation structure used in the simulation.

5.2 Station layouts of past earthquakes

We perform similar Monte Carlo simulations to those presented in Sect. 5.1, but here the ground-motion fields are simulated only at stations that recorded past earthquakes. We take four different stations layouts as references, corresponding to four events

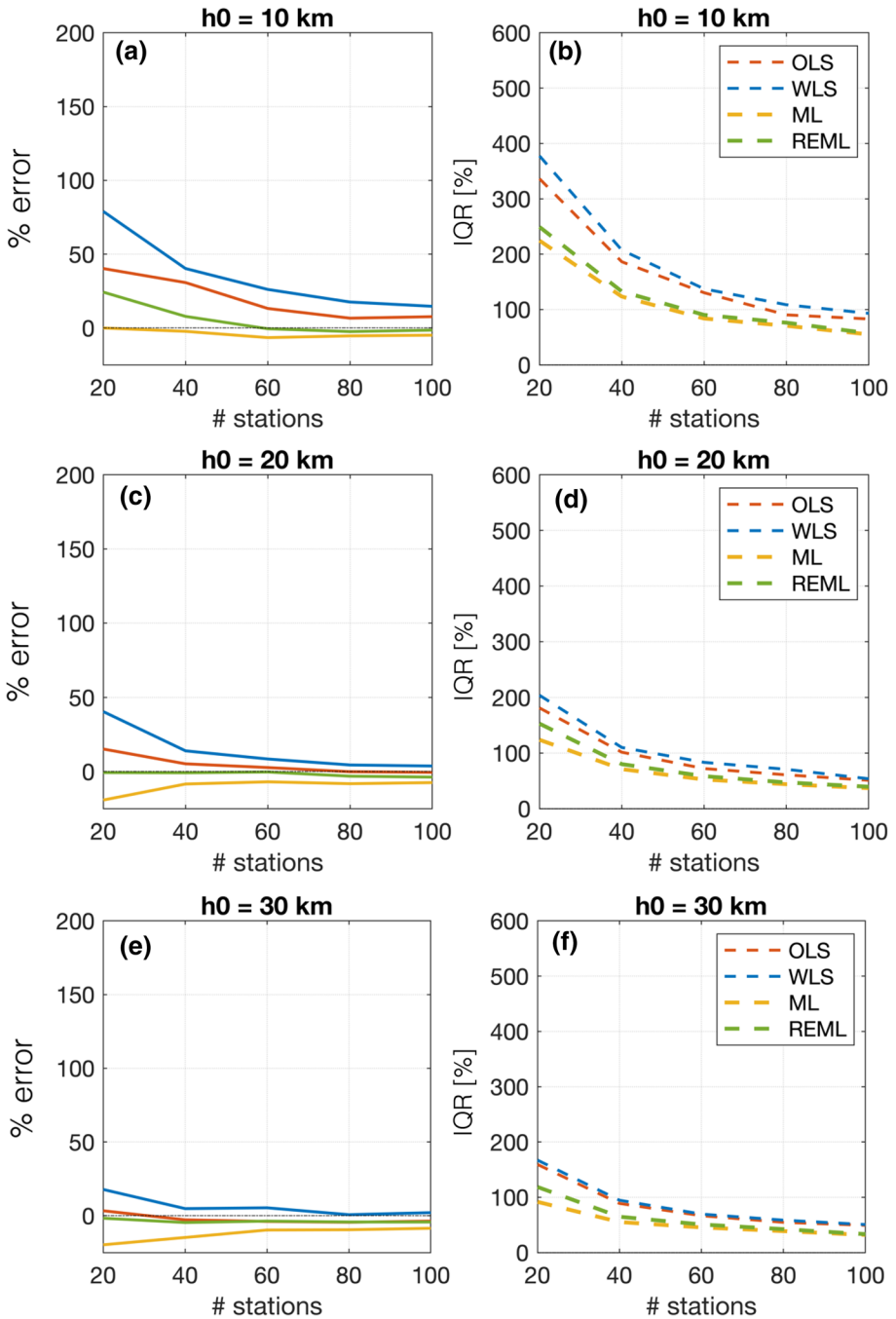


Fig. 8 Bias (left panel) and Interquartile (right panel) range values as a function of the number of stations. The different rows refer to the different initial h_0 values. Different colours refer to the approaches employed in the correlation estimation

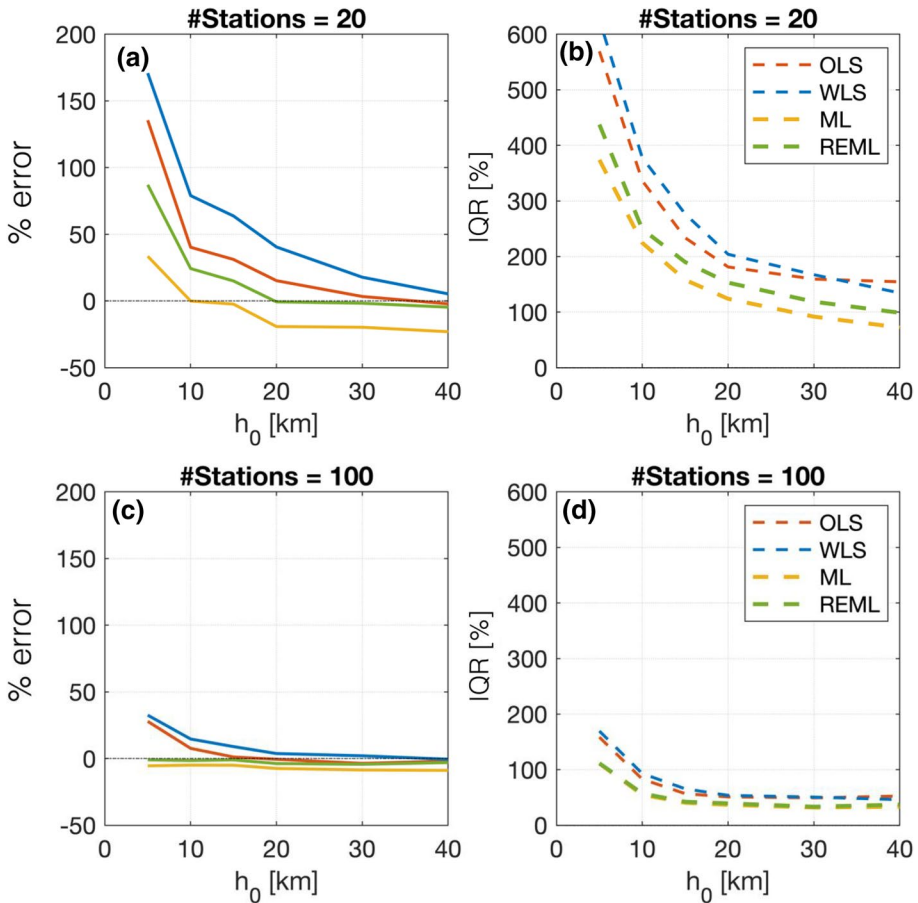


Fig. 9 Bias (left panel) and Interquartile (right panel) range values as a function of h_0 . The different rows refer to the different number of stations considered. Different colours refer to the approaches employed in the correlation estimation

selected within both the ESM (Lanzano et al. 2018) and NGA-West2 (Ancheta et al. 2014) strong-motion flat-files. We consider only well-recorded earthquakes with more than 100 observations within 100 km from the epicentre. Figure 11 show the four station layouts: (1) ESM1 is the M_w 6.0 29th May 2012 Emilia (Italy) event; (2) ESM2 is an M_w 4.3 event that occurred on 23rd September 2016 in Central Italy; (3) NGA1 is the M_w 6.9 13th June 2008 Iwate (Japan) event; NGA2 is a M_w 4.7 event that occurred on 18th May 2009 in California. We believe such layouts are a good sample of the type of station distributions often seen in practice. We adopt a Monte Carlo approach, generating 1000 simulations of ground motion residuals for each h_0 [10, 20, 30, 40] and for different numbers of stations [20, 40, 60, 80, 100], which are randomly selected within each configuration. We note that for each correlation structure (e.g. each h_0), the only varying parameter is the number of selected stations, so that the comparison is not affected by other factors. We also imposed the same seed, which sets the starting number used to generate a sequence of random numbers. Hence, the underlying residuals distribution is

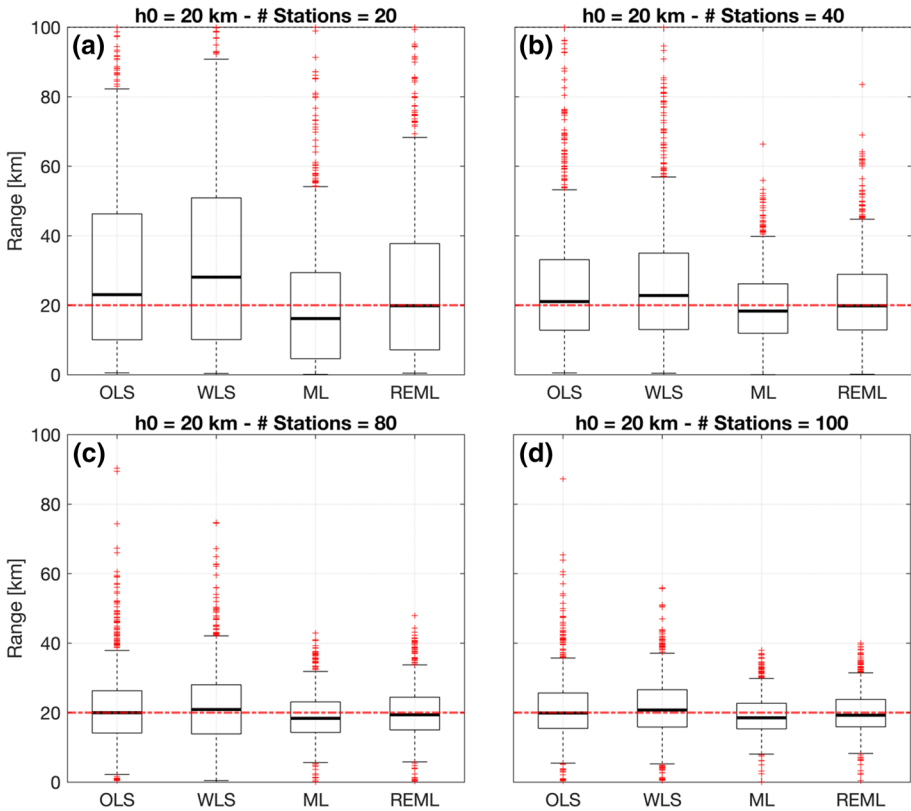


Fig. 10 Boxplots of range estimates for different number of stations and $h_0 = 20$ km: **a** 20 stations; **b** 40 stations; **c** 80 stations; **d** 100 stations

the same, although the station configurations are different in each layout. This allows us to isolate the effect of the station layout.

Figures 12 and 13 compare the median % error and the IQR computed for the different station layouts and $h_0 = 10$ km by using the different methodologies proposed in this study. Three main observations can be highlighted. Generally, ML and REML have the smallest median % error and they lead to lower uncertainties compared to the OLS and WLS approaches, independently of the correlation structure imposed. Such outcomes agree with the findings presented in Sect. 5.1, where residuals are simulated on a fine grid and stations are randomly selected. Similar conclusions can be drawn for the other h_0 values (20, 30, 40 km, Figure S3 and Figure S4 in the supplementary material).

Furthermore, ML and REML feature similar values both in terms of median and variability among the different station layouts. By contrast, OLS and WLS show a strong variability among the four layouts. This applies to all h_0 values, although differences are less pronounced for ground motions correlated over larger distances (i.e. higher h_0). We believe that such behaviour is mainly due to the semivariogram computation, whose robustness depends on both the bin size, as demonstrated in Sect. 4, and the number of stations within each bin. These parameters are strictly related to the station layout so that more homogeneous station distributions would provide more reliable range estimates.

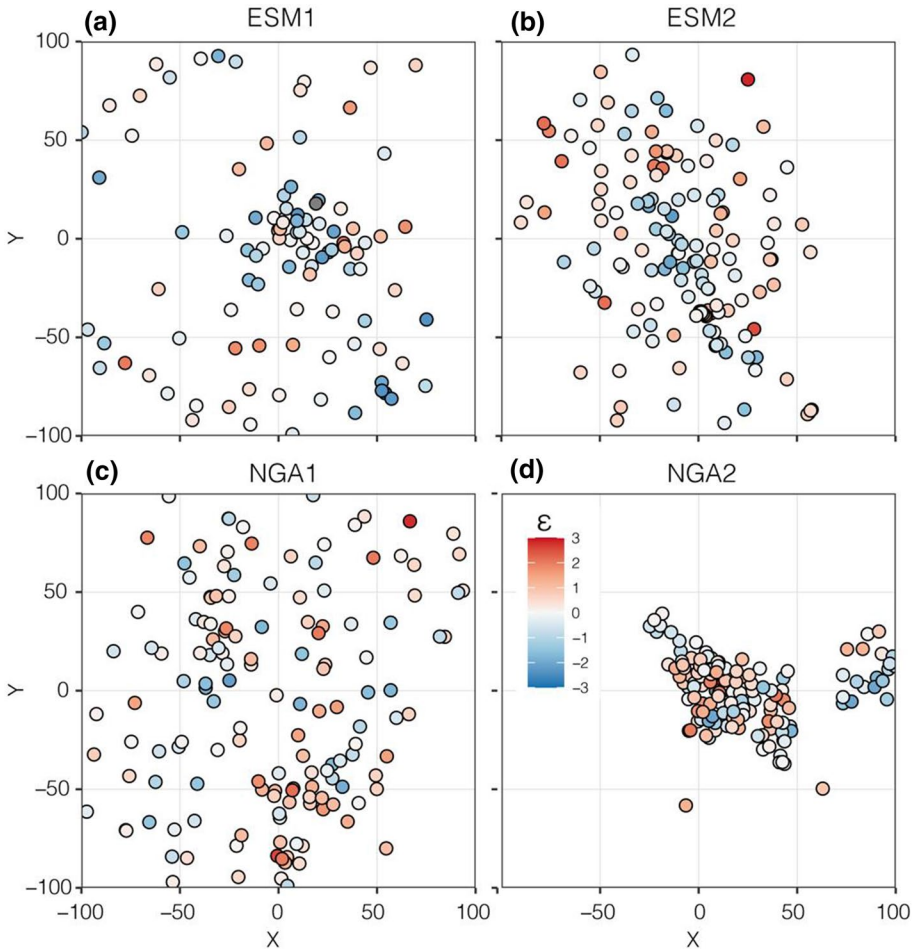


Fig. 11 Station layouts for the four selected events: **a** ESM1—event id 'IT-2012-0011'; **b** ESM2—event id 'ESMC-20160903_0000063'; **c** NGA1—event id '279'; **d** NGA2—event id '1011'. Dots are colour-coded based on the within-event residuals (one out 1000 simulated ground-motion fields). Coordinates are normalized with respect of the epicentre of each event

Finally, we note that generally the NGA2 configuration has the lowest median % error and IQR among all the considered layouts. We believe that this result lies with the more homogeneous and denser distribution of stations with respect to the other three layouts, which leads to more accurate range estimates. To demonstrate this, we plot in Fig. 14 the number of station pairs within 12 km as a function of the number of stations. Systematically, the NGA2 configuration has the largest number of pairs in the first four bins (12 km/3 km=4), independently of the number of available stations. Conversely, the NGA1 layout has the lowest number of pairs and as a consequence it features the highest median % error and IQR, especially when lower h_0 are considered. Such observations apply to the OLS and WLS approaches, whereas the ML and REML are not strongly affected by the different station layouts. This is a promising outcome and demonstrates how maximum-likelihood methods may outperform the least-squares approaches.

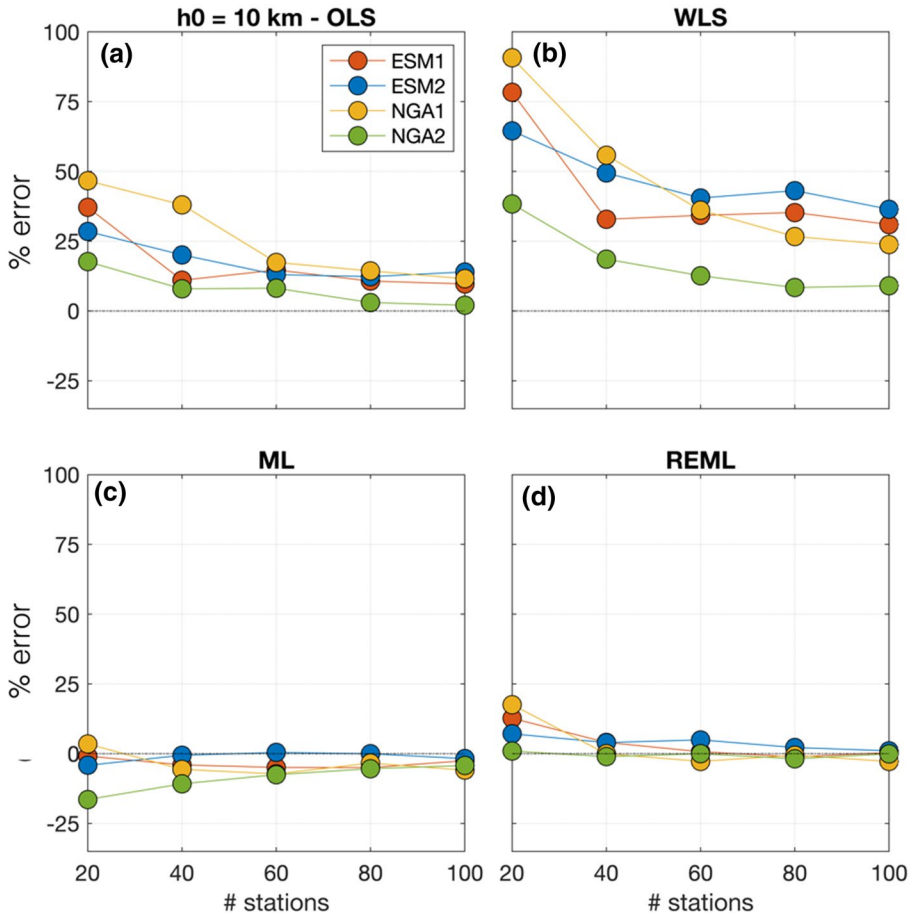


Fig. 12 Median % error of the range estimates for the four different stations layouts as a function of the number of stations. **a** OLS; **b** WLS; **c** ML; **d** REML. The initial value of the range is set to 10 km

6 Discussion

In this study, we show that there is uncertainty in modelling the spatial correlation and the size of this uncertainty depends on the availability of data as well as the derivation technique. Specifically, spatial correlation models for areas with limited data (e.g. regions without dense strong-motion networks and/or low seismicity) are more uncertain than those with extensive observations. Consequently, regional seismic hazard models should account not only for the spatial correlation, but they should also capture its associated uncertainty, which depends on the region.

This regional-dependent uncertainty leads to the following two questions. Which correlation model or models should we use for regions with sparse observations? Is a global model truly able to capture the correlation of that specific area? These are similar questions faced by hazard analysts concerning the selection, modification or development of ground-motion models (e.g. Douglas 2018).

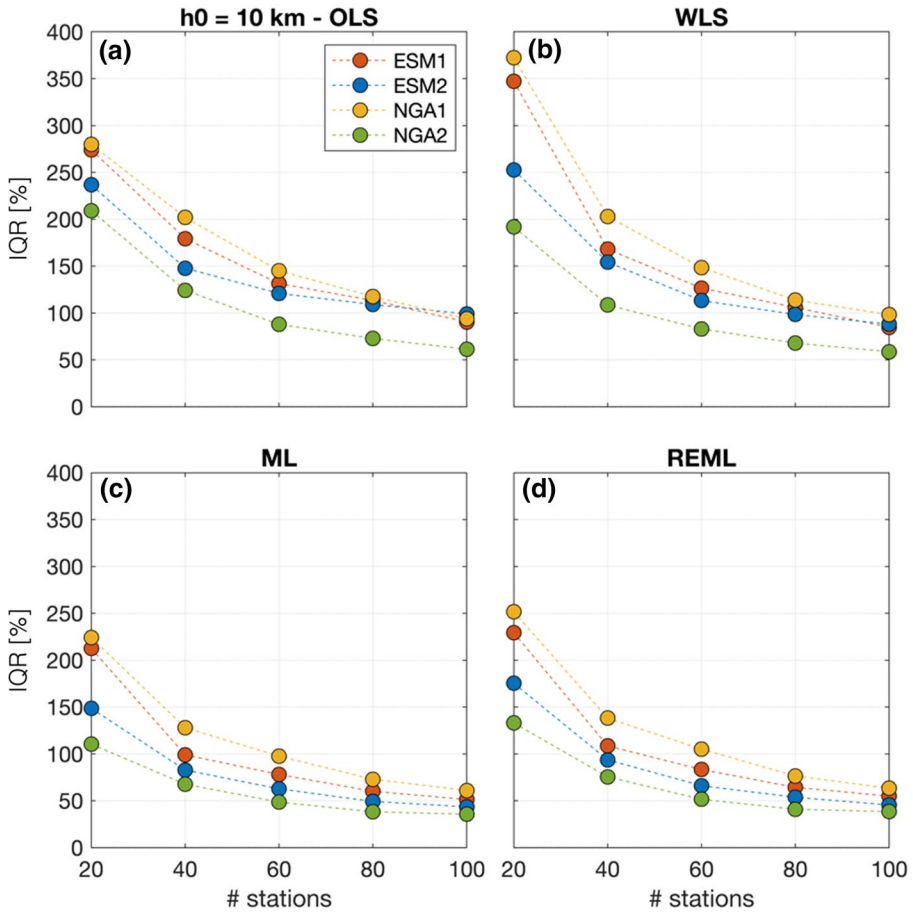


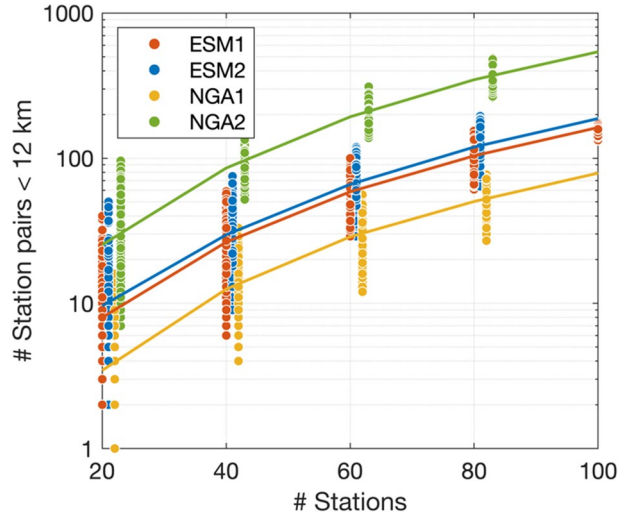
Fig. 13 IQR of the range estimates for the four different station layouts as a function of the number of stations. **a** OLS; **b** WLS; **c** ML; **d** REML. The initial value of the range is set to 10 km

Here, we propose some guidelines to model the spatial correlation uncertainty based on the availability of recordings. In particular, we propose a logic tree with symmetrical lower, middle and upper branches using a standard three-point distribution with weights equal to 0.185, 0.63 and 0.185, respectively (Keefer and Bodily 1983). Table 1 reports the 5%, 50% and 95% percentile values of the range estimates as a function of both the number of stations and h_0 for the different approaches used in this study. This logic tree should capture the spread in the correlation estimates, thus leading to a more informed seismic risk assessment.

We are aware that this table does not cover all possible stations- h_0 combinations, but its aim is to provide a first-order estimate of the spatial correlation uncertainty that one should consider when modelling spatial correlations for both regions where observations are abundant and for those characterized by sparse recordings.

We show here an example in which we compute the spatial correlation for six different earthquakes recorded by 40 (3 events) and 80 (3 events) stations. The events are selected within the ESM (Lanzano et al. 2018) flat-file and we use the OLS and REML

Fig. 14 Number of stations pairs within 12 km as a function of the number of stations for the four station layouts. The dots indicate the number of stations pairs in each simulation, whereas the solid lines represent the average number of pairs as a function of the number of stations



approaches to estimate the range. We note that the scope of this section is to discuss correlation uncertainty and not the optimal method that best estimates the range.

Figure 15 presents the experimental semivariograms and the theoretical models computed on the residuals of three different events recorded by 40 stations. If we look at the REML estimates, the three events feature ranges equal to 41, 15 and 5.5 km. Based on our simulations, such median estimates have the following confidence intervals, given the number of available stations: (1) [15–80] km, (2) [1–40] km, and (3) [0.6–28] km. At the same time, OLS provides median estimates equal to 21, 16.1 and 12.1 km, respectively, which correspond to the following confidence intervals: (1) [1.9–67.2] km, (2) [1.3–64.8] km, and (3) [1.1–60.3] km.

Baker and Chen (2020) provide estimates of the correlation computed for a set of events within the NGA-West2 database (Ancheta et al. 2014) and the corresponding number of stations that recorded each event. For instance, the M_w 6.5 Big Bear-01 (1992) event has a range equal to 15.7 km computed based on 45 observations. According to Table 1 (WLS, $h_0=15$ km, 40 stations), the range estimate is within the [1.1–55.7] km confidence interval. Analogously, the M_w 6.2 Christchurch (2011) event features a range of 18.9 km evaluated on the basis of 80 stations. The corresponding confidence interval from Table 1 is [3.6–41] km. These confidence intervals appear wide but it should be recalled that these are for the 5 to 95th percentiles.

While the median range estimates of the considered events differ from each other, the confidence intervals overlap, suggesting that these earthquakes may, in fact, have similar correlation structures. Similar conclusions can also be drawn from Fig. 16, where we show the results for three different events recorded by 80 stations.

These findings, while preliminary, may help us to answer the second question posed at the beginning of this section about global versus local models. A global model calibrated on data from multiple events may be suitable to describe the correlation structure of a region for which observations are currently sparse since denser datasets would provide more constrained range estimates. On the other hand, pooling data from multiple earthquakes would not enable the study of the correlation due to a particular event to be

Table 1 5%, 50% and 95% percentile values of the range estimates for different h_0 and different number of stations for the approaches used in this study. Estimates are in km

h0	# Stations	OLS			WLS			ML			REML		
		5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
5 km	20	1.3	11.8	131.7	1.2	13.5	91.0	0.2	6.7	46.5	0.9	9.4	70.6
	40	0.8	8.2	39.5	0.8	9.3	35.9	0.1	6.2	24.9	0.6	6.8	27.9
	60	0.6	7.5	26.2	0.6	8.6	26.9	0.1	5.2	18.8	0.5	5.7	21.1
	80	0.6	6.5	19.7	0.5	7.1	20.1	0.1	4.9	13.7	0.4	5.1	14.4
	100	0.6	6.4	16.3	0.5	6.6	16.6	0.1	4.7	12.0	0.4	5.0	12.6
10 km	20	1.6	14.0	319.1	1.3	17.9	154.4	0.2	10.0	48.7	1.0	12.4	65.8
	40	1.1	13.1	60.3	0.8	14.0	50.7	0.2	9.8	28.7	0.7	10.8	32.4
	60	0.8	11.3	34.3	0.7	12.6	33.1	0.2	9.3	21.6	0.8	9.9	23.5
	80	0.7	10.7	24.2	0.7	11.8	26.1	0.2	9.5	18.1	0.8	9.8	19.2
	100	0.8	10.8	22.6	0.8	11.5	24.0	2.5	9.5	17.1	2.5	9.9	17.8
15 km	20	1.6	19.7	196.5	1.4	24.6	121.6	0.3	14.6	56.5	1.1	17.2	80.4
	40	1.3	16.3	64.8	1.1	17.8	55.7	0.2	14.3	34.3	1.0	15.6	39.6
	60	1.1	16.2	41.0	0.9	18.2	41.3	1.9	14.4	28.0	2.2	15.2	30.9
	80	4.1	15.9	33.2	1.4	16.7	33.5	5.3	14.2	25.1	5.8	15.0	26.7
	100	5.2	15.2	29.0	3.4	16.3	29.1	7.2	14.3	22.5	7.4	14.8	23.9
20 km	20	2.1	23.0	175.3	1.6	28.1	110.3	0.3	16.2	53.8	1.5	19.9	74.9
	40	1.9	21.0	67.2	1.4	22.8	59.0	0.4	18.3	38.5	1.5	19.8	43.9
	60	4.9	20.5	49.2	1.7	21.7	47.3	7.2	18.6	33.7	7.7	19.9	37.6
	80	6.3	20.0	41.4	3.6	20.9	41.0	8.3	18.4	30.9	8.4	19.4	33.1
	100	9.1	19.9	36.2	6.0	20.8	36.3	11.1	18.5	28.8	11.4	19.3	30.4
30 km	20	2.7	31.0	340.9	1.7	35.3	148.3	0.4	24.1	67.4	1.8	29.5	103.3
	40	5.3	29.2	93.6	1.6	31.4	79.9	5.6	25.6	52.7	6.5	28.6	61.6
	60	10.8	28.8	70.6	7.9	31.6	66.7	11.9	27.1	47.2	12.3	28.9	53.1
	80	13.2	28.7	59.5	11.4	30.2	55.7	15.6	27.1	44.2	16.5	28.7	48.2
	100	15.4	28.9	53.8	13.4	30.6	53.7	16.7	27.4	41.6	17.5	28.7	45.1
40 km	20	3.6	39.1	578.8	2.4	42.1	207.2	0.5	30.8	78.8	2.2	38.1	123.7
	40	10.2	39.2	144.2	3.1	42.3	104.7	12.8	34.6	65.4	15.1	39.2	79.9
	60	16.7	39.2	105.4	14.7	40.9	84.9	18.9	34.4	58.8	20.1	37.6	67.2
	80	19.8	38.7	94.2	18.4	40.6	77.7	20.9	35.8	58.8	21.9	38.3	65.6
	100	21.8	39.3	80.4	20.3	39.8	72.4	24.1	36.5	57.5	25.1	38.8	63.4

investigated and therefore it would increase the uncertainty in the estimates related to the specificity of the region of interest.

While investigating the estimation uncertainty related to a single event, in this study we have not explored multiple earthquakes uncertainties. Heresi and Miranda (2019) computed the central tendency and the dispersion of the range values from well-recorded NGA-West2 events. For instance, the spectral acceleration at $T=0.1s$ shows an average range of 14.3 km and a standard deviation computed on the natural logarithm of the range values equal to 0.83. We performed similar analysis for the Central Italy region, and we obtained an average range of 27.8 km and a standard deviation equal to 0.75 for the same IM. While these studies account for the event-to-event dispersion, they do not investigate the estimation uncertainty. Therefore, further work is required

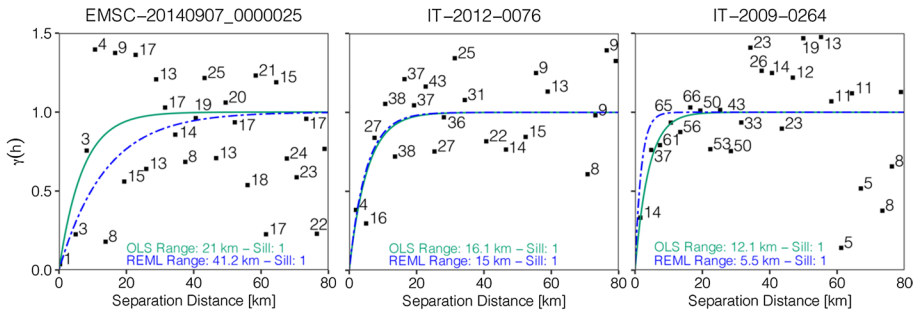


Fig. 15 Experimental semivariograms and theoretical models for three different events recorded by 40 stations. The number of pairs in each bin is reported close to each semivariance estimate (black dots)

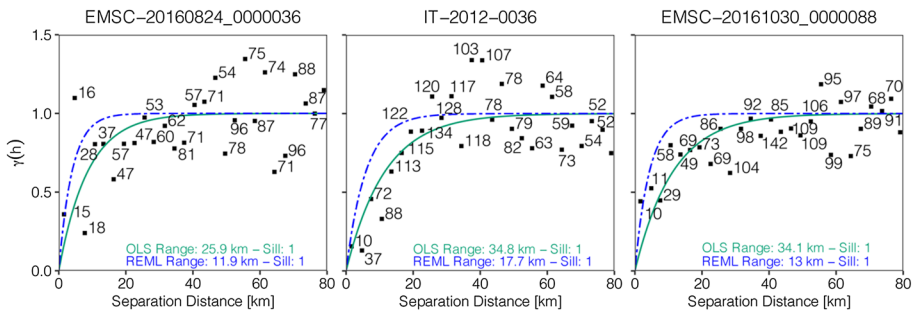


Fig. 16 Experimental semivariograms and theoretical models for three different events recorded by 80 stations. The number of pairs in each bin is reported close to each semivariance estimate (black dots)

to establish the fraction of uncertainty which derives from pooling data from different events.

Finally, we report here a scheme showing the application of the proposed logic tree for a median $h_0 = 20$ km. We consider the estimates obtained by using the REML approach when 60 stations are available. Based on the correlation estimates in Table 1, we simulate n spatially correlated random fields for each range to use in conjunction with predicted median ground motions. The different branches (after assessing the risk, e.g. after estimating the corresponding losses for each range) would then be averaged with weights equal to 0.185, 0.63 and 0.185, respectively, to obtain the final risk assessment (e.g. resulting losses) which account for the uncertainties in the correlation estimates (Fig. 17).

7 Conclusions

In this work, we introduce alternative methods to classic least-squares regression for the estimation of the correlation structure of earthquake ground motions. In particular, we employ two maximum likelihood-based approaches, namely the Gaussian maximum likelihood (ML) and the restricted maximum likelihood (REML). These have gained increasing importance in geostatistics, particularly when spatial trends exist in the data (Diggle and Ribeiro 2007; Oliver and Webster 2014; Li et al. 2018). However, estimation methods

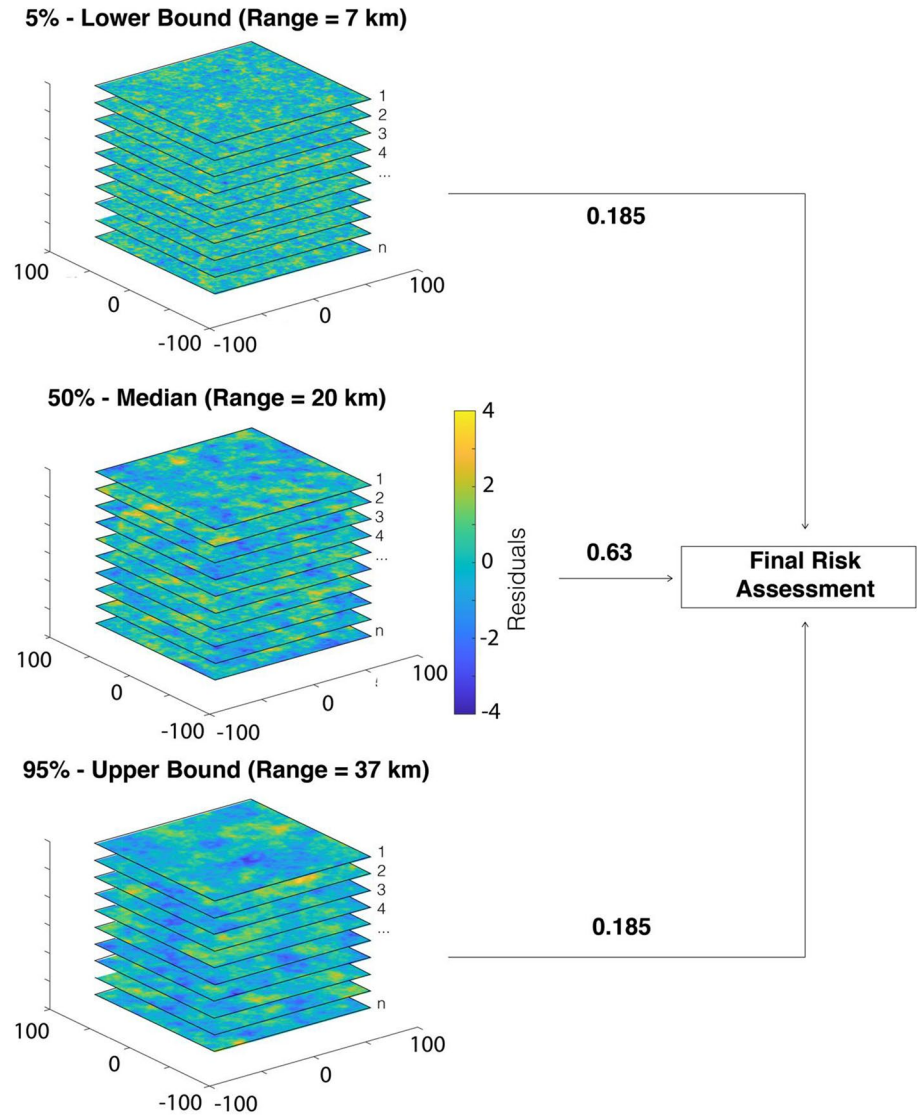


Fig. 17 Application of the proposed logic tree for $h_0=20$ km and number of stations equal to 60 to use within a seismic risk assessment study. We use the estimates obtained by using the REML approach. The graphs on the left-hand side correspond to n spatially correlated random fields simulated for each range value, 7 km, 20 km and 37 km, respectively

based on the likelihood function have not been commonly used to assess ground-motion correlation. One of the main advantages of such approaches is that they are straightforward and do not require further steps for the estimation of the correlation parameters.

We first showed that ML and REML estimates do not depend on the bin size, unlike ordinary and weighted least-squares regression (OLS and WLS). Indeed, the latter requires the definition of the experimental semivariogram, whose robustness depends on

both the bin width and the number of stations within each bin. Our outcomes are promising as they show how ML and REML may outperform the least-squares approaches. There is a trade-off between the bin width and the estimate robustness: wider bins include a larger number of residuals pairs, which increases the robustness of the semi-variance estimates, but at the same time wider bins may mask shorter correlation lengths.

We then performed two different studies to show the dependence of the correlation on the number of available stations and on the station layout. Firstly, we carried out simulations of within-event residuals on a fine grid, varying both the h_0 and the number of stations available. Generally, ML and REML feature lower percentage errors and dispersion compared to OLS and WLS, independently of the number of stations and of the underlying spatial correlation structure (h_0). This is a rather interesting result, especially for those regions characterised by sparse strong-motion networks. Second, we carried out simulations of within-event residuals only at recording stations of past earthquakes. We chose four different station layouts, which are considered as good examples of the type of station distributions often seen in practice. Our outcomes suggest that OLS and WLS are more affected by the station configuration because their estimates are based on the computation of the experimental semivariogram. Thus, more homogeneous station layouts would provide more reliable range estimates. By contrast, ML and REML seem to be less influenced by the station layout both in terms of median percentage error and interquartile range.

This article intends to be a continuation of the work of Schiappapietra and Douglas (2020), as it further analyses the dependency of correlation on different factors such as the bin size and the station configuration. We shed light on alternative approaches to characterize the spatial correlation structure of earthquake ground motions, providing useful insights for users and researchers interested in investigating ground-motion spatial correlation.

Finally, we proposed some guidelines to model the spatial correlation uncertainty based on the availability of recordings, following a logic-tree approach. The main idea is to provide hazard and risk assessment by using the median range estimate and the lower and upper bounds range values (5% and 95% confidence intervals). The resulting risk analysis accounting for the correlation uncertainties is eventually obtained by averaging all the branches with suitable weights (0.185, 0.63 and 0.185, respectively).

These findings, while preliminary, may help researchers to model the spatial correlation uncertainty that one should consider when performing regional seismic hazard and risk assessment. Application of the proposed logic-tree to a specific case study may indicate features for further developments.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10518-021-01179-w>.

Acknowledgements We thank Dr Stella Pytharouli and Prof. Jack Baker for helpful feedback and advice at early stages of this research. We also thank two anonymous reviewers for their comments on a previous version of this study. This work has been supported by the University of Strathclyde (Ph.D. of first author).

Authors' contribution E. Schiappapietra: conceptualization, methodology, software, formal analysis, investigation, resources, data curation, Writing—original draft, Writing—review and editing, visualization. J. Douglas: supervision, writing—review and editing.

Funding University of Strathclyde (Ph.D. of Erika Schiappapietra).

Availability of data and materials The ESM (Lanzano et al. 2018) strong-motion flat-file is available at <https://esm.mi.ingv.it/flatfile-2018/>, whereas the NGA-West2 flat-file is available at <https://peer.berkeley.edu/research/data-sciences/databases>.

Declarations

Conflict of interest The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ancheta TD, Darragh RB, Stewart JP, Seyhan E, Silva WJ, Chiou BSJ, Donahue JL (2014) NGA-West2 database. *Earthq Spectra* 30(3):989–1005
- Baker JW, Chen Y (2020) Ground motion spatial correlation fitting methods and estimation uncertainty. *Earthq Eng Struct Dyn* 49:1662–1681
- Baker JW, Jayaram N (2008) Correlation of spectral acceleration values from NGA ground motion models. *Earthq Spectra* 24(1):299–317. <https://doi.org/10.1193/1.2857544>
- Chen Y, Baker J (2019) 'Spatial correlations in cybershake physics-based ground motion simulations. *Bull Seismol Soc Am*. <https://doi.org/10.1785/0120190065>
- Cressie N (1985) Fitting variogram models by weighted least squares. *J Int Assoc Math Geol* 17(5):563–586. <https://doi.org/10.1007/BF01032109>
- Diggle PJ, Ribeiro PJ (2007) *Model-based geostatistics*. 1st edn, Springer Series in Statistics. Springer, New York, NY. <https://doi.org/10.1007/978-0-387-48536-2>.
- Douglas J (2018) Capturing geographically-varying uncertainty in earthquake ground motion models or what we think we know may change. In: European conference on earthquake engineering Thessaloniki, Greece (pp 153–181). Springer, Cham
- Douglas J, Edwards B (2016) 'Recent and future developments in earthquake ground motion estimation. *Earth-Sci Rev* 160:203–219. <https://doi.org/10.1016/j.earscirev.2016.07.005>
- Esposito S, Iervolino I (2012) Spatial correlation of spectral acceleration in European data. *Bull Seismol Soc Am* 102(6):2781–2788. <https://doi.org/10.1785/0120120068>
- Goda K (2011) Interevent variability of spatial correlation of peak ground motions and response spectra. *Bull Seismol Soc Am* 101(5):2522–2531. <https://doi.org/10.1785/0120110092>
- Heresi P, Miranda E (2019) 'Uncertainty in intraevent spatial correlation of elastic pseudo-acceleration spectral ordinates. *Bull Earthq Eng* 17(3):1099–1115. <https://doi.org/10.1007/s10518-018-0506-6>
- Iervolino I (2013) Probabilities and fallacies: why hazard maps cannot be validated by individual earthquakes. *Earthq Spectra*. <https://doi.org/10.1193/1.4000152>.
- Infantino M, Smerzini C, Lin J (2021) Spatial correlation of broadband ground motions from physics-based numerical simulations. *Earthq Eng Struct Dyn*. <https://doi.org/10.1002/eqe.3461>
- Jayaram N, Baker JW (2009) Correlation model for spatially distributed ground-motion intensities. *Earthq Eng Struct Dyn*. <https://doi.org/10.1002/eqe>
- Huang C, Tarbali K, Galasso C, Paolucci R (2020) Spatial correlation validation of 3D physics-based ground-motion simulations. In: 17th world conference in earthquake engineering (17WCEE), Japan
- Keefer DL, Bodily SE (1983) Three-point approximations for continuous random variables. *Manag Sci* 29(5):595–609. <https://doi.org/10.1287/mnsc.29.5.595>
- Künsch H, Papritz AJ, Schwier C, Stahel WA (2013) Robust estimation of the external drift and the variogram of spatial data. In: ISI 58th World Statistics Congress of the International Statistical Institute. Eidgenössische Technische Hochschule Zürich
- Lanzano G, Puglia R, Russo E, Luzi L, Bindi D, Cotton F, D'Amico M, Felicetta C, Pacor F, ORFEUS WG5 (2018) ESM strong-motion flat-file 2018. Istituto Nazionale di Geofisica e Vulcanologia (INGV), Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum (GFZ), Observatories & Research Facilities for European Seismology (ORFEUS). PID: 11099/ESM_flatfile_2018

- Li Z et al (2018) An automatic variogram modeling method with high reliability fitness and estimates. *Comput Geosci* 120:48–59. <https://doi.org/10.1016/j.cageo.2018.07.011>
- Matheron G (1962) *Traité de géostatistique appliquée 1* (1962) Editions Technip. Cited in Esposito and Iervolino (2012)
- Ming D et al (2019) An advanced estimation algorithm for ground-motion models with spatial correlation. *Bull Seismol Soc Am* 109(2):541–566. <https://doi.org/10.1785/0120180215>
- Oliver MA, Webster R (2014) A tutorial guide to geostatistics: computing and modelling variograms and kriging. *CATENA* 113:56–69. <https://doi.org/10.1016/j.catena.2013.09.006>
- Papritz A (2020) georob: robust geostatistical analysis of spatial data. R package version 0.3–13. <https://CRAN.R-project.org/package=georob>
- Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. *Comput Geosci* 30:683–691
- R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Ribeiro PJ, Diggle PJ, Schlather M, Bivand R, Ripley B (2020) geoR: analysis of geostatistical data. R package version 1.8-1. <https://CRAN.R-project.org/package=geoR>
- Schiappapietra E, Douglas J (2020) Modelling the spatial correlation of earthquake ground motion: Insights from the literature, data from the 2016–2017 Central Italy earthquake sequence and ground-motion simulations. *Earth Sci Rev*. <https://doi.org/10.1016/j.earscirev.2020.103139>
- Schiappapietra E, Smerzini C (2021) Spatial correlation of broadband earthquake ground motion in Norcia (Central Italy) from physics-based simulations. *Bull Earthq Eng*. <https://doi.org/10.1007/s10518-021-01160-7>
- Schwanghart W (2021a) Experimental (Semi-) Variogram, MATLAB Central File Exchange. Retrieved May 13, 2021. (<https://www.mathworks.com/matlabcentral/fileexchange/20355-experimental-semi-variogram>)
- Schwanghart W (2021b) variogramfit, MATLAB Central File Exchange. Retrieved May 13, 2021. (<https://www.mathworks.com/matlabcentral/fileexchange/25948-variogramfit>)
- Sokolov V, Ismail-Zadeh A (2016) On the use of multiple-site estimations in probabilistic seismic-hazard assessment. *Bull Seismol Soc Am* 106(5):2233–2243. <https://doi.org/10.1785/0120150306>
- Sokolov V, Wenzel F (2019) ‘Areal exceedance of ground motion as a characteristic of multiple-site seismic hazard: sensitivity analysis. *Soil Dyn Earthq Eng* 126(June):105770. <https://doi.org/10.1016/j.soildyn.2019.105770>
- Stafford PJ et al (2018) ‘Extensions to the Groningen ground-motion model for seismic risk calculations: component-to-component variability and spatial correlation. *Bull Earthq Eng* 17(8):4417–4439. <https://doi.org/10.1007/s10518-018-0425-6>
- Strasser FO, Abrahamson N, Bommer J (2009) Sigma: issues, insights, and challenges. *Seismol Res Lett* 80(1):40–56. <https://doi.org/10.1785/gssrl.80.1.40>
- Weatherill GA et al (2015) ‘Exploring the impact of spatial correlations and uncertainties for portfolio analysis in probabilistic seismic loss estimation. *Bull Earthq Eng* 13(4):957–981. <https://doi.org/10.1007/s10518-015-9730-5>

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.