FINAL TECHNICAL REPORT

USGS AWARD G10AP00046

SPATIAL CROSS-CORRELATION OF SPECTRAL ACCELERATIONS AT MULTIPLE PERIODS: MODEL DEVELOPMENT AND RISK ASSESSMENTS CONSIDERING SECONDARY EARTHQUAKE EFFECTS

PI: Jack W. Baker
Report co-author: Christophe Loth
Dept. of Civil & Environmental Engineering
Yang & Yamasaki Environment & Energy Building
473 Via Ortega, Room 283
Stanford CA 94305-4020
650-725-2573 (phone)
650-723-7514 (fax)
bakerjw@stanford.edu

August 2011

Research supported by the U. S. Geological Survey (USGS), Department of Interior, under USGS award number G10AP00046. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U. S. Government.

Characterizing spatial cross-correlation between ground motion spectral accelerations at multiple periods

By Christophe Loth and Jack W. Baker

Abstract

Many seismic loss problems (such as damage of distributed infrastructure and losses to portfolios of structures) are dependent upon the regional distribution of ground motion intensity, rather than intensity at only a single site. Quantifying ground-motion shaking over a spatially-distributed region therefore requires information on the correlation between the ground-motion intensities at different sites during a single event. The focus of the present study is to assess the spatial correlation between ground motion spectral accelerations at different periods. Ground motions from eight well-recorded earthquakes were used to study the spatial correlations. Based on obtained empirical correlation estimates, the authors propose a geostatistics-based method to formulate a predictive model that is suitable for simulation of spectral accelerations at multiple sites and multiple periods. While the calibration of this model and investigation of its implications were somewhat complex, it should be emphasized that the model is very simple to use for making correlation predictions. A user of this model only needs to evaluate Equation (42), with the needed coefficients from Table 4 and Table 5, to compute a correlation coefficient for spectral values at two periods at a specified separation distance. These results may then be used in evaluating the seismic risk of portfolios of structures with different fundamental periods.

Introduction

Quantifying ground-motion shaking over a spatially-distributed region rather than at just a single site is of interest for a variety of applications relating to risk of infrastructure or portfolios of properties. This requires information on the correlation between the ground-motion intensities at different sites during a single event. Researchers have previously estimated the correlations between residuals of spectral accelerations at the same spectral period at two different sites. But

there is little knowledge about cross-correlations between residuals of spectral accelerations at different periods (or more generally between residuals of two different intensity measures) at two different sites, which becomes important, for instance, when assessing the risk of a portfolio of buildings with different fundamental periods.

This research relies on the general framework of ground-motion models (e.g., Boore and Atkinson 2008, Abrahamson and Silva 2008, Chiou and Youngs 2008, Campbell and Bozorgnia 2008), that are defined as follows: for an earthquake j at a site i,

$$\ln Y_{ij} = \ln \overline{Y}_{ij} + \sigma_{ij}\varepsilon_{ij} + \tau_j \eta_j \tag{1}$$

where Y_{ij} refers to the ground-motion parameter of interest (e.g. $S_a(T)$ the spectral acceleration at period T); $\overline{Y}_{ij}$ denotes the predicted (by the ground-motion model) median ground-motion intensity, a function of various parameters such as magnitude, distance, period and local-site conditions; ε_{ij} refers to the intra-event residual, a random variable of mean zero and unit standard deviation; and η_j denotes the inter-event residual, also a random variable of mean zero and unit standard deviation. The standard deviations σ_{ij} and τ_j are included in the ground-motion model prediction and depend on the spectral period of interest (in some models, they are also a function of the earthquake magnitude and the distance of the site from the rupture). For a given earthquake j, the inter-event residual η_j computed at any particular period is a constant across all the sites.

Previous studies have established that a vector of spatially distributed intra-event residuals $\boldsymbol{\varepsilon}_j = \left(\varepsilon_{1j}, \varepsilon_{2j}, ..., \varepsilon_{nj}\right)$ follows a multivariate normal distribution (Jayaram and Baker 2008). Consequently, one can fully define the $\boldsymbol{\varepsilon}_j$ by specifying their mean vector and the covariance between all considered pairs. In our particular case, the mean vector of $\boldsymbol{\varepsilon}_j$ is 0 and hence we only need to know the variance-covariance matrix: for a given earthquake j,

$$\boldsymbol{\Sigma}(event\ j) = \begin{bmatrix} \operatorname{cov}\left(\varepsilon_{1j}, \varepsilon_{1j}\right) & \dots & \operatorname{cov}\left(\varepsilon_{1j}, \varepsilon_{nj}\right) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}\left(\varepsilon_{nj}, \varepsilon_{1j}\right) & \cdots & \operatorname{cov}\left(\varepsilon_{nj}, \varepsilon_{nj}\right) \end{bmatrix} \tag{2}$$

where Σ_{kl} is the covariance between ε_{kj}, the residual at site k due to earthquake j, and ε_{lj}, the residual at site l due to earthquake j.

Spatial modeling of earthquake intensities has been investigated in the past by various researchers. For instance, the recent work of Foulser-Piggott and Stafford (2011) aims at modeling the spatial correlation for $Z = I_a$, the Arias intensity. The modeling of the spatial correlation of the residuals of a single spectral acceleration period $Z = S_a(T)$ has also been addressed in previous contributions (e.g., Goda and Hong 2008, Jayaram and Baker 2009, Wang and Takada 2005, Boore et al. 2003). Jayaram and Baker (2009) formulated a predictive equation of the correlation coefficient as a function of the period of interest and the separation distance between two considered sites. The present work will generalize the modeling to a multivariate framework that accounts for several intensity measures $\{Z_i = S_a(T_i) \mid i = 1...n\}$, where less study has been done.

This study will begin with a presentation of geostatistical concepts relevant to the spatial modeling of correlations. The next section will describe the authors' first attempt at using empirical data to estimate these correlations, along with the encountered limitations and issues. A technique also borrowed from geostatistics will then be introduced as an improved solution, and predictive models for covariance and correlation will be derived using it. Finally, as Goda and Hong (2008) proposed to use the single period result combined with a Markov-type hypothesis to formulate a correlation model for the multi-period case, the authors will also discuss the influence of such hypothesis in the last section, in comparison with the formulated model.

Geostatistical modeling of correlations

Spatially distributed random variables are often described by a variogram, which is a very popular tool in the geostatistics domain (Journel and Huijbregts 1978, Goovaerts 1997). The variogram is a so-called two-point statistic that characterizes the spatial decorrelation or dissimilarity. Its general formulation is given below for a pair of locations $\mathbf{u}, \mathbf{u}'$:

$$\gamma(\mathbf{u}, \mathbf{u}') = \frac{1}{2} E\left[\left(Z(\mathbf{u}) - Z(\mathbf{u}') \right)^2 \right] \quad (3)$$

where $Z(\mathbf{u})$ is a random variable representing the value of interest at location $\mathbf{u}$, $E[\ ]$ denotes the expectation, and $\gamma(\mathbf{u},\mathbf{u}')$ is the variogram value.

Since one often does not possess several observations of a random variable at a given pair of sites, the assumption of stationarity has to be made in order to evaluate Equation (3): one will typically retain that the variogram does not depend on the site locations $(\mathbf{u},\mathbf{u}')$ but only on their separation vector $\mathbf{h} = \mathbf{u} - \mathbf{u}'$. Thus, for a stationary random variable Z, for instance $Z = \varepsilon(T)$, the variogram is defined as follows:

$$\gamma(\mathbf{h}) = \frac{1}{2} E\left[\left(Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u})\right)^2\right] \tag{4}$$

where $\mathbf{h}$ represents a given separation vector. This variogram function can be empirically estimated with:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left[Z(\mathbf{u}_\alpha + \mathbf{h}) - Z(\mathbf{u}_\alpha)\right]^2 \tag{5}$$

where $\hat{}$ denotes an empirical value, $\mathbf{u}_\alpha$ a recording location from the data, and $N(\mathbf{h})$ the number of pairs at separation vector $\mathbf{h}$ available in the data.

Previous studies have indicated that the correlation structure of residuals from ground motion models was not dependent on the considered direction, and was therefore isotropic (Jayaram and Baker 2009; Bazzurro and Luco 2004; Wang and Takada 2005; Goda and Hong 2008). This translates in Equations (4) and (5) by simply "removing" all vector notations, so that $h = \|\mathbf{h}\|$. It should also be noted that in the practical computation of the variogram with Equation (5), it is unlikely that two data points will be separated by the exact distance h. Therefore, a tolerance parameter Δ will have to be considered such that for a given lag distance h, all the pairs of points separated by a distance included in the interval $[h-\Delta, h+\Delta]$ will contribute to the evaluation of the empirical variogram $\hat{\gamma}(h)$. Figure 1 shows the isotropic variogram function computed for data from the Northridge earthquake.

Furthermore, the covariance function can be defined as:

$$C(h) = \mathrm{cov}\left(Z(u), Z(u+h)\right) = E\left[\left(Z(u)-m\right)\left(Z(u+h)-m\right)\right] \quad (6)$$

where m is the mean of $Z(u)$ (and is also equal to the mean of $Z(u+h)$ under the stationarity hypothesis). This spatial covariance is directly related to the variogram function with:

$$C(h) = C(0) - \gamma(h) \quad (7)$$

Similarly, it can be noted that the correlation coefficient is defined as:

$$\rho(h) = \frac{C(h)}{C(0)} \quad (8)$$

Thus, variogram and covariance have "opposite" behaviors: the covariance is a measure of spatial similarity between $Z(u)$ and $Z(u+h)$. While one could conduct a covariance study on either one of those functions, the variogram is often preferred in geostatistical practice, as it does not require a prior estimation of the mean of the random field m.

In this research, the authors consider the cross-covariance structure of residuals of spectral acceleration at multiple periods. This means that one needs to extend the previous definitions to the multivariate case in order to estimate all spatial cross-correlation terms between $\varepsilon(T_i)$ and $\varepsilon(T_j)$, $T_i \neq T_j$. First, the definition of the variogram can easily be generalized to the multivariate case. Denoting two stationary random variables $Z_1 = \varepsilon(T_1)$ and $Z_2 = \varepsilon(T_2)$, one defines their cross-variogram:

$$\gamma_{12}(\mathbf{h}) = \frac{1}{2} E\left[\left(Z_1(\mathbf{u}+\mathbf{h}) - Z_1(\mathbf{u})\right)\left(Z_2(\mathbf{u}+\mathbf{h}) - Z_2(\mathbf{u})\right)\right] \quad (9)$$

which may again be empirically evaluated with:

$$\hat{\gamma}_{12}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left[\left(Z_1(\mathbf{u}_\alpha+\mathbf{h}) - Z_1(\mathbf{u}_\alpha)\right)\left(Z_2(\mathbf{u}_\alpha+\mathbf{h}) - Z_2(\mathbf{u}_\alpha)\right)\right] \quad (10)$$

Again, it should be noted that the previously explained definitions (9) and (10) are true only for second order stationary random variables, meaning:

$$\begin{cases} E\left[Z_i(\mathbf{u})\right] = m_i \quad \forall i \in 1,...,n \text{ and for all locations } \mathbf{u} \\ E\left[\left(Z_i(\mathbf{u}) - m_i\right)\left(Z_j(\mathbf{u}+\mathbf{h}) - m_j\right)\right] = C_{ij}(\mathbf{h}) \quad \forall i, j \in 1,...,n \text{ and for all locations } \mathbf{u},\ \mathbf{u}+\mathbf{h} \end{cases} \tag{11}$$

Equation (7) will extend in the multivariate case, by defining the isotropic variogram matrix function $\boldsymbol{\Gamma}(h)$:

$$\boldsymbol{\Gamma}(h) = \left[\gamma_{ij}(h)\right] = \begin{pmatrix} \gamma_{11}(h) & \cdots & \gamma_{1n}(h) \\ \vdots & \ddots & \vdots \\ \gamma_{n1}(h) & \cdots & \gamma_{nn}(h) \end{pmatrix} \tag{12}$$

Similarly, one denotes the isotropic covariance matrix function $\mathbf{C}(h)$ as follows:

$$\mathbf{C}(h) = \left[C_{ij}(h)\right] = \begin{pmatrix} C_{11}(h) & \cdots & C_{1n}(h) \\ \vdots & \ddots & \vdots \\ C_{n1}(h) & \cdots & C_{nn}(h) \end{pmatrix} \tag{13}$$

where $C_{ij}(h)$ is the covariance function between $Z_i = \varepsilon(T_i)$ and $Z_j = \varepsilon(T_j)$. With these notations, one obtains the following relationship:

$$\mathbf{C}(h) = \mathbf{C}(0) - \boldsymbol{\Gamma}(h) \tag{14}$$

These concepts will be the basis of the quantification of the spatial correlation between pairs of various spectral accelerations, presented in the following sections.

Ground motion data

In this study, the authors used recorded ground-motion data from the Pacific Earthquake Engineering Research (PEER) Center's NGA2 database (http://peer.berkeley.edu/products/strong_ground_motion_db.html). Figure 2 shows plots of the station locations for eight earthquakes which have been considered in this study (Table 1): Northridge, Chi-Chi, Tottori, Niigata, Parkfield, Chuetsu, Iwate, El Mayor Cucapah.

Table 1: The eight considered earthquakes in the spatial correlation study

Earthquake recordings from the NGA database				
Name	Location	Year	Magnitude	Number of recordings
Northridge	California	1994	6.7	152
Chi-Chi	Taiwan	1999	7.6	401
Tottori	Japan	2000	6.6	235
Niigata	Japan	2004	6.6	365
Parkfield	California	2004	6.0	90
Chuetsu	Japan	2007	6.8	403
Iwate	Japan	2008	6.9	280
El Mayor Cucapah	California	2010	7.2	154

Histograms of the number of pairs of stations with separation distance h in each earthquake, to be used in the computation of the empirical variograms with Equation (10), are plotted on Figure 3. The available data from the NGA library can also be plotted in terms of spectra as described on Figure 4, where one can compare the median predictions at each site from the attenuation model on the left with the actual observed spectra on the right. In this study, the authors used the Boore and Atkinson ground motion prediction model (Boore and Atkinson 2008).

Direct variogram fit of empirical data

Fitting technique

Geostatistics literature recommends a manual fit of the variogram, warning against regression methods that might misrepresent the actual information provided by the variogram (Journel and Huijbregts 1978). Each estimated point of the variogram $\hat{\gamma}(h)$ is subject to an error inherent to that point. This error will vary with the considered separation distance h, the extent of the region used in the variogram calculation, etc. For these reasons, variogram fitting cannot be reduced to a simple regression problem. However, given the quantity of data to be analyzed in the multivariate case (we consider 9 periods and 8 earthquakes, resulting in 360 different variograms), it appeared reasonable to develop an automated fitting algorithm to speed up the process, as long as the result of the fit was consistent with independently obtained manual fits.

Not any function can be chosen to fit an empirical variogram. The covariance function, directly related to the variogram through Equation (7), must be positive definite. This is due to the fact that the variance of any linear combination of the covariance computed at p sites must be non-negative; in other words, for any set of locations $\mathbf{u}_\alpha$ and any set of weights ω_α, the covariance function must satisfy:

$$\mathrm{var}\left(\sum_{\alpha=1}^{p}\omega_\alpha Z\left(\mathbf{u}_\alpha\right)\right)=\sum_{\alpha=1}^{p}\sum_{\beta=1}^{p}\omega_\alpha\omega_\beta C\left(\left\|\mathbf{u}_\alpha-\mathbf{u}_\beta\right\|\right)\geq 0 \tag{15}$$

where $\mathrm{var}(\ )$ denotes the variance.

In practice, one models a variogram with a positive linear combination of admissible variogram models. These standard models include, but are not limited to, the four following models. The exponential model is defined as:

$$\tilde{\gamma}(h)=S\left[1-\exp\left(\frac{-3h}{R}\right)\right] \tag{16}$$

where $\tilde{}$ refers to the value from a model, S is the sill and R is the range of the variogram. The sill of a bounded variogram is equal to the variance of Z; for the exponential variogram, it represents the value to which $\tilde{\gamma}(h)$ asymptotically converges as h tends to infinity. The range is then defined as the separation distance h at which $\tilde{\gamma}(h)$ is equal to 95% of the sill of the exponential variogram. This means that the range represents the distance at which 95% of the correlation is lost. The spherical model is defined as:

$$\tilde{\gamma}(h)=\begin{cases}S\left[\frac{3}{2}\left(\frac{h}{R}\right)-\frac{1}{2}\left(\frac{h}{R}\right)^3\right] & \text{if } h\leq R\\ S & \text{if } h>R\end{cases} \tag{17}$$

With this model, the sill S is attained at $h=R$. The third common variogram model is the Gaussian model:

$$\tilde{\gamma}(h)=S\left[1-\exp\left(\frac{-3h^2}{R^2}\right)\right] \tag{18}$$

The sill and the range of the Gaussian variogram are defined as for the exponential variogram. Finally, the nugget effect model is defined as:

$$\tilde{\gamma}(h) = \begin{cases} 0 & \text{if } h = 0 \\ S & \text{if } h > 0 \end{cases} \tag{19}$$

This variogram induces a complete lack of correlation at non-zero separation distance, therefore no range can be defined for the nugget effect. The first three variogram models are shown in Figure 5. The entire correlation structure of the variables of study will be completely defined by the variogram model, which itself depends only on the corresponding sills and ranges.

In this work, we first assumed that each cross-variogram γ_{ij} associated with $\varepsilon(T_i)$ and $\varepsilon(T_j)$ can be modeled with an isotropic exponential function, such that:

$$\tilde{\gamma}_{ij}(h) = S_{ij}\left[1 - \exp\left(\frac{-3h}{R_{ij}}\right)\right] \tag{20}$$

where S_{ij} is the sill and R_{ij} the range. This choice is motivated by results obtained by researchers in the past (e.g., Jayaram and Baker 2009; Wang and Takada 2005), who observed an exponential decay of the correlation coefficient in the univariate case. Indeed, it can be shown using Equations (7) and (8) that the variogram and the correlation coefficient are related as follows:

$$\gamma_{ij}(h) = \rho_{ij}(0) - \rho_{ij}(h) \tag{21}$$

thus, $\tilde{\rho}_{ij}(h) = S_{ij}\exp\left(-3h / R_{ij}\right)$. It can be noted that other functional forms for the correlation coefficient have been proposed, such as the more general $\tilde{\rho}(h) = \exp\left(-\alpha h^{\beta}\right)$ by Goda and Hong (2008), where α and β are constants (for $\beta = 1$, this model is equivalent to the exponential functional form). Boore et al. (2003) used a particular case of Goda and Hong's model with $\beta = 0.5$.

Previous studies have proposed empirical equations to predict the sill (e.g., Baker and Jayaram 2008; Abrahamson, Kammerer, and Gregor 2003; Baker and Cornell 2006; Inoue and

Cornell 1990), as it is equal to the correlation coefficient between $\varepsilon(T_i)$ and $\varepsilon(T_j)$ at the same site ($h = 0$):

$$S_{ij} = \rho_{ij}(0) \tag{22}$$

Various methods were investigated in order to achieve a robust estimation of the sill, among which can be cited: (*i*) a direct computation of $\rho_{ij}(0)$ of the empirical data; (*ii*) calculating the mode of the histogram of the variogram values themselves; (*iii*) a refinement of (*ii*) using a Gaussian kernel function. The last approach proved to be the most robust one, and it has been retained in lieu of the predictive model. Indeed, when fitting a least squares regression, it is critical to assess as correctly as possible the value of the sill, in order to achieve a correct estimate of the range.

The empirical estimation of S with a kernel function relies on a discretization of the observed variogram values, followed by a computation of a kernel weighted function:

$$\begin{aligned}
&y_0 = 0, y_1 = 0.01, ..., y_i = 0.01i, ..., y_{100} = 1 \\
&\text{kernel}(i) = \sum_{k=1}^{k_{\max}} \exp\left(-\frac{\left(\hat{\gamma}(h_k) - y_i\right)^2}{\sigma} \right), \; h_{k_{\max}} = 100 \text{ km}, \;\; \sigma = \text{constant} \\
&S = y_{i_0} \quad \text{s.t.} \;\; \max_i \left(\text{kernel}(i)\right) = \text{kernel}(i_0)
\end{aligned} \tag{23}$$

Variations of the constant σ did not have a significant impact on the final result of the sill value. In this study, a value of $\sigma = 0.1$ has been used.

Once S is accurately determined, the range can be derived using weighted least squares regression. With the exponential variogram model, the problem can be linearized as follows:

$$\hat{\gamma}(h) = S \exp\left(\frac{-3h}{R}\right) \Rightarrow \ln \hat{\gamma}(h) = ah + b \quad \text{with} \quad \begin{cases} a = \dfrac{-3}{R} \\ b = \ln S \end{cases} \tag{24}$$

The regression algorithm will evaluate the weighted sum of squares, as a function of the range R:

$$WSS(R) = \sum_k \omega(h_k) \left[\ln \hat{\gamma}(h_k) - \ln \tilde{\gamma}(h_k) \right]^2 = \sum_k \frac{1}{h_k} \left[\ln \hat{\gamma}(h_k) - \left(\left(\frac{-3}{R} \right) h_k + \ln S \right) \right]^2 \tag{25}$$

where $\omega(h_k)$ is a weighting function giving more importance to the smallest separation distances (an inverse distance weighting has been used here, such that the weight on $\hat{\gamma}(h_k)$ is equal to $1/h_k$). The value of R yielding the minimum of this WSS will be retained as the range of the experimental variogram.

Observed results

Figure 6 shows the result of the kernel fitting for the cross-variogram between $T_1 = 1s$ and $T_2 = 2.5s$ from the Northridge earthquake. The fitted variogram proves to be a good match with the data while also representing a likely outcome of a manual fitting. The kernel fitting provides accurate estimates of the sills (Figure 7) in agreement with Equation (22). Results for cross-variograms between all periods are easily obtained with a similar approach and are shown in Figure 8. However, numerical instabilities may be encountered with residuals having low correlation, e.g. between ε's with very short and very long periods. These cross-variograms are often just "noise" with an almost zero sill and lead to irrelevant estimates of the range due to a non-convergence of the least squares regression. Thus, raw results of the direct variogram fit have to be filtered, as shown in Figure 9. When filtering the results, the authors could observe clusters of data in the range vs. sill plane, where the observations could be distinguished with respect to the value of the corresponding period pairs. It appeared that cross-variograms for two long periods ($T_1 > 1s, T_2 > 1s$) have a higher range of around 50 km, while they show a shorter range of approximately 25 km for two short periods ($T_1 < 1s, T_2 < 1s$). The presence of two structures of different range is indicated in Figure 10 to Figure 12, which show both short and long range components.

The 1999 Chi-Chi earthquake provides many recordings (Figure 2) and thus is one of the most useful events for this study. The direct variogram fitting technique provides adequate representation of the data as can be seen on Figure 13. The results did not show two different spatial structures as it was the case for the Northridge earthquake, but longer variogram ranges were noticed in average, meaning that the correlation between spectral accelerations generally holds for longer distances.

Similar work has been done for the aforementioned six other earthquakes. The same quality of fit could be observed, although the empirical variograms did not demonstrate such a clear exponential trend in some cases, possibly due to the relative lack of data (e.g., for Parkfield and El Mayor). While a simple average of the sills and ranges over all earthquakes may be proposed for the development of a predictive model, some limitations persist as to the exploitation of these results for the formulation of the covariance matrix.

Limitations

The direct variogram fit developed in this study proved to be a useful tool to evaluate the spatial correlation of our empirical data. One may very well use these results to estimate any correlation coefficient between spectral acceleration at two different periods at two different sites. However, a more general objective of this study was to formulate a predictive model for the covariance matrix of a given set of ε's, based on these estimations. When attempting to compute such a model, one only needs the specification of the variogram matrix $\boldsymbol{\Gamma}$ computed before (see Equation (14)).

However, for $\mathbf{C}$ to be an acceptable covariance matrix, the same condition of positive definiteness as in the univariate case (see Equation (15)) must be satisfied: the variance of any weighted linear combination of n variables at p sites must be non-negative. This results in the following requirement for the multivariate case (Wackernagel 1995):

$$\operatorname{var}\left(\sum_{i=1}^{n} \sum_{\alpha=1}^{p} \omega_{\alpha}^{i} Z_i\left(\mathbf{u}_{\alpha}\right) \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{\alpha=1}^{p} \sum_{\beta=1}^{p} \omega_{\alpha}^{i} \omega_{\beta}^{j} C_{ij}\left(\left\| \mathbf{u}_{\alpha} - \mathbf{u}_{\beta} \right\| \right) \geq 0 \tag{26}$$

where ω_{α}^{i} is the weight associated with the value of Z_i at location $\mathbf{u}_{\alpha}$. Unfortunately, the direct variogram fitting approach described above takes no such constraint into account when evaluating empirical sills and ranges, and thus will not lead to a positive definite covariance matrix in most cases. It is possible to "fix" this matrix by merely changing its eigenvalues to make it positive definite. In practice, this is achieved by performing an eigenvalue decomposition of $\mathbf{C}$, such that:

$$\mathbf{C} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1} \tag{27}$$

where each column of $\mathbf{Q}$ is the eigenvector $\mathbf{q}_i$ of $\mathbf{C}$ and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, i.e., $\mathbf{\Lambda}_{ii} = \lambda_i$. We then modify the eigenvalue matrix $\mathbf{\Lambda}$ into $\mathbf{\Lambda}^+$ by changing the negative coefficients to 0 (see section 6.2 from Jäckel 2002):

$$\mathbf{\Lambda}^+_{ii} = \begin{cases} \lambda_i & \text{if } \lambda_i \geq 0 \\ 0 & \text{if } \lambda_i < 0 \end{cases} \tag{28}$$

Finally, $\mathbf{\Lambda}^+$ is recombined with the eigenvector matrix $\mathbf{Q}$ to obtain the positive definite matrix $\mathbf{C}^+$:

$$\mathbf{C}^+ = \mathbf{Q}\mathbf{\Lambda}^+\mathbf{Q}^{-1} \tag{29}$$

The authors have observed that very little changes need to be made to the fitted covariance matrix in order to transform it into a positive definite one (i.e., $\mathbf{C}^+ \approx \mathbf{C}$).

While modifying the eigenvalues is relatively easy to do, it does not allow much control on how much the covariance matrix will be changed. It also makes it difficult to access the "new" actual values of the ranges of the different cross-variograms. Other approaches have been established to generate admissible covariance models. One involves the computation of cross-covariance terms from convolution integrals of the direct covariances (Majumdar and Gelfand 2007) such that $C_{ij}(\mathbf{h}) = \int_{\square^2} \rho_i(\mathbf{u})\rho_j(\mathbf{u}+\mathbf{h})\mathbf{du}$. The integrals can then be evaluated by using polar coordinates and Monte-Carlo integration. Although this approach will lead to valid models, it makes it quite difficult to fit the cross-covariance coefficients to empirical data.

A sufficient solution to remedy the problem of positive definiteness is to impose a single range for all direct and cross-variograms. The covariance matrix function will become:

$$\mathbf{C}(h) = \rho(h) \cdot \mathbf{C}(0) \tag{30}$$

where $\rho(h)$ is a scalar function accounting for the loss of correlation with increasing distance (for instance in the present case, $\rho(h) = \exp(-3h/R)$). This formulation of the covariance matrix function is called the separable model (Banerjee, Gelfand, and Carlin 2004). For the full covariance matrix to be positive definite in this case, one has only to ensure that the covariance matrix at a single site $\mathbf{C}(0)$ is positive definite. This is a much simpler task than ensuring the

full covariance of size the number of periods n times the number of sites p to be positive definite, since $\mathbf{C}(0)$ is only of size $n \times n$. Unfortunately, fitting a single range to the data is not possible, as it does not reflect the underlying structures discovered in this section. The next section will introduce an extension of this separable model that can incorporate more than one range.

The Linear Model of Coregionalization

While fitting independently each empirical variogram may not provide an admissible correlation model, it does give some insight into the spatial characteristics of the considered variables. From the direct variogram fit developed for the Northridge earthquake, we noticed very clear contributions of two different ranges: a short range component acting on small periods and a large range component acting on longer periods. To take the effect of multiple spatial scales into account, a more global model was proposed, which models all variables as linear combinations of the same basic structural components. Analytically, for a given set of n random variables $(Z_1, Z_2, ..., Z_n)$:

$$Z_i(\mathbf{u}) = \sum_{l=0}^{L} \sum_{k=1}^{n_l} a_{ik}^l Y_k^l(\mathbf{u}) + m_i \qquad \forall i = 1, ...n \tag{31}$$

with

- $E\left[Z_i(\mathbf{u})\right] = m_i$
- $E\left[Y_k^l(\mathbf{u})\right] = 0 \quad \forall k, l$
- $\operatorname{cov}\left(Y_k^l(\mathbf{u}), Y_{k'}^{l'}(\mathbf{u}+\mathbf{h})\right) = \begin{cases} c_l(\mathbf{h}) & \text{if } k = k' \text{ and } l = l' \\ 0 & \text{otherwise} \end{cases}$

(Journel and Huijbregts 1978). This is the so-called linear model of coregionalization. This model has become a widely used tool in multivariate geostatistics. The decomposition of the random field into independent components Y^l yields to the following formulation of the variogram matrix (in the isotropic case):

$$\boldsymbol{\Gamma}(h) = \sum_{l=0}^{L} \mathbf{B}^l g^l(h) \tag{32}$$

where $\mathbf{B}^l$ are the coregionalization matrices, $g^l(h)$ are the spatial components to be chosen a priori by the user, and only need to be taken from the admissible variogram functions listed in Equations (16) to (20). One can note that the case of $L = 0$ corresponds to the separable model previously explained. The coregionalization matrices can be interpreted as specific contribution to the sill or variance of each structure $g^l(h)$. It can be shown that in order to ensure the positive definiteness of the covariance matrix, one only needs to provide positive definite $\mathbf{B}^l$ matrices. This is a much simpler task than trying to directly define a np by np covariance matrix.

Fitting technique

Goulard and Voltz (1992) proposed an automated algorithm to fit a Linear Model of Coregionalization (LMC) in a positive definite manner. Its objective is to minimize the weighted sum of squares comparable to the one presented in Equation (25):

$$WSS = \sum_{k=1}^{K} \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} \omega(h_k) \cdot \frac{\left[\tilde{\gamma}_{ij}(h_k) - \hat{\gamma}_{ij}(h_k)\right]^2}{\hat{\sigma}_i \cdot \hat{\sigma}_j} \tag{33}$$

where $\tilde{\gamma}_{ij}(h_k)$ denotes the value of the variogram model, $\hat{\gamma}_{ij}(h_k)$ is the actual variogram empirical value, $\omega(h_k)$ the weight at lag h_k, $\hat{\sigma}_i$ the observed standard deviation of Z_i. The WSS is simply a weighted sum of the standardized squared errors between the empirical variogram and the model, over all periods and all discrete separation distances. The Goulard algorithm has become popular in multivariate geostatistics involving coregionalization studies, as it provides a fast and elegant way to fit all crossvariograms while ensuring the positive definiteness of the resulting covariance matrix.

The Goulard algorithm is executed as follows:

1. Initialize the $L+1$ coregionalization matrices $\mathbf{B}^l$ with any values.

2. Take out one of the $L+1$ variogram models ($g^{l_0}(h)$), compute the difference between the initial empirical model and the LMC deprived of the $l_0{}^{\text{th}}$ structure:

$$\boldsymbol{\Lambda}_{l_0}\boldsymbol{\Gamma}(h_k) = \hat{\boldsymbol{\Gamma}}(h_k) - \sum_{\substack{l=0 \\ l\neq l_0}}^{L} \hat{\mathbf{B}}^{l} g^{l}(h_k) \tag{34}$$

3. Compute the symmetric matrix:

$$\mathbf{G}_{l_0} = \sum_{k=1}^{K} \omega(h_k) \cdot \boldsymbol{\Delta}_{l_0}\boldsymbol{\Gamma}(h_k) \cdot g^{l_0}(h_k) \tag{35}$$

4. Obtain the spectral decomposition of $\mathbf{G}_{l_0} = \mathbf{Q}_{l_0}\boldsymbol{\Lambda}_{l_0}\mathbf{Q}_{l_0}^{\mathbf{T}}$. Set all negative eigenvalues to 0 by forming: $\mathbf{G}_{l_0}^{+} = \mathbf{Q}_{l_0}\boldsymbol{\Lambda}_{l_0}^{+}\mathbf{Q}_{l_0}^{\mathbf{T}}$ where $\boldsymbol{\Lambda}_{l_0}^{+}$ is $\boldsymbol{\Lambda}_{l_0}$ with all the negative diagonal terms changed to 0 (this is similar to what was discussed in Equations (27) to (29)).
5. Compute the new coregionalization matrix corresponding to the $l_0{}^{\text{th}}$ structure:

$$\hat{\mathbf{B}}^{l_0} = \frac{\mathbf{G}_{l_0}^{+}}{\sum_{k=1}^{K} \omega(h_k) \cdot \left[g^{l_0}(h_k) \right]^2} \tag{36}$$

6. Increment $l_0 \leftarrow l_0 + 1$ ($l_0 \leftarrow 0$ if $l_0 > L$) and loop over steps 2 to 5 until WSS is smaller than a user-specified threshold.

This algorithm is equivalent to fitting one structure at the time to the empirical data, while ensuring positive definiteness of each coregionalization matrix at step 4. The procedure is not guaranteed to converge in theory, but the experience has shown that the algorithm almost always converges whatever the initial choice of the coregionalization matrices at step 1 (Goovaerts 1997). The authors were able to confirm this experience with the ground motion data considered here.

The first step in fitting a coregionalization model is to choose a set of basic structures $g^{l}(h)$ among the admissible functions described in the previous section. At this point, insights

from the earlier direct variogram fits become more useful to identify which structures should be included in the model. While one could think of considering a nested model composed of all the different exponential functions previously fitted for each period pair, it is better to minimize the number of structures to simplify both calculation and further interpretation. Thus, the authors propose to keep one short range exponential function of (20 kilometers) and one long range exponential function (70 kilometers), so that the variogram matrix function can be expressed as:

$$\boldsymbol{\Gamma}(h) = \mathbf{B}^1\left(1-\exp\left(\frac{-3h}{20}\right)\right) + \mathbf{B}^2\left(1-\exp\left(\frac{-3h}{70}\right)\right) \tag{37}$$

This choice is motivated by the observations obtained from the direct variogram fit, where two distinct ranges emerged from the overall fit of all cross-covariance terms. The values of 20 and 70 km were picked based on the analysis of the 8 studied earthquakes, in order for all variogram ranges to fall approximately within these boundaries while ensuring an adequate fit at small separation distances. The retained weighting was the same as for the direct variogram fitting, such that $\omega(h_k) = 1/h_k$.

Northridge

We present here the fitting of an LMC to the empirical variograms previously examined. We observe that the coregionalization model matches quite well with the observed data. However, we noted a relatively high value of the *WSS* compared to the other earthquakes. This is mainly explained by the noise in the empirical variograms, which can be observed for instance on the bottom-right plot of Figure 14. The reason for this noise is the lack of enough available data at high periods ($T > 3s$). Nevertheless, it should be noted that the *WSS* cannot be considered as an absolute measure of the goodness of fit of a coregionalization model for a variety of reasons (Goovaerts 1997). Indeed, from Equation (33), it can be seen that the *WSS* directly depends on the number of considered lags K, such that the linear model that yields the smallest *WSS* for a given K might not necessarily minimize *WSS* for a different K'. Also, a small *WSS* value may be artificially obtained by sacrificing the goodness of fit of the direct variograms (relative to a single period) to the fit of the cross-variograms. All results have to be checked visually in order to ensure an acceptable model. As mentioned before, one desires the best fit at the smaller separation distances, which is what is observed for the particular cross-variogram of Figure 15.

On this plot of the cross-variogram between $\varepsilon(1s)$ and $\varepsilon(2s)$ for the Northridge earthquake, one obtains a very good fit for distances smaller than 50 km, while the larger distances show much more noise in the empirical variogram and a poorer fit. However, the weighted fitting is not sensitive to these large distance values and so provides a robust estimation of the data at short distances.

The shape of the coregionalization matrices over the different period pairs are shown on Figure 16. The plots demonstrate that the short range matrix $\mathbf{B}^1$ makes a larger contribution to the variogram at small periods, while the long range matrix $\mathbf{B}^2$ has a more significant impact on large periods. This result is in agreement with the observations made in the direct variogram fit.

Chi-Chi

Direct variogram fit of the Chi-Chi earthquake residuals showed a more unique spatial behavior, in a sense that the authors could not identify two structures as clearly as with the Northridge earthquake. However, Figure 17 shows that the same linear coregionalization model used with the Northridge data still provided very accurate fits of every variogram, mainly because the observed ranges in the direct variogram fit were also within 20 to 70 kilometers. Furthermore, the obtained *WSS* is dramatically lower than for the Northridge earthquake data, due to the availability of many more data at high periods, thus reducing the variability of the empirical variograms. Once again, it is important to check the accuracy of the fitting for each variogram, especially at small distances. Figure 18 shows the particular fitting of the cross-variogram between 1s and 2s, where one observes that the coregionalization model provides a particularly good fit of the empirical data, particularly at distances of less than 50 km.

Other earthquakes

Other earthquakes were investigated, although they did not provide as much data as the Chi-Chi recordings. Still, the linear model of coregionalization lead once again to acceptable results as far as the quality of the fit is concerned. Figure 19 and Figure 20 show the 3D plots of the coregionalization matrices obtained for each earthquake. It can be observed that the shapes of the matrices look quite similar from earthquake to earthquake, except for the high periods of the

Northridge earthquake ($T > 3s$), due to the relative insufficiency of the data. Note also that the lack of systematically differing patterns between these cases suggests that this data set does not provide evidence to build a model that varies by region, or by earthquake magnitude. This lack of observed variation is not proof that no such trends exist, but rather that if they exist they are subtle enough that they cannot be detected using currently available earthquake strong motion data. The individual fits for each event provide useful results that will be incorporated in the proposed predictive equation.

Consistency

The empirical covariance matrix functions developed for the Chi-Chi earthquake were then used to generate ε data via Monte-Carlo simulations at the locations of the recordings from the same Chi-Chi earthquake. From the simulated ε, another model of coregionalization was fit, and compared to the model of coregionalization which the empirical covariance was based on. Figure 21 shows the surface plots of the difference between the initial coregionalization matrices and the ones fitted to the generated data:

$$\begin{cases} \boldsymbol{\Delta}_{ij}^{\mathrm{SR}} = \mathbf{B}_{\mathbf{initial},ij}^{1} - \mathbf{B}_{\mathbf{fitted},ij}^{1} \\ \boldsymbol{\Delta}_{ij}^{\mathrm{LR}} = \mathbf{B}_{\mathbf{initial},ij}^{2} - \mathbf{B}_{\mathbf{fitted},ij}^{2} \end{cases} \tag{38}$$

Very little differences were observed (i.e. $\boldsymbol{\Delta}^{\mathrm{SR}}$ and $\boldsymbol{\Delta}^{\mathrm{LR}}$ are both close to 0), which indicates the robustness and unbiasedness of the method.

Observations

Extending the simple framework of the separable model, the linear model of coregionalization proved to be a reliable technique to fit many cross-covariances at once. The Goulard algorithm is both fast and easy to use, as it does not require any other input than the empirical variograms and the set of basic structures $g^{l}(h)$. The goodness of fit obtained with this new method is somewhat comparable to the results from the direct variogram fitting from the authors' previous work. While one could derive correlation estimates from the former method, the linear model of coregionalization also provides an admissible model for simulation purposes.

(the positive definiteness of the full covariance matrix is ensured as long as each coregionalization matrix $\mathbf{B}^l$ is positive definite).

The next step of this study will be to use the fitted coregionalization models to build a predictive equation for the covariance matrix function, $\mathbf{C}(h)$, and the matrix of correlation coefficients at lag h, $\mathbf{R}(h)$.

Formulation of a predictive model

From all investigated earthquake data, the authors propose a model to predict the covariance matrix function $\mathbf{C}(h)$ from a sampling of nine periods ranging from 0 to 10 seconds, by averaging all the fitted coregionalization matrices over the various earthquakes (Figure 22).

One can extract any subsample of periods and use the corresponding coregionalization sub-matrices for simulation purposes. In the case one might want to consider an extra period that does not belong to the proposed sample, linear interpolation between periods can be used as long as the positive definiteness of the resulting coregionalization matrices is verified. If the resulting coregionalization matrix is not positive definite, then setting the eigenvalues of the non-positive definite matrix to 0 will lead to an admissible model (see procedure described in Equations (27) to (29)).

The variogram matrix function is first modeled using Equation (37) with the coregionalization matrices $\mathbf{B}^1$ and $\mathbf{B}^2$. The covariance matrix function $\mathbf{C}(h)$ can be obtained from the variogram matrix with Equation (14) by noting that:

$$\mathbf{C}(0) = \lim_{h\to\infty} \boldsymbol{\Gamma}(h) = \mathbf{B}^1 + \mathbf{B}^2 \tag{39}$$

which yields the following simple formulation:

$$\mathbf{C}(h) = \mathbf{B}^1 \exp\left(\frac{-3h}{20}\right) + \mathbf{B}^2 \exp\left(\frac{-3h}{70}\right) \tag{40}$$

The authors report the $\mathbf{B}^l$ matrices in Table 2 and Table 3 below, for a set of nine sampling periods.

Table 2: Short range coregionalization matrix, $\mathbf{B}^1$

$\mathbf{B}^1$									
Period	PGA	0.1	0.2	0.5	1	2	5	7.5	10
PGA	0.54								
0.1	0.47	0.55							
0.2	0.46	0.41	0.56				sym.		
0.5	0.34	0.22	0.28	0.54					
1	0.19	0.08	0.10	0.33	0.46				
2	0.09	0.01	0.02	0.17	0.29	0.44			
5	0.04	0.00	-0.01	0.09	0.17	0.30	0.44		
7.5	0.05	0.02	-0.01	0.09	0.17	0.28	0.40	0.46	
10	0.06	0.04	0.02	0.07	0.12	0.23	0.34	0.38	0.37

Table 3: Long range coregionalization matrix, $\mathbf{B}^2$

$\mathbf{B}^2$									
Period	PGA	0.1	0.2	0.5	1	2	5	7.5	10
PGA	0.30								
0.1	0.25	0.28							
0.2	0.26	0.23	0.29				sym.		
0.5	0.23	0.14	0.23	0.32					
1	0.16	0.06	0.13	0.26	0.36				
2	0.10	0.01	0.07	0.21	0.32	0.39			
5	0.07	0.00	0.02	0.15	0.21	0.29	0.41		
7.5	0.06	-0.01	0.01	0.13	0.18	0.26	0.39	0.45	
10	0.05	-0.02	0.00	0.14	0.21	0.30	0.41	0.47	0.58

Predictive model for the correlation coefficient

It is straightforward to obtain the corresponding matrix of correlation coefficients at lag h, also called correlogram matrix, denoted by:

$$\mathbf{R}(h) = \left[\rho_{ij}(h)\right] = \begin{pmatrix} \rho_{11}(h) & \cdots & \rho_{1n}(h) \\ \vdots & \ddots & \vdots \\ \rho_{n1}(h) & \cdots & \rho_{nn}(h) \end{pmatrix} \tag{41}$$

While the correlation coefficient is the most commonly used in practice, there exist various other coefficients measuring spatial linear dependency between random variables such as the

codispersion coefficient or the structural correlation coefficient (Goovaerts 1992). One can show that this correlogram matrix has the following formulation:

$$\mathbf{R}(h) = \mathbf{P}^1 \exp\left(\frac{-3h}{20}\right) + \mathbf{P}^2 \exp\left(\frac{-3h}{70}\right) \tag{42}$$

where the $\mathbf{P^1}$ and $\mathbf{P^2}$ matrices are standardized versions of the $\mathbf{B^1}$ and $\mathbf{B^2}$ matrices, such that:

$$\mathbf{P}_{ij}^{l} = \frac{\mathbf{B}_{ij}^{l}}{\left(\sqrt{\mathbf{B}_{ii}^{1} + \mathbf{B}_{ii}^{2}}\right) \times \left(\sqrt{\mathbf{B}_{jj}^{1} + \mathbf{B}_{jj}^{2}}\right)}, \quad l = 1, 2 \tag{43}$$

This result is simply obtained by dividing the covariance matrix coefficients by the product of the standard deviations at the two considered periods, since:

$$\sigma_i = \sqrt{C_{ii}(0)} = \sqrt{\mathbf{B}_{ii}^{1} + \mathbf{B}_{ii}^{2}} \tag{44}$$

The authors report the $\mathbf{P}^l$ matrices in Table 4 and Table 5 below, for a set of 9 sampling periods.

Table 4: Short range standardized coregionalization matrix, $\mathbf{P}^1$

P¹									
Period	PGA	0.1	0.2	0.5	1	2	5	7.5	10
PGA	0.64								
0.1	0.57	0.66							
0.2	0.54	0.49	0.66				sym.		
0.5	0.40	0.26	0.33	0.63					
1	0.23	0.10	0.12	0.39	0.56				
2	0.11	0.01	0.02	0.20	0.36	0.53			
5	0.05	0.01	-0.01	0.10	0.20	0.35	0.52		
7.5	0.05	0.02	-0.01	0.10	0.19	0.32	0.46	0.51	
10	0.07	0.05	0.02	0.08	0.14	0.26	0.38	0.41	0.39

Table 5: Long range standardized coregionalization matrix, $\mathbf{P}^2$

$\mathbf{P^2}$									
Period	PGA	0.1	0.2	0.5	1	2	5	7.5	10
PGA	0.36								
0.1	0.30	0.34							
0.2	0.31	0.27	0.34				sym.		
0.5	0.27	0.17	0.27	0.37					
1	0.19	0.08	0.16	0.31	0.44				
2	0.12	0.01	0.08	0.25	0.38	0.47			
5	0.08	0.00	0.03	0.17	0.25	0.35	0.48		
7.5	0.06	-0.01	0.02	0.15	0.21	0.30	0.45	0.49	
10	0.05	-0.03	0.00	0.15	0.24	0.33	0.46	0.51	0.61

Example: Suppose one intends to quantify the correlation coefficient between $\ln S_a(1s)$ at site A and $\ln S_a(2s)$ at site B in a given earthquake, where sites A and B are separated by a distance of $h = 15$ kilometers. One reads in Table 4 and Table 5 that $\mathbf{P}_{12}^1 = 0.36$ and $\mathbf{P}_{12}^2 = 0.38$, and substitutes these values in Equation (42) to obtain:

$$\rho_{12}(15) = 0.36\exp\left(\frac{-3\times 15}{20}\right) + 0.38\exp\left(\frac{-3\times 15}{70}\right) = 0.24 \tag{45}$$

This calculation is clearly rather simple, indicating that while the calibration of the model was quite complex, it is very easy to apply.

Case study: an evaluation of the Markov-type screening hypothesis

In this section, the authors present an application of the use of the proposed covariance model. While showing the general principles of the construction of the spatial covariance matrix, this study will also evaluate the impact of accounting for different sets of other ground motion intensities (e.g. spectral accelerations at different sites or different periods) in the variance of the final prediction of one ground motion intensity at a given site.

Models that involve conditioning on a smaller set of variables rather than the full considered set are called Markov models. Journel (1999) introduced a Markov-type model to be used in the joint modeling of two random variables Z_1 and Z_2, considering the "screening" hypothesis stated as follows:

$$E\left[Z_2(u) \mid Z_1(u); Z_1(u+h)\right] = E\left[Z_2(u) \mid Z_1(u)\right] \tag{46}$$

In words, this hypothesis assumes that the dependence of the variable Z_2 on the primary variable Z_1 is limited to the co-located primary variable. In practice, Z_1 would have a larger correlation range than Z_2. Under this hypothesis, the spatial correlation between the two can be shown equal to:

$$\rho_{12}(h) = \rho_{12}(0) \cdot \rho_1(h) \tag{47}$$

Goda and Hong (2008) proposed such a model to characterize the spatial correlation between spectral accelerations at different periods ($Z_1 = \ln S_a(T_1), Z_2 = \ln S_a(T_2)$, with $T_1 > T_2$). This is consistent with the definition of the primary variable above, since the authors often observed larger correlation ranges for higher periods. In the following, the authors evaluate the accuracy of this screening hypothesis by comparing predictions from the Markov dependence model to corresponding predictions from the full linear model of coregionalization derived above.

Accuracy of the Markov approximations

The model presented in Equation (47) is examined in this section. Figure 23 shows a comparison of the correlation coefficients obtained from the full linear model of coregionalization and from the Markov model of case 2 (Equation (47)), at $T_1 = 2s$ and $T_2 = 1s$. The latter model can be considered as a "reduced" coregionalization model, because it is still based on the previously developed LMC, but only one of the periods is involved in the spatial decaying part. Also plotted is the case $T_1 = 1s$ and $T_2 = 2s$, for which one observes a slightly greater difference with the full LMC: this confirms that the primary period should generally be the larger one. One observes a very good match between the two approaches over all separation distances. While it provides a simpler way to estimate spatial correlation, this relative result should be put into perspective, as it still relies on the developed coregionalization model. However, as can be seen on Figure 24, this Markov approximation is not as good for periods more spread apart (plotted are the cross-correlations corresponding to $T_1 = 2s$ and $T_2 = 0.2s$). In such a case, using the full coregionalization model is the better option.

Precision of the Markov approximations

While the accuracy of a Markov-type approximation was just discussed, it is also important to study the resulting variance of such estimation. The first case considered here is the situation where one observes spectral accelerations at various periods but at one site, and wants to predict the spectral acceleration at a single period at another site. In other words, one desires to know $\varepsilon_{\text{site A}}(T^*)$ conditioned on the observations $\left\{\varepsilon_{\text{site B}}(T_1),...,\varepsilon_{\text{site B}}(T^*),...,\varepsilon_{\text{site B}}(T_n)\right\}$. A problem of interest is how $\text{var}\left[\varepsilon_{\text{site A}}(T^*)\,|\,\varepsilon_{\text{site B}}(T^*)\right]$ compares to $\text{var}\left[\varepsilon_{\text{site A}}(T^*)\,|\,\varepsilon_{\text{site B}}(T_1),...,\varepsilon_{\text{site B}}(T^*),...,\varepsilon_{\text{site B}}(T_n)\right]$, which is actually the evaluation of the amount of extra information brought by incorporating additional conditioning periods at a remote site to assess the primary residual of interest. It can be theoretically shown that accounting for multiple conditioning periods rather that a single one will reduce the variance of $\varepsilon_{\text{site A}}(T^*)$, thereby resulting in an increase in the accuracy of the intensity estimates (Goovaerts 1997). Due to the multivariate normal distribution of a vector of spatially distributed epsilons, one can easily compute the presented conditional variances; denoting $\boldsymbol{\varepsilon}_1$ the set of residuals to be predicted, conditioned on the set of residuals $\boldsymbol{\varepsilon}_2$, one can express their joint distribution as follows:

$$\begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right) \tag{48}$$

where $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The covariance matrix is obtained using a spatial correlation model described earlier. Given this model for the joint distribution, the distribution of $\boldsymbol{\varepsilon}_1$ conditional on $\boldsymbol{\varepsilon}_2$ is obtained as follows:

$$\boldsymbol{\varepsilon}_1 \,|\, \boldsymbol{\varepsilon}_2 \sim N\left(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{e}, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right) \tag{49}$$

where $\mathbf{e}$ is the vector of observed values of $\boldsymbol{\varepsilon}_2$ at the recording stations. As a further application, the expected ground motion intensities at all sites are then obtained by combining the median intensities with the expected value of the residuals obtained from Equation (49). Denoting h the

separation distance between site A and site B, one can form the covariance matrices of interest to evaluate $\text{var}\left[\varepsilon_{\text{site A}}(T^*) \mid \varepsilon_{\text{site B}}(T^*)\right] = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ with:

$$\begin{cases} \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22} = \left[C_{ss}(0)\right] \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = \left[C_{ss}(h)\right] \end{cases} \tag{50}$$

with $C_{ss}(h)$ the covariance matrix coefficient corresponding to the period T^*. Similarly, in order to estimate $\text{var}\left[\varepsilon_{\text{site A}}(T^*) \mid \varepsilon_{\text{site B}}(T_1),...,\varepsilon_{\text{site B}}(T^*),...,\varepsilon_{\text{site B}}(T_n)\right]$, the corresponding submatrices will be:

$$\begin{cases} \boldsymbol{\Sigma}_{11} = \left[C_{ss}(0)\right] \\ \boldsymbol{\Sigma}_{22} = \begin{bmatrix} C_{11}(0) & \cdots & C_{1s}(0) & \cdots & C_{1n}(0) \\ & \ddots & & & \vdots \\ & & C_{ss}(0) & & C_{sn}(0) \\ & \text{sym.} & & \ddots & \vdots \\ & & & & C_{nn}(0) \end{bmatrix} \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^{\mathrm{T}} = \left[C_{11}(h),...,C_{1s}(h),...,C_{1n}(h)\right] \end{cases} \tag{51}$$

Figure 25 shows a plot of the relative variance reduction $r_{\text{case 1}}$ for different choices of the primary period T^*, over a varying separation distance h:

$$r_{\text{case 1}} = \frac{\text{var}\left[\varepsilon_{\text{site A}}(T^*) \mid \varepsilon_{\text{site B}}(T_1),...,\varepsilon_{\text{site B}}(T^*),...,\varepsilon_{\text{site B}}(T_n)\right] - \text{var}\left[\varepsilon_{\text{site A}}(T^*) \mid \varepsilon_{\text{site B}}(T^*)\right]}{\text{var}\left[\varepsilon_{\text{site A}}(T^*)\right] - \text{var}\left[\varepsilon_{\text{site A}}(T^*) \mid \varepsilon_{\text{site B}}(T^*)\right]} \tag{52}$$

with T^* chosen among the sample $[0.1, 0.2, 0.5, 1, 2, 5, 7.5, 10]$, and T_1 to T_n the remaining periods of that sample. One observes that $r_{\text{case 1}}$ is equal to zero at $h = 0$, since site A and site B are at the same location, and the two variances from the numerator are equal to 0. Also, $r_{\text{case 1}}$ tends to 0 as h tends to infinity, because the conditional set of observations at a far away site does not bring any information about the residual at the primary site. Overall, little variance reduction (around 2 to 3%) is achieved when incorporating multiple periods in the conditional set of observations, supporting the reasonableness of a screening hypothesis in the joint modeling of spectral accelerations.

A second case, directly derived from Equation (46), was investigated in a similar manner. Its accuracy has been evaluated in the previous section (Figure 23 and Figure 24). The problem is now to predict the residual $\varepsilon_{\text{site A}}(T_1)$ conditioned on the residual at the same site but at a different period $\varepsilon_{\text{site A}}(T_2)$, and then to quantify the variance reduction generated by considering the residual at the conditioning period and at a remote site $\varepsilon_{\text{site B}}(T_2)$. Equation (49) still applies, one will estimate $\text{var}\left[\varepsilon_{\text{site A}}(T_1) \mid \varepsilon_{\text{site A}}(T_2)\right] = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ with:

$$\begin{cases} \boldsymbol{\Sigma}_{11} = \left[C_{11}(0)\right] \\ \boldsymbol{\Sigma}_{22} = \left[C_{22}(0)\right] \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = \left[C_{12}(0)\right] \end{cases} \tag{53}$$

with $C_{12}(0)$ the covariance matrix coefficient corresponding to the periods T_1 and T_2. Similarly, one can compute $\text{var}\left[\varepsilon_{\text{site A}}(T_1) \mid \varepsilon_{\text{site A}}(T_2), \varepsilon_{\text{site B}}(T_2)\right]$, denoting h the separation distance between site A and site B:

$$\begin{cases} \boldsymbol{\Sigma}_{11} = \left[C_{11}(0)\right] \\ \boldsymbol{\Sigma}_{22} = \begin{bmatrix} C_{22}(0) & C_{22}(h) \\ C_{22}(h) & C_{22}(0) \end{bmatrix} \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^{\mathbf{T}} = \begin{bmatrix} C_{12}(0) & C_{12}(h) \end{bmatrix} \end{cases} \tag{54}$$

A plot of the relative variance reduction $r_{\text{case 2}}$ for different choices of the primary period T_1 is shown on Figure 26:

$$r_{\text{case 2}} = \frac{\text{var}\left[\varepsilon_{\text{site A}}(T_1) \mid \varepsilon_{\text{site A}}(T_2), \varepsilon_{\text{site B}}(T_2)\right] - \text{var}\left[\varepsilon_{\text{site A}}(T_1) \mid \varepsilon_{\text{site A}}(T_2)\right]}{\text{var}\left[\varepsilon_{\text{site A}}(T_1)\right] - \text{var}\left[\varepsilon_{\text{site A}}(T_1) \mid \varepsilon_{\text{site A}}(T_2)\right]} \tag{55}$$

T_1 was selected among the list of periods $[0.1, 0.2, 0.5, 1, 2, 5, 7.5, 10]$, and T_2 was chosen as the closest inferior period to T_1 in that same set (for the first period of the set $T_1 = 0.1s$, $T_2 = 0.01s$ was considered). Again, $r_{\text{case 2}}$ has the same properties as $r_{\text{case 1}}$ as h tends to 0 and to infinity. Even less variance reduction is achieved as compared to the previously investigated case

($r_{\text{case 2}} \approx r_{\text{case 1}}$), which means that the estimation variance is not affected by the incorporation of the extra information $\varepsilon_{\text{site B}}(T_2)$. This corroborates the screening hypothesis from Equation (46).

Summary

This research has investigated various techniques to model the spatial correlation of spectral accelerations at multiple periods. Quantifying this correlation was done with geostatistical tools involving variogram modeling, a common measure of spatial dissimilarity. Ground motions recordings from eight different earthquakes (Northridge, Chi-Chi, Tottori, Niigata, Parkfield, Chuetsu, Iwate, El Mayor Cucapah) were used to compute empirical variograms of spectral acceleration residuals at different periods.

The authors' first focus was to fit independently each cross-variogram with an exponential function fully characterized by a sill (asymptotical value of the variogram) and a range (distance at which correlation is effectively zero), which provided relevant insight of the data. An automated least squares algorithm was developed, with a robust estimation of the sill using a kernel method. This approach allows evaluating a correlation coefficient between spectral accelerations at different periods and at different sites.

This first result is informative, but is not compatible with the generation of simulated ground motion maps, which requires a positive definite covariance matrix. Based on the direct fit results, two underlying structures were identified (short- and long-range functions both accounting for the spatial decaying of the correlation as distance increases) that became input of a linear model of coregionalization, equivalent to the modeling of each cross-variogram with a linear combination of those same two exponential functions. Extending the simple framework of the separable model (in which only one range is used for all cross-variograms), the linear model of coregionalization proved to be a reliable technique to fit many cross-covariances at once. The Goulard algorithm, involved in the automated fitting of the model, is both fast and easy to use, as it does not require any other input than the empirical variograms and the set of basic structures $g^l(h)$. The goodness of fit obtained with this new method is somewhat comparable to the results from the direct variogram fitting from the authors' previous work. This allowed generating a new admissible covariance model applicable for ground motion simulation purposes.

Using this model, the correlation coefficient between any pair of spectral accelerations at different periods and at different sites may also be easily retrieved as shown in a simple example.

The robustness of the model calibration approach was evaluated using a novel approach, by simulating a synthetic set of ground motion data from the estimated cross-variogram model, and attempting to re-estimate the model from the synthetic data. The estimated cross-variograms obtained from the synthetic data were very similar to the cross-variogram model, indicating that the algorithm is able to accurately detect spatial correlation features from observed ground motions.

The developed covariance model was then used, to examine the validity of a Markovian screening hypothesis in the case of ground motion residuals. After investigating two different Markov models, the authors focused on the one formulating the cross-correlation coefficient as a product of the cross-correlation at a single site times the spatial correlation coefficient of the highest period. This approach proved to be compatible with the developed coregionalization model, and can therefore be considered as a possible simplification of the full linear model of coregionalization, as long as the two considered periods are relatively close to one another.

Even though the calibration of this model and investigation of its implications were somewhat complex, it should be emphasized that the model is very simple to use for making correlation predictions. A user of this model only needs to evaluate Equation (42), with the needed coefficients from Table 4 and Table 5, to compute a correlation coefficient for spectral values at two periods at a specified separation distance. While this model is more general than most previous models that considered only single-period correlations or used a Markov-type assumption to compute multi-period correlations, the model proposed here is not significantly more complex to use than those earlier models, and so should be a useful resource for those interested in predicting correlations of spectral values at differing locations and periods.

Acknowledgements

This work was supported by the U.S. Geological Survey (USGS) via External Research Program award G10AP00046. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the USGS.

References

Abrahamson, N. A., Annie Kammerer, and Nick Gregor. 2003. Summary of scaling relations for spectral damping, peak velocity, and average spectral acceleration: Report for the PEGASOS project. Personal communication.

Abrahamson, Norman, and Walter Silva. 2008. "Summary of the Abrahamson & Silva NGA Ground-Motion Relations". Earthquake Spectra 24 (1): 67-97.

Baker, Jack W, and C Allin Cornell. 2006. "Correlation of response spectral values for multi-component ground motions". Bulletin of the Seismological Society of America 96 (1): 215-227.

Baker, Jack W, and Nirmal Jayaram. 2008. "Correlation of spectral acceleration values from NGA ground motion models". Earthquake Spectra 24 (1): 299-317. doi:10.1193/1.2857544.

Banerjee, S, Alan Gelfand, and Bradley Carlin. 2004. Multivariate spatial modeling. In Hierarchical Modeling and Analysis for Spatial Data. C&H/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC.

Bazzurro, P, and N. Luco. 2004. Effects of Different Sources of Uncertainty and Correlation on Earthquake-Generated Losses. In Proc. International Forum on Engineering Decision Making, Stoos, Switzerland.

Boore, David M., and Gail M. Atkinson. 2008. "Ground-Motion Prediction Equations for the Average Horizontal Component of PGA, PGV, and 5%-Damped PSA at Spectral Periods between 0.01 s and 10.0 s". Earthquake Spectra 24 (1): 99-138.

Boore, David M., James F. Gibbs, William B. Joyner, John C. Tinsley, and Daniel J. Ponti. 2003. "Estimated Ground Motion From the 1994 Northridge, California, Earthquake at the Site of the Interstate 10 and La Cienega Boulevard Bridge Collapse, West Los Angeles, California". Bulletin of the Seismological Society of America 93 (6): 2737-2751.

Campbell, Kenneth W., and Yousef Bozorgnia. 2008. "NGA Ground Motion Model for the Geometric Mean Horizontal Component of PGA, PGV, PGD and 5% Damped Linear Elastic Response Spectra for Periods Ranging from 0.01 to 10 s". Earthquake Spectra 24 (1): 139-171.

Chiou, Brian S. J., and Robert R. Youngs. 2008. "An NGA Model for the Average Horizontal Component of Peak Ground Motion and Response Spectra". Earthquake Spectra 24 (1): 173-215.

Foulser-Piggott, Roxane, and Peter J. Stafford. 2011. "A predictive model for Arias intensity at multiple sites and consideration of spatial correlations". Earthquake Engineering & Structural Dynamics: n/a.

Goda, K., and H. P. Hong. 2008. "Spatial Correlation of Peak Ground Motions and Response Spectra". Bulletin of the Seismological Society of America 98 (1): 354-365. doi:10.1785/0120070078.

Goovaerts, Pierre. 1992. "Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information". Journal of Soil Science 43 (4) (December 1): 597-619. doi:10.1111/j.1365-2389.1992.tb00163.x.

Goovaerts, Pierre. 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, USA, September 18.

Goulard, M., and M. Voltz. 1992. "Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix". Mathematical Geology 24 (3) (April 1): 269-286-286.

Inoue, Takashi, and C Allin Cornell. 1990. Seismic Hazard Analysis of Multi-Degree-of-Freedom Structures. Report #RMS-8. Stanford, CA: Reliability of Marine Structures.

Jäckel, Peter. 2002. Monte Carlo methods in finance. Chichester, West Sussex, England; J. Wiley & Sons.

Jayaram, Nirmal, and Jack W Baker. 2008. "Statistical Tests of the Joint Distribution of Spectral Acceleration Values". Bulletin of the Seismological Society of America 98 (5): 2231-2243. doi:10.1785/0120070208.

Jayaram, Nirmal, and Jack W. Baker. 2009. "Correlation model for spatially distributed ground-motion intensities". Earthquake Engineering & Structural Dynamics 38 (15): 1687-1708. doi:10.1002/eqe.922.

Journel, A. G. 1999. "Markov Models for Cross-Covariances". Mathematical Geology 31 (8) (November 1): 955-964.

Journel, A. G., and Ch. J. Huijbregts. 1978. Mining geostatistics. London; New York: Academic Press.

Majumdar, Anandamayee, and Alan Gelfand. 2007. "Multivariate Spatial Modeling for Geostatistical Data Using Convolved Covariance Functions". Mathematical Geology 39 (2) (February 1): 225-245-245.

Wackernagel, Hans. 1995. Multivariate geostatistics: an introduction with applications. Berlin: Springer.

Wang, Min, and Tsuyoshi Takada. 2005. "Macrospatial Correlation Model of Seismic Ground Motions". Earthquake Spectra 21 (4): 1137-1156.

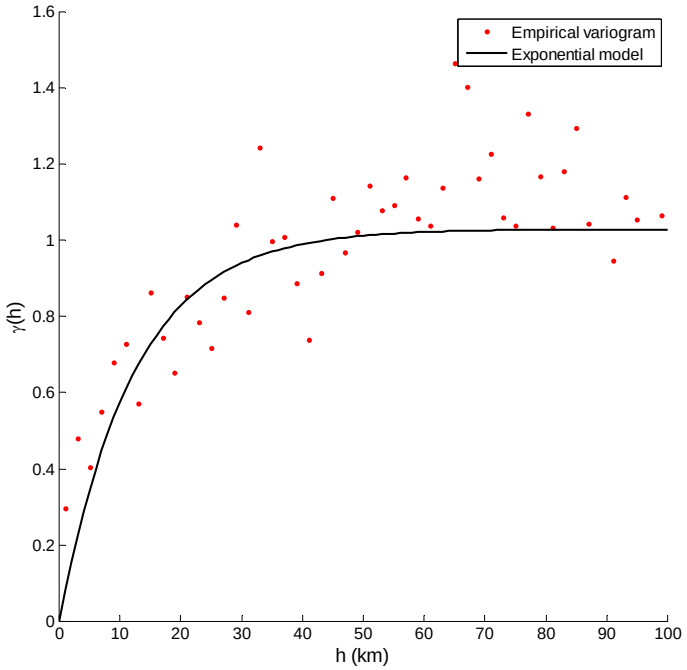

Figure 1: Empirical variogram and fitted exponential model of the normalized residuals (ε) from the Northridge earthquake data, at T_1=T_2=1s.

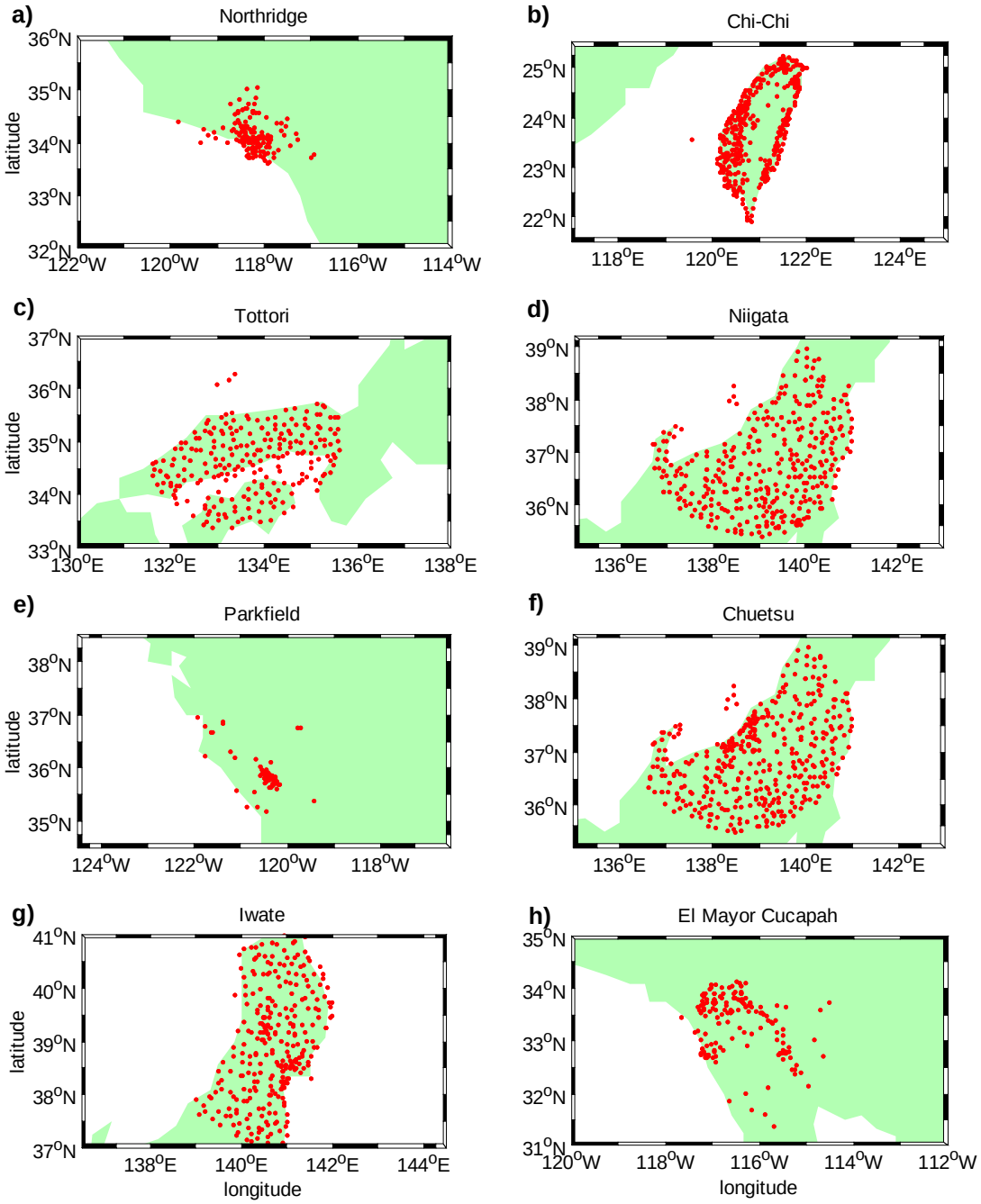

Figure 2: Locations of recordings from the eight considered earthquakes.

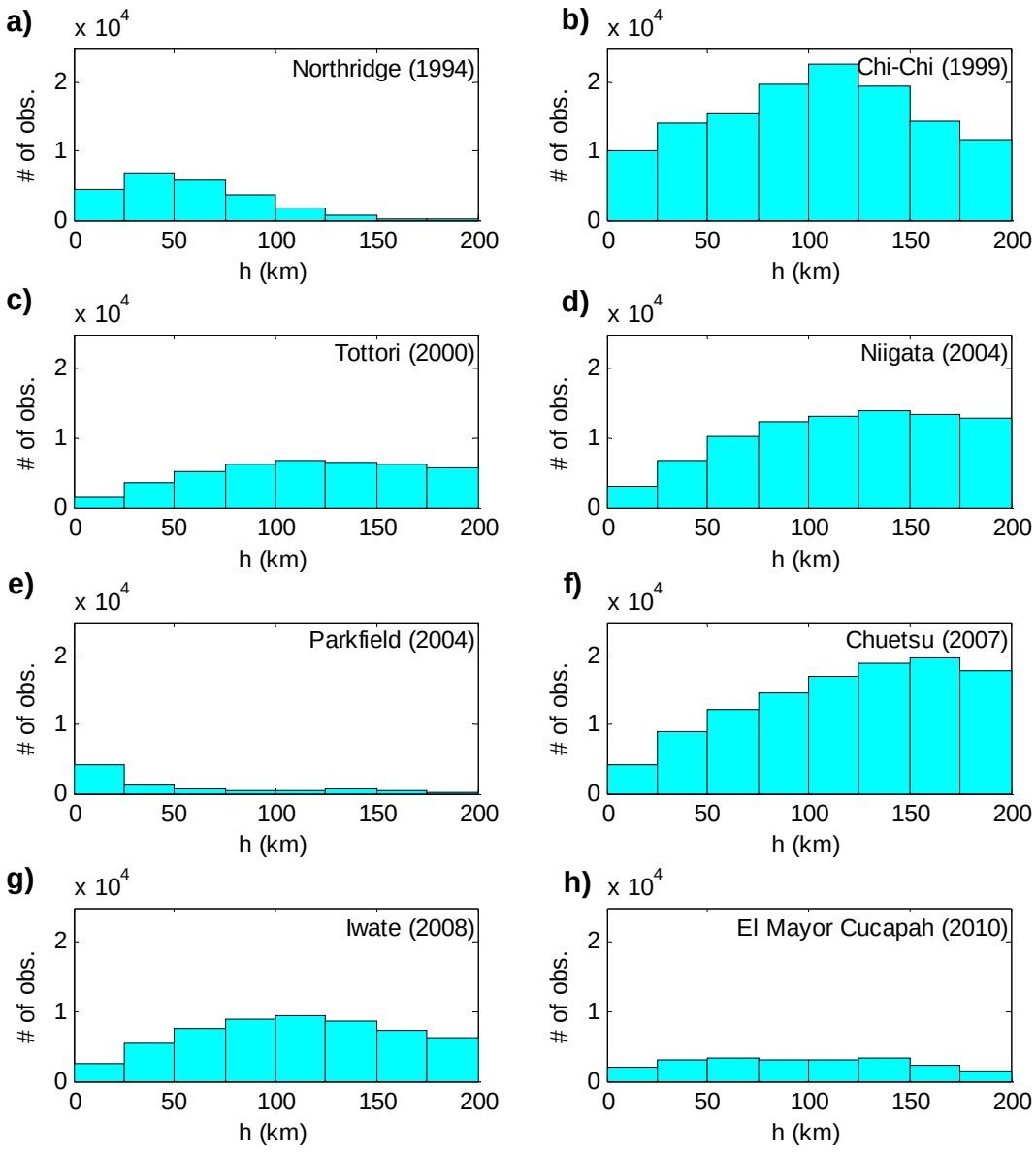

Figure 3: Histogram of the number of station pairs per separation distance for the eight considered earthquakes.

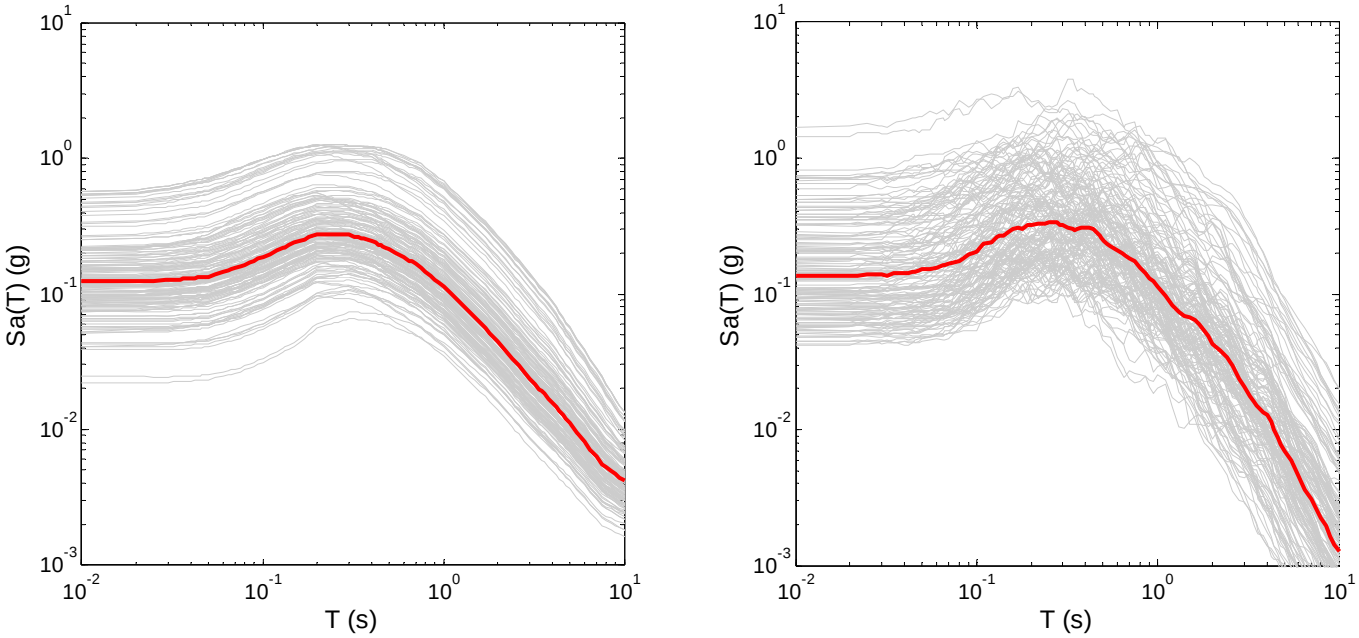

Figure 4: Northridge earthquake median predictions (left) and observations (right). Each thin line represents the median prediction (resp. observation) at a given site available in the Northridge earthquake recordings. The red thick line is the median of all predictions (resp. observations).

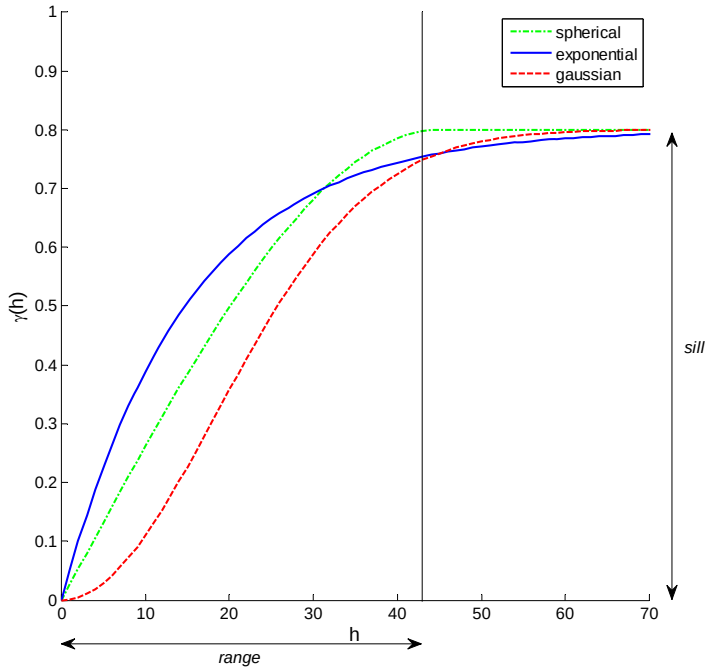

Figure 5: Spherical, exponential and Gaussian variograms with $S = 0.8$ and $R = 45$. For the spherical model, the range represents the distance at which 100% of the correlation is lost, whereas for the Gaussian and exponential variograms, it represents the distance at which 95% of the correlation is lost

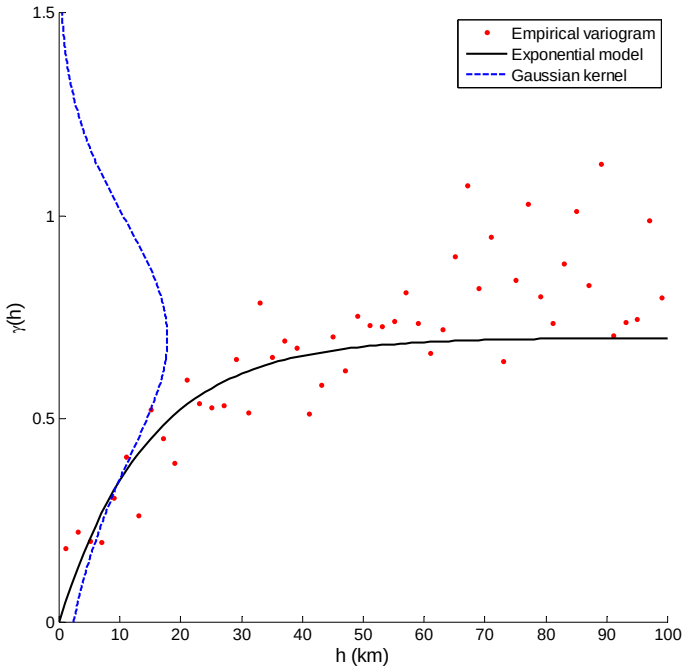

Figure 6: Direct cross-variogram fitting using the Northridge earthquake data for T_1=1 s and T_2=2.5 s. The sill is determined as the y-axis value at which the Gaussian kernel attains its maximum.

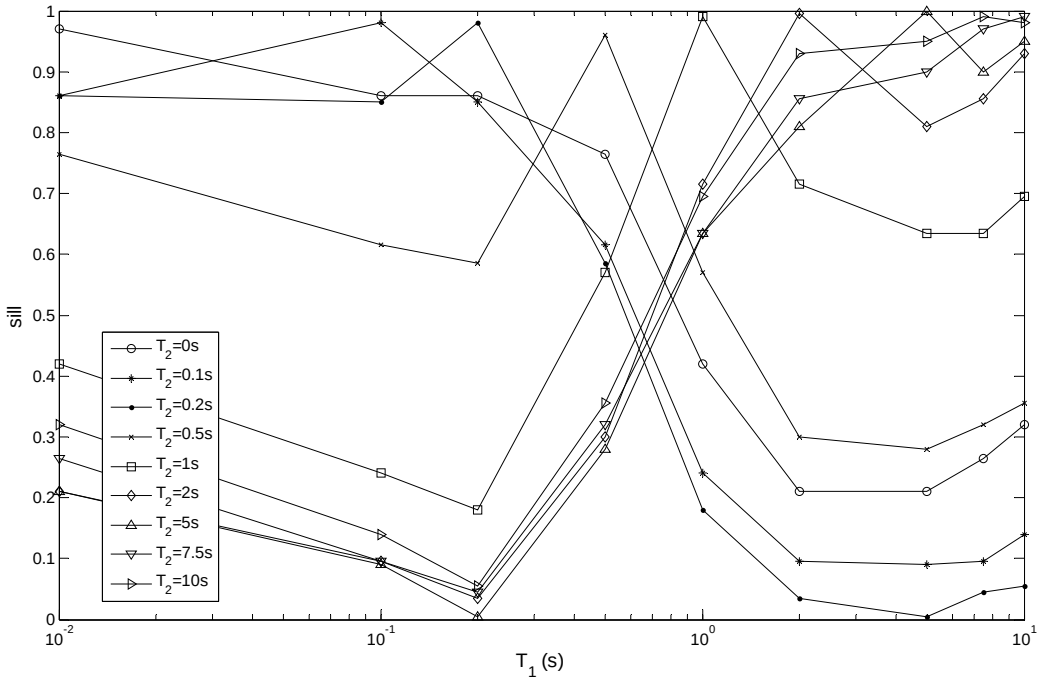

Figure 7: Sills of the crossvariograms obtained with the direct variogram fitting technique for the Northridge earthquake. Each line attains its maximum at $T_1=T_2$.

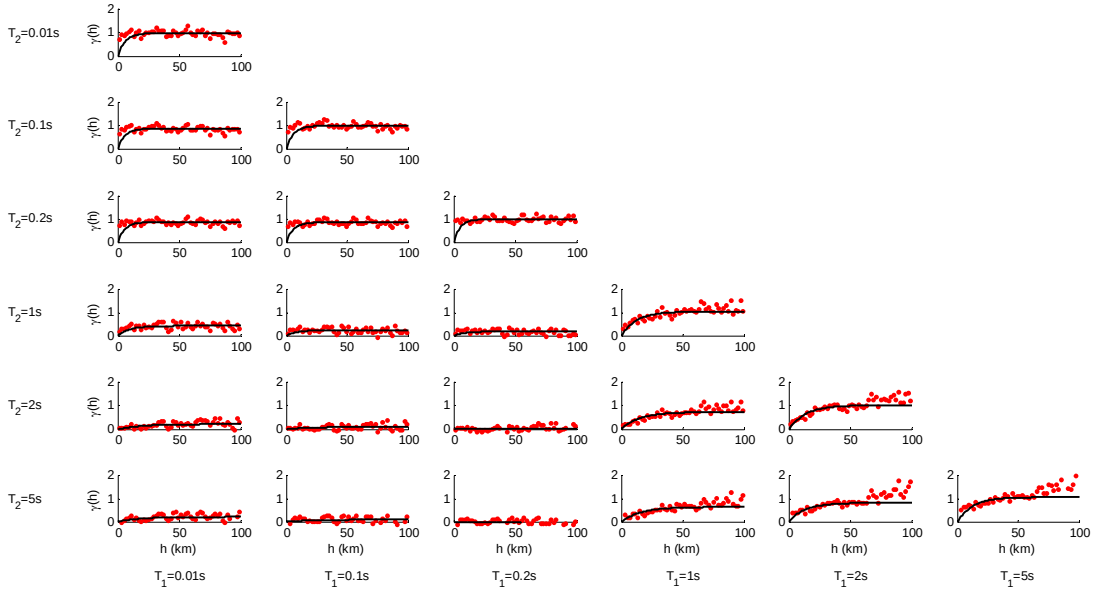

Figure 8: Direct cross-variogram fits for six pairs of periods from the Northridge earthquake data.

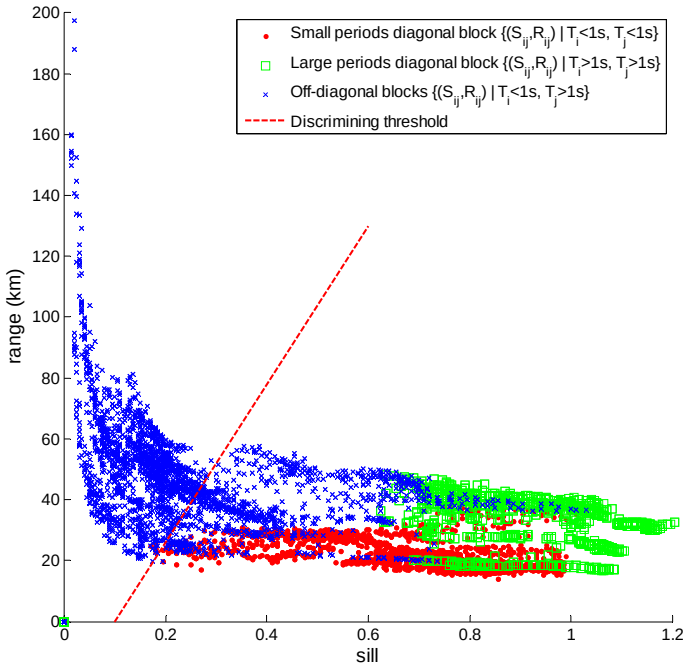

Figure 9: Sills and ranges for fitted cross-variograms from the Northridge earthquake data (105 periods were considered). The fitted data at the left of the dashed line will not be taken into account.

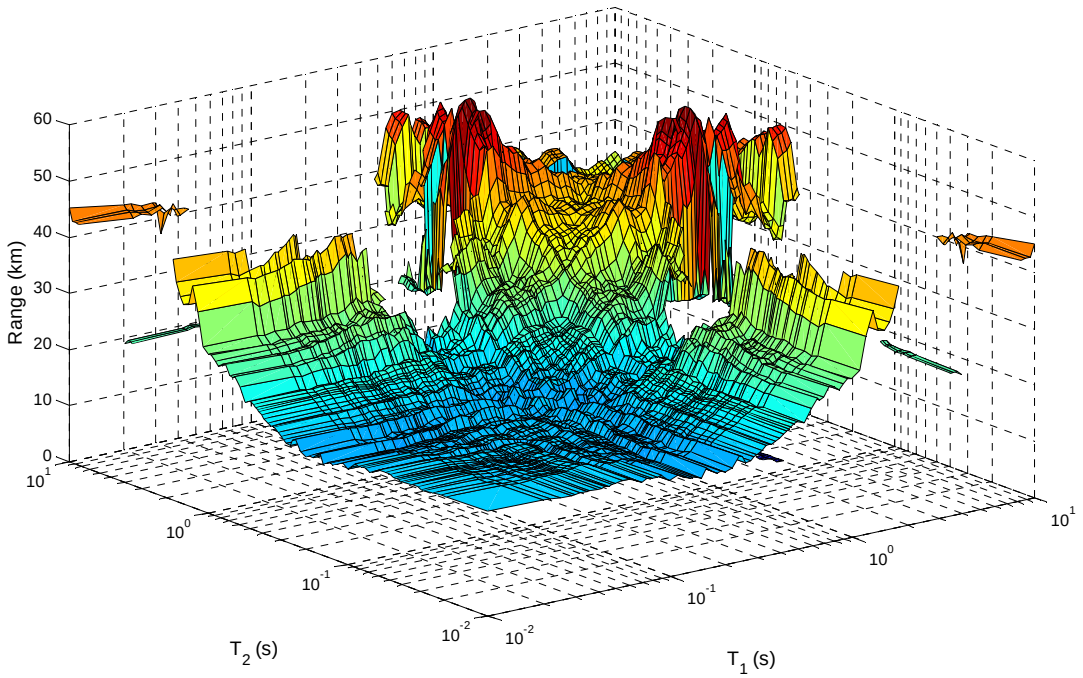

Figure 10: Filtered ranges of the cross-variograms from the Northridge earthquake data (105 periods were considered).

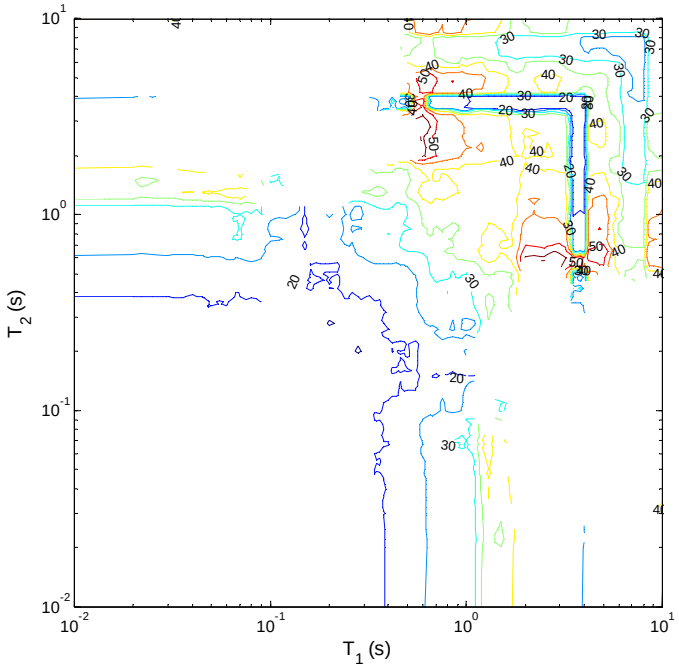

Figure 11: Contour plot of the filtered ranges of the cross-variograms from the Northridge earthquake data (105 periods were considered).

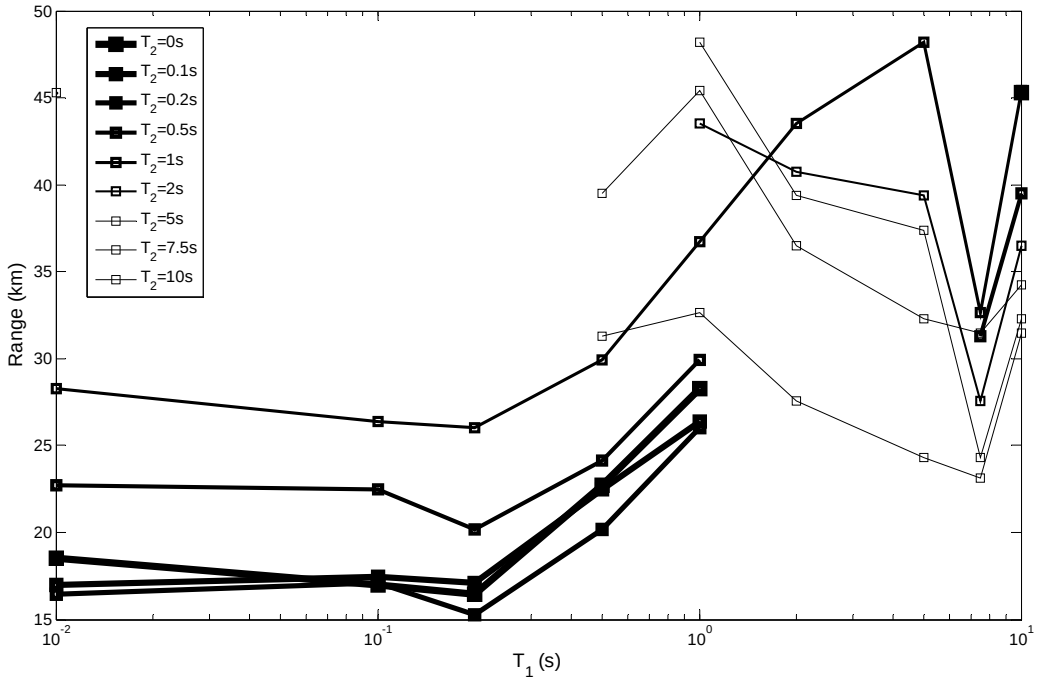

Figure 12: Filtered ranges of the cross-variograms from the Northridge earthquake data (9 periods were considered).

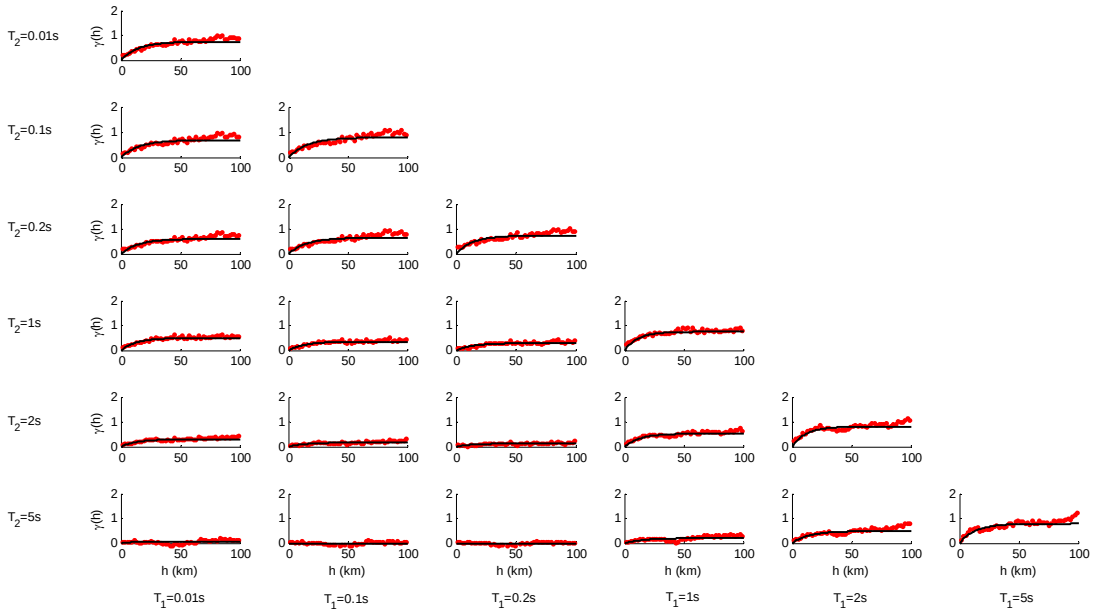

Figure 13: Direct cross-variogram fits for six pairs of periods from the Chi-Chi earthquake data.

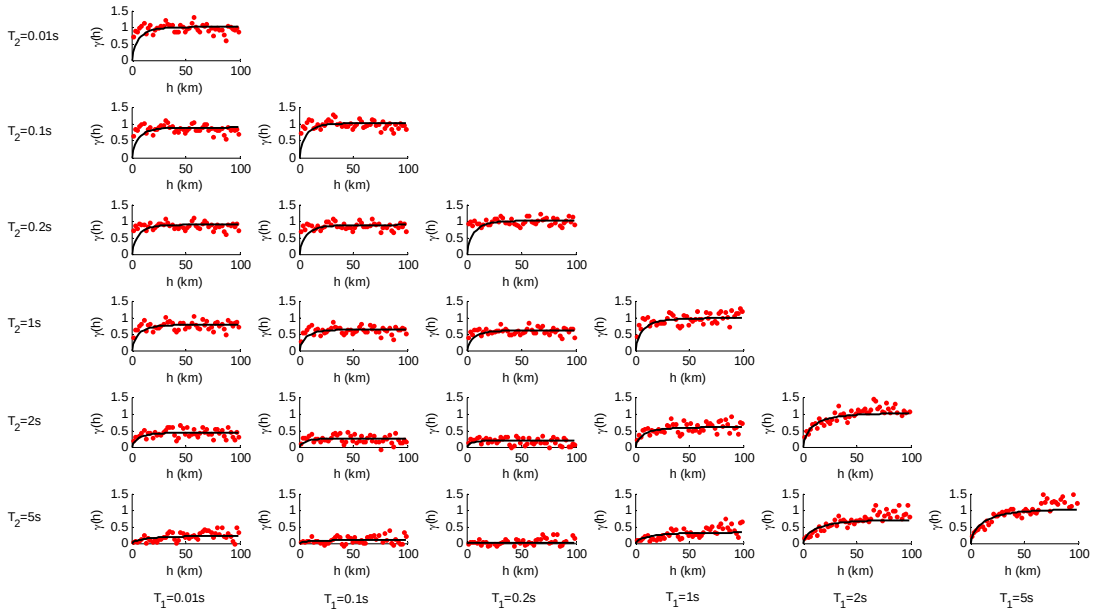

Figure 14: Northridge earthquake cross-variograms obtained using the Linear Model of Coregionalization.

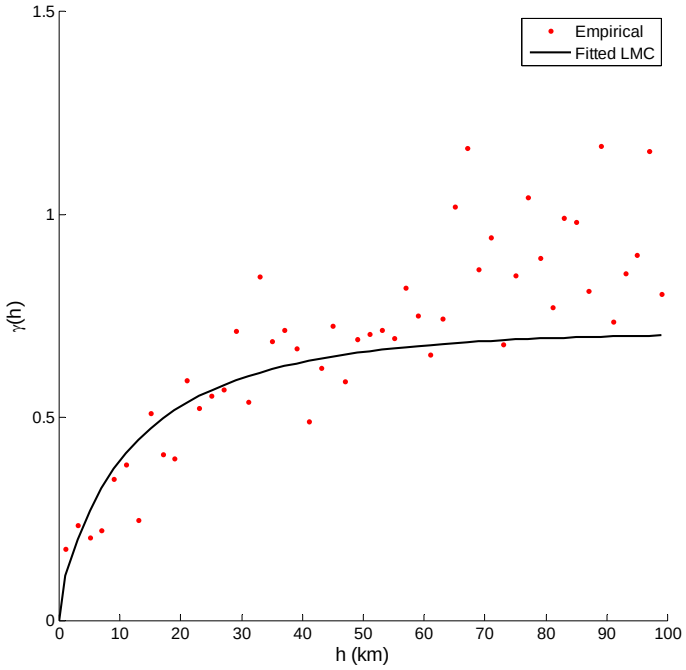

Figure 15: Northridge earthquake cross-variogram obtained using the Linear Model of Coregionalization for T_1=1 s and T_2=2 s.

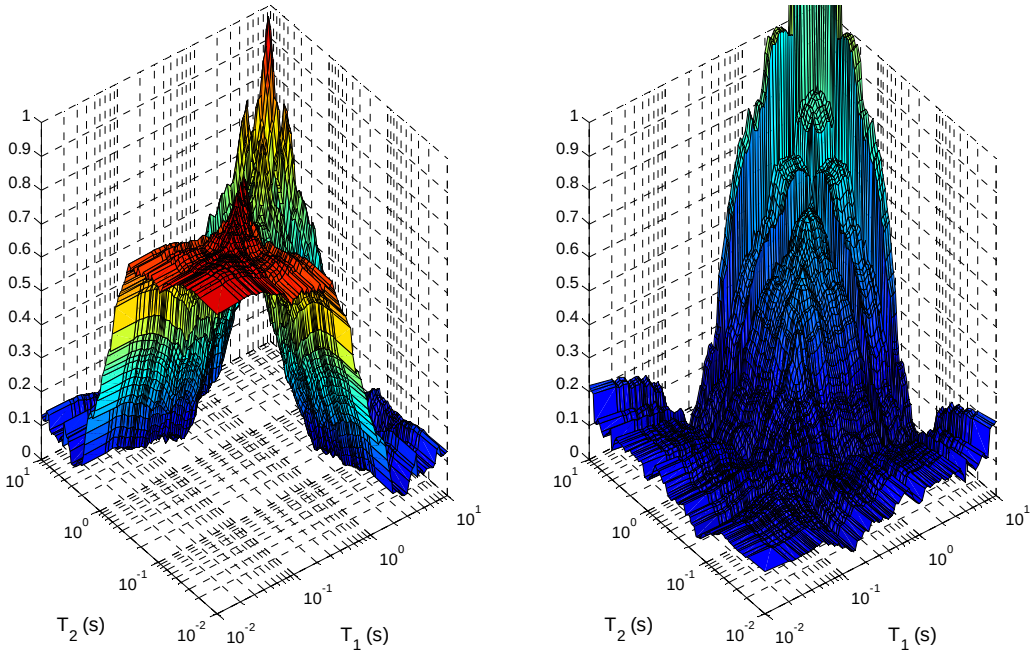

Figure 16: Coregionalization matrices obtained using the Northridge earthquake data. On the left, the short range coregionalization matrix $\mathbf{B}^1$; on the right, the long range coregionalization matrix $\mathbf{B}^2$.

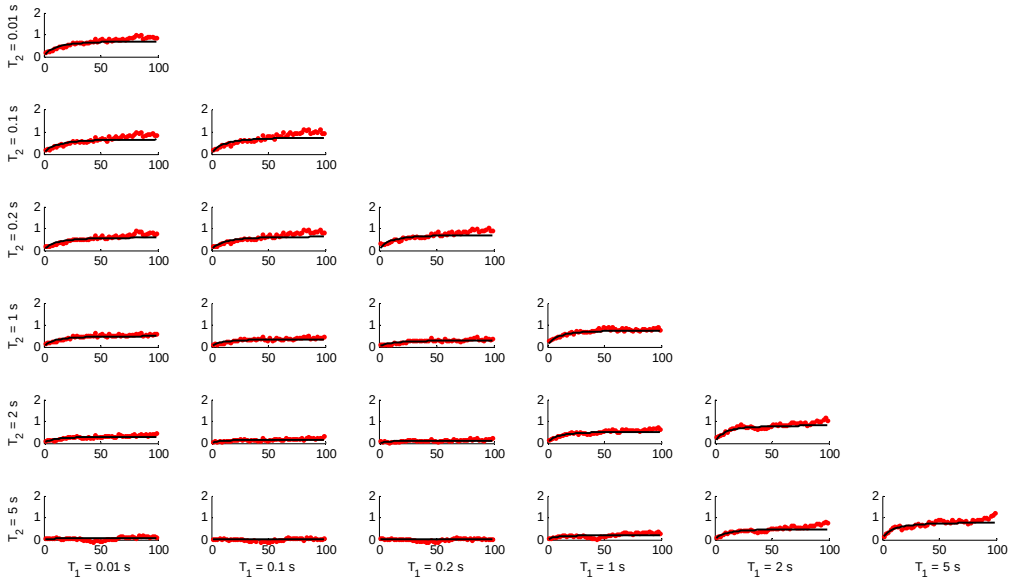

Figure 17: Chi-Chi earthquake cross-variograms obtained using the Linear Model of Coregionalization.

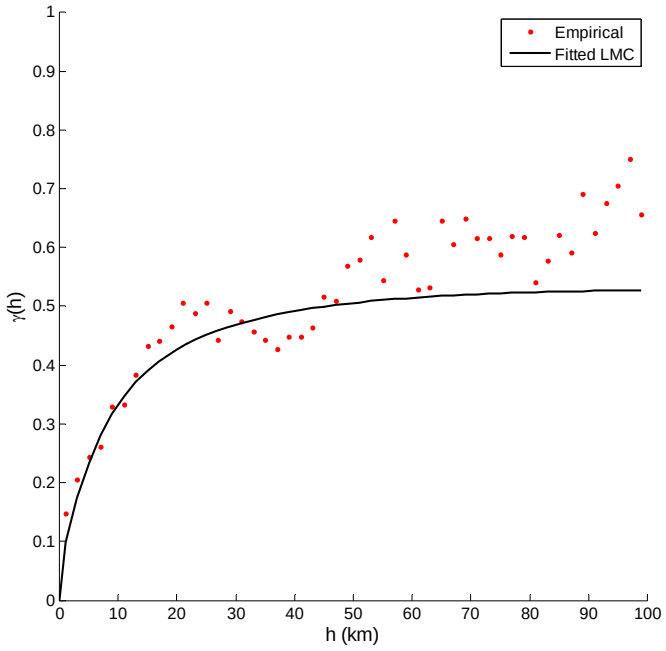

Figure 18: Chi-Chi earthquake cross-variogram obtained using the Linear Model of Coregionalization for T_1=1 s and T_2=2 s.

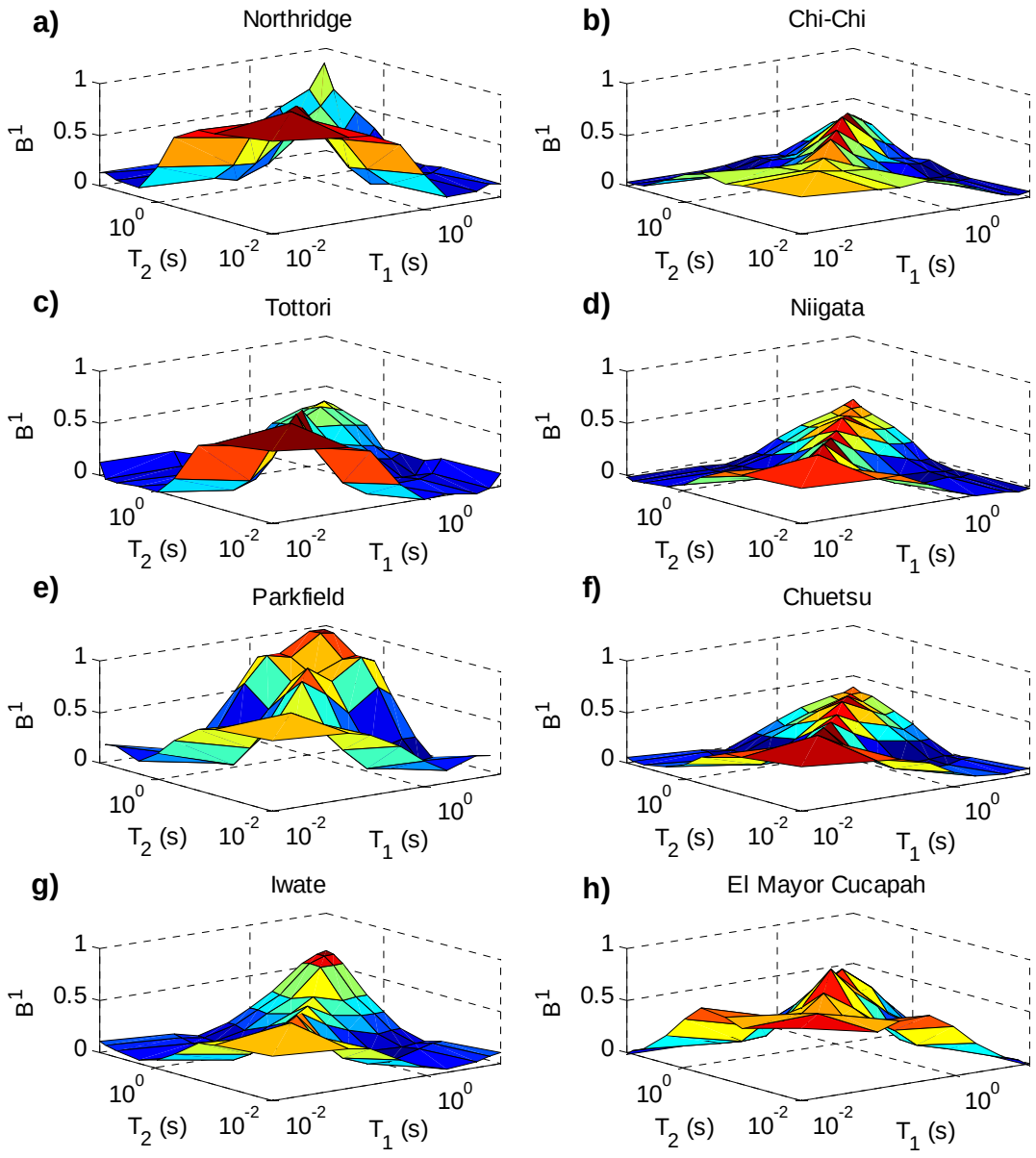

Figure 19: Short range coregionalization matrices for the eight investigated earthquakes, $\mathbf{B^1}$.

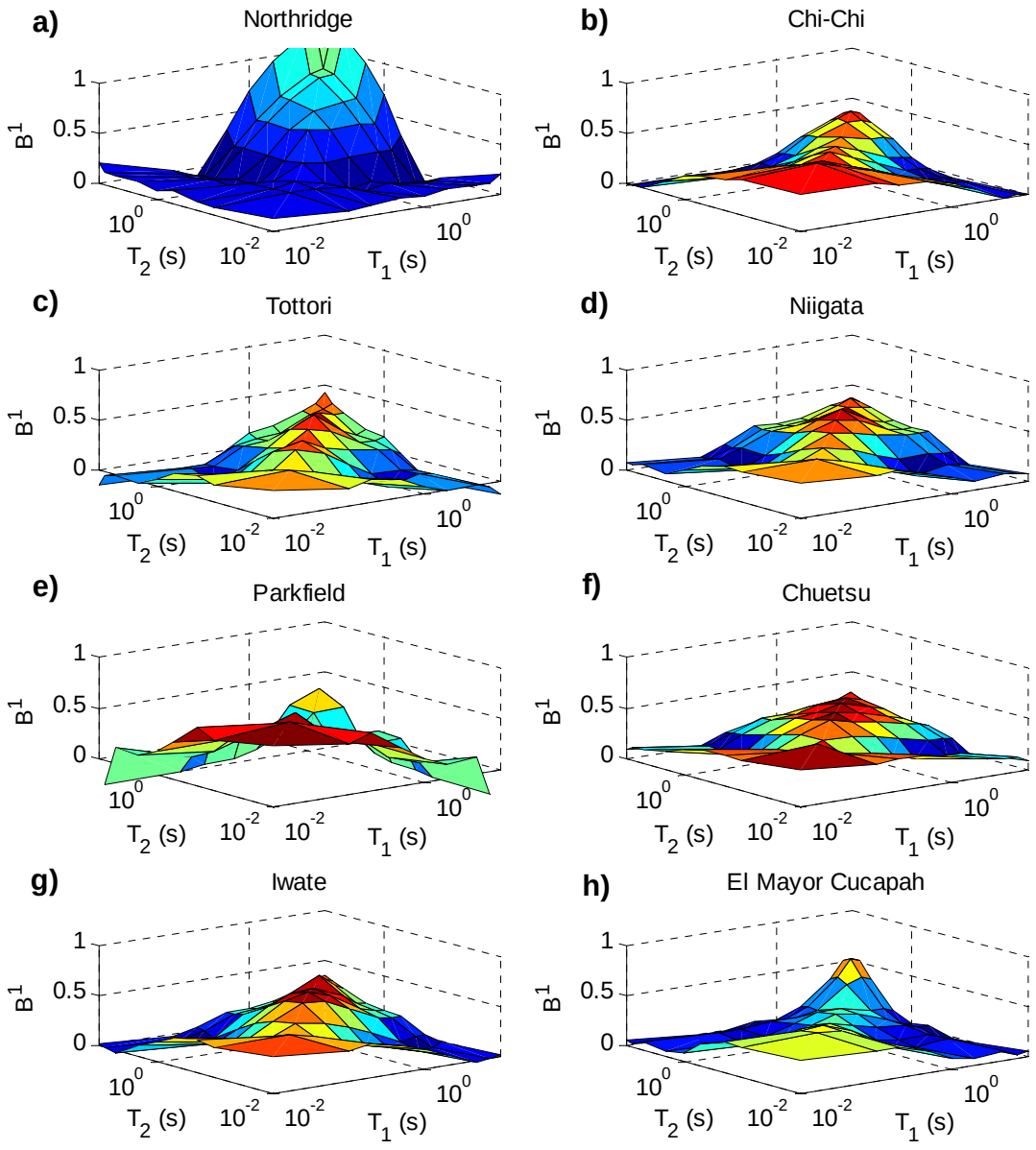

Figure 20: Long range coregionalization matrices for the eight investigated earthquakes, $\mathbf{B}^2$.

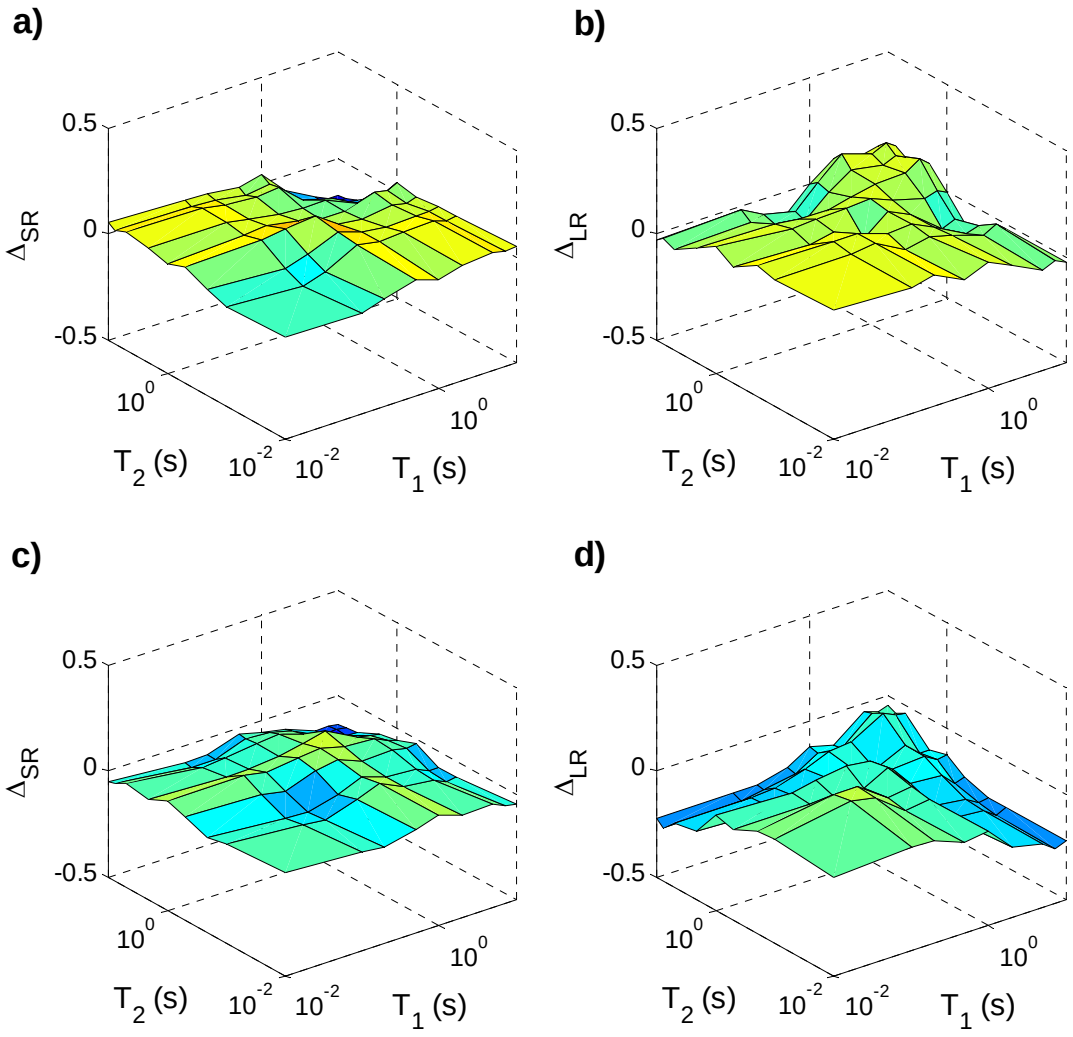

Figure 21: Each row represents a different realization of the differences between back-fitting of simulated epsilons using the covariance model (a) Simulation 1, Δ^{SR}; b) Simulation 1, Δ^{LR}; c) Simulation 2, Δ^{SR}; d) Simulation 2, Δ^{LR}).

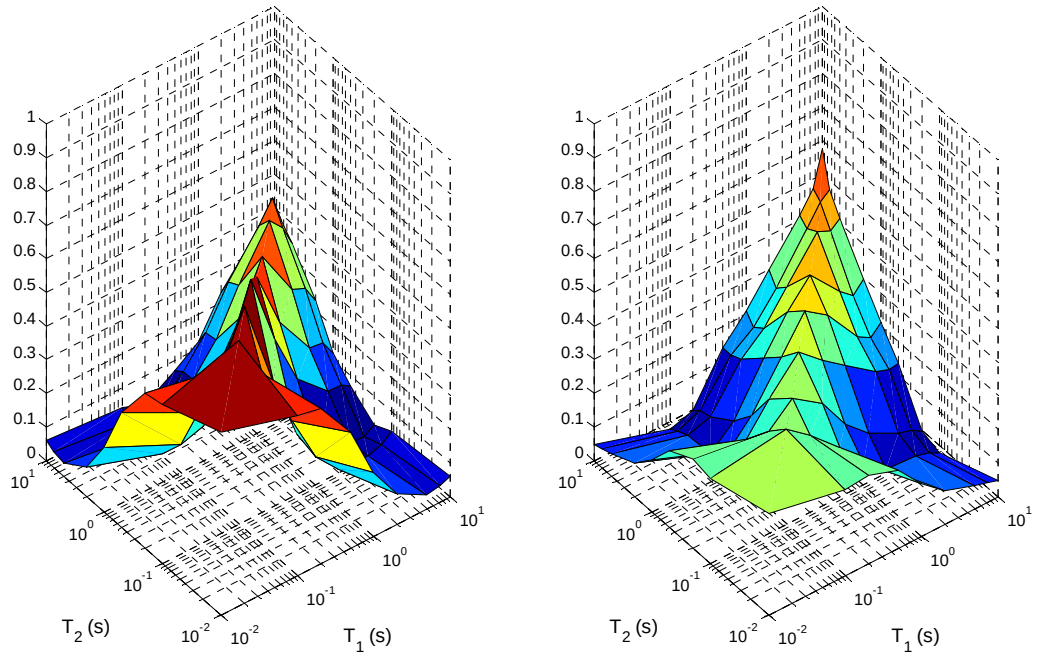

Figure 22: Average over the various earthquakes of the short range $\mathbf{B}^1$ (left) and long range $\mathbf{B}^2$ (right) coregionalization matrices

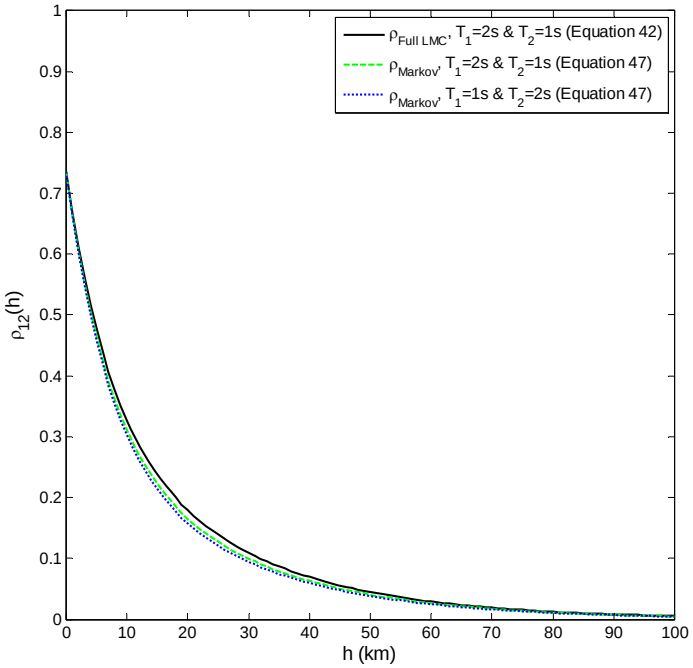

Figure 23: Comparison between the correlation coefficient obtained with the full linear model of coregionalization and the one computed from the reduced Markov-type model (1s and 2s)

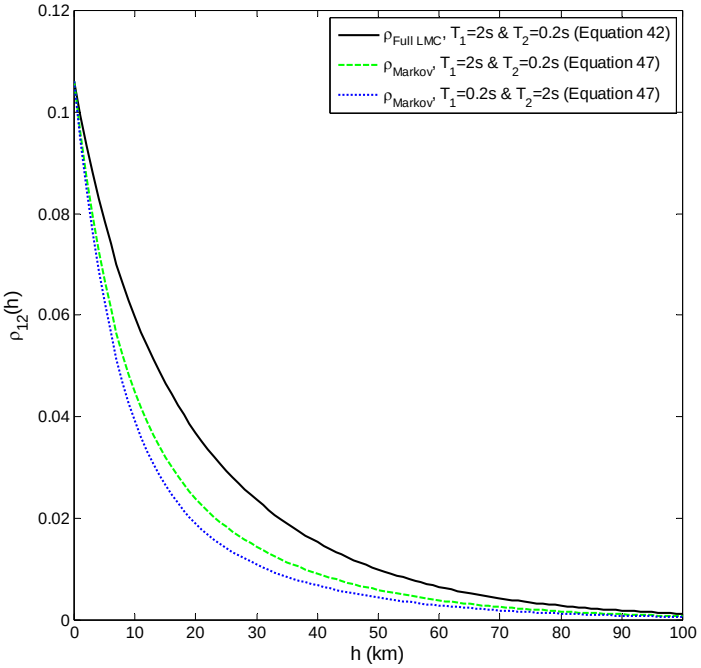

Figure 24: Comparison between the correlation coefficient obtained with the full linear model of coregionalization and the one computed from the reduced Markov-type model (2s and 0.2s)

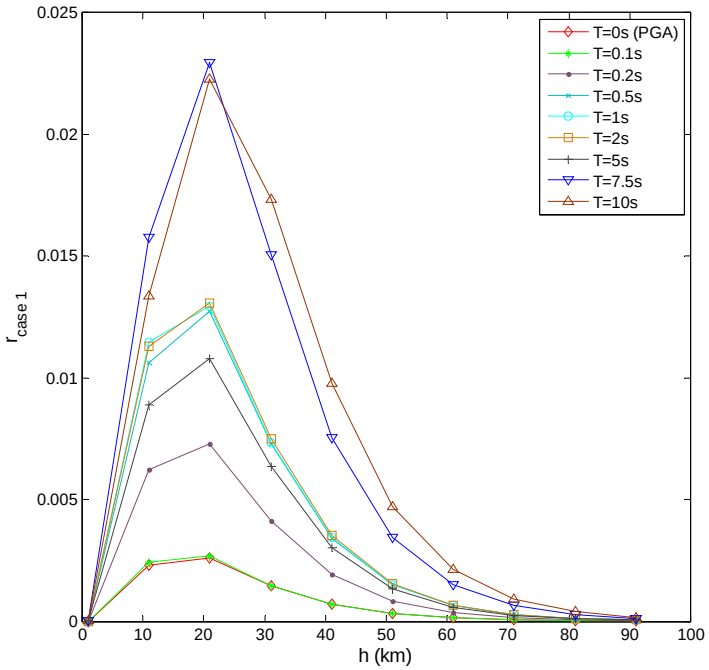

Figure 25: Fractional reduction in variance obtained by considering additional spectral periods at a remote site when computing the variance at a given site of interest (case 1)

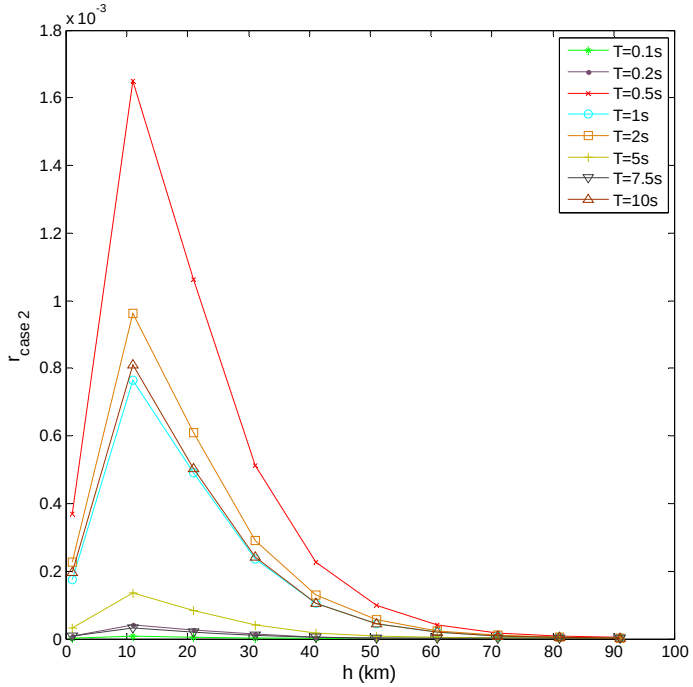

Figure 26: Fractional reduction in variance obtained by considering an additional site when computing the variance at a given site of interest (case 2)