PGA Semi-Empirical Correlation Models Based on European Data

S. Esposito, I. Iervolino & G. Manfredi
Dipartimento di Ingegneria Strutturale, Università degli Studi di Napoli Federico II, Naples, Italy.

Ohrid
30.08-03.09
14 ECEE 2010

ABSTRACT:
Spatial modeling of ground motion intensity measures (IMs) is required for seismic risk analysis of distributed systems as lifelines. In fact, when a spatially distributed infrastructure is of concern, site-specific hazard tools, which treat IMs at different locations independently, may be not adequate to assess accurately the seismic risk and the joint distribution of ground motion parameters at different sites may be needed. This work focuses on semi-empirical estimation of the correlation coefficient, as a function of inter-site separation distance, of peak ground acceleration (PGA). In particular, subsets of the European Strong-motion Database is employed to evaluate the intra-event correlation based on multiple earthquakes. Starting from the residuals, with respect to a ground motion prediction equation (GMPE), calibrated on the same dataset, a statistical analysis of correlation based on geostatistical tools is performed and semi-empirical models, for horizontal and vertical PGA components, are proposed. Finally, possible criteria for including spatial correlation in *regional* probabilistic seismic hazard analysis are discussed.

Keywords: ground-motion correlation, lifeline risk assessment, spatial variability, hazard analysis, geostatistics.

1.INTRODUCTION

Seismic risk analysis of distributed systems and infrastructures requires a different approach from the one commonly used for site-specific structures. One of the key issues, at least on the hazard side, is to account for the existence of a statistical correlation between ground motion intensity measures (IMs) at the different sites where the components of the system are located. Correlation models available in literature are usually based on dense observations of single earthquake records and depend uniquely on inter-site separation distance. For example, Boore *et al.* (2003) report a spatial correlation model for peak ground acceleration (PGA) from data of the 1994 Northridge earthquake. Wang and Takada (2005) computed different correlation models using peak ground velocity (PGV) observations of Japanese earthquakes occurred from 2000 to 2003, and the 1999 Chi-Chi earthquake. Goda and Hong (2008) computed correlation for PGA, PGV and pseudo-spectral acceleration (PSA) residuals based on the California dataset and the Chi-Chi earthquake. Jayaram and Baker (2009) refer to PGA and PSA from Northridge earthquake records, Chi-Chi, and Japanese earthquakes. This paper uses the European Strong-motion Database (ESD) to estimate a spatial correlation model for PGA residuals. Each earthquake in the dataset is characterized by a relatively small number of records, and the requirement for a large number of data for a good estimation of the correlation is coped with the option of using records from multiple events. The ground motion prediction equation (GMPE) with respect to which residuals are computed is that of Ambraseys *et al.* (2005a,b). A geostatistical analysis for the correlation assessment is performed and some aspects related to the estimation methodology are discussed. Finally, criteria for hazard analysis for spatially distributed systems are briefly presented.

2. DATASET

The estimation of the correlation model starts from the characterization of residuals of empirical data with respect to a GMPE. To this aim, part of the European Strong-motion Database (Ambraseys *et. al*, 2002) was considered (http://www.isesd.cv.ic.ac.uk/). In particular, a subset of the records used for the Ambraseys *et al*. (2005a,b) GMPE was used. In fact, the dataset is comprised[1] of 496 records from 88 events recorded between 1973 and 2003. Moment magnitude (M_w) ranges from 5 to 7.6 and epicentral distance (R_{epi}) from 3 to 137 km. Characteristics of the dataset, with respect to explanatory variables of the considered prediction equation, magnitude, distance and local site conditions, are shown in Fig. 1.

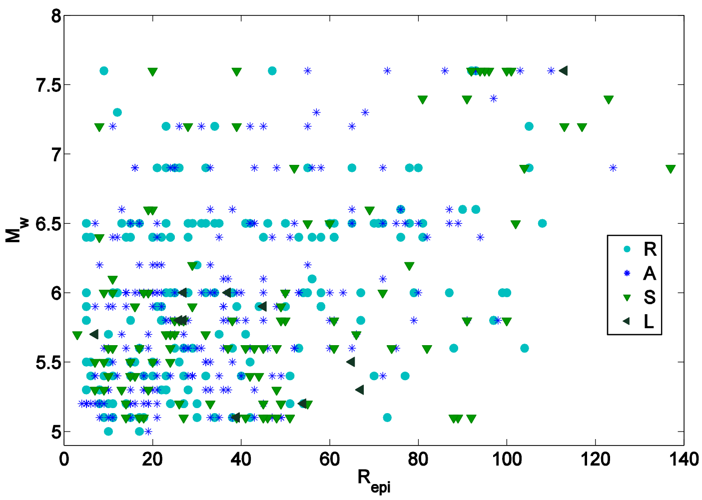

Figure 1. Strong-motion subset with respect to magnitude, distance and local site conditions: rock (R), stiff soil (A), soft soil (S), and very soft soil (L).

3. SPATIAL CORRELATION ESTIMATION

GEMPEs have been developed to predict mean of the logs of ground-motion intensities, and related variability, at a site *p* due to earthquake *j*; i.e.:

$$\log Y_{pj} = \overline{\log Y_{pj}} + \eta_j + \varepsilon_{pj} \tag{1}$$

Y_{pj} denotes the observed IM of interest; $\overline{Y_{pj}}$ is the mean of the logs; η_j denotes the inter-event residual, which is unique in a given earthquake; and ε_{ij} represents the intra-event heterogeneity of ground motion. ε_{pj} and η_j are usually assumed to be independent, normally distributed with zero mean and standard deviation σ_{intra} and σ_{inter}, respectively. These standard deviations, which are estimated while fitting the GMPE, may also be a function of the explanatory variables. For example, in Ambraseys *et al.* (2005a) the intra-event standard deviation is function of M_w.

Generally, $\log Y_{pj}$ is modeled as a Normal random variable; i.e. $\log Y_{pj} \sim N(\overline{\log Y_{pj}};\sigma)$ where $\sigma^2 = \sigma^2_{intra} + \sigma^2_{inter}$. For a single earthquake the joint distribution of $\log Y_{pj}$ for all sites of interest can be considered a multivariate normal distribution; i.e., a Gaussian random field (Jayaram and Baker, 2008), characterized by the following covariance matrix $\mathbf{\Sigma}$:

[1] Earthquakes considered here are those for which more than one record is available.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2_{inter} & \cdots & \sigma^2_{inter} \\ \vdots & \ddots & \vdots \\ \sigma^2_{inter} & \cdots & \sigma^2_{inter} \end{bmatrix} + \sigma^2_{intra} \begin{bmatrix} 1 & \cdots & \rho\left(h_{pq}\right) \\ \vdots & \ddots & \vdots \\ \rho\left(h_{pq}\right) & \cdots & 1 \end{bmatrix} \tag{2}$$

where the inter-event variability produces fully correlated residuals, and the intra-event variability produces correlated residuals (Malhotra, 2008). In fact, the correlation coefficient $\rho(h_{pq})$ denotes the spatial correlation between ε_{pj} and ε_{qj} at two different sites, p and q, distant by h_{pq}. In this model it is assumed that, under specific hypotheses, the spatial variability of intra-event residuals can be assumed as a function only of the inter-site separation distance, and as the separation distance increases, the correlation asyntotically tends to disappear (see Section 3.1).

Note also that GMPEs, usually, do not incorporate the correlation of intra-event residuals in regression analysis. The importance of accounting for spatial correlation in modeling has been discussed elsewhere (e.g., Cressie, 1993; Albert and McShane, 1995). Herein, traditional GMPE models, in which residuals are independent, are considered to define a spatial correlation function.

3.1 Geostatistical analysis of intra-event residuals

Spatial correlation modeling is performed through tools that are widely used in several fields characterized by data distributed in space. In fact, geostatistical estimation usually relies on semivariograms to characterize the correlation structure. However, its use is based on the assumption that the data follow a multivariate Gaussian distribution and may lead to inaccurate results if this assumption is violated (Cressie, 1993).

To understand the semivariogram tool, let consider $\mathbf{u_p} \in R^2$ the generic location in a 2-dimensional Euclidian space and suppose that the intra-event residual in a specific earthquake, $\varepsilon_j\left(\mathbf{u}_p\right) = \varepsilon_{pj}$, is a random variable in a way that it forms a multivariate Gaussian random field in a domain $S \in R^2$ (the region of interest). Considering two data locations $\mathbf{u}_p$ and $\mathbf{u}_q$, the semivariogram can be defined as in Eqn. 3.

$$\gamma_j\left(\mathbf{u}_p, \mathbf{u}_q\right) = \frac{1}{2} Var\left[\varepsilon_j\left(\mathbf{u}_p\right) - \varepsilon_j\left(\mathbf{u}_q\right)\right] \tag{3}$$

Under the hypothesis of a second-order stationary random field, the two points covariance function, $C_j\left(\mathbf{h}\right) = Cov\left[\varepsilon_j\left(\mathbf{u}+\mathbf{h}\right) - \varepsilon_j\left(\mathbf{u}\right)\right]$, depends only on the separation vector ($\mathbf{h}$) between $\mathbf{u}_p$ and $\mathbf{u}_q$ and the reference to a particular location $\mathbf{u}$ can be dropped. Moreover if the random field is isotropic, the covariance function (and so the semivariogram) depends only on the separation distance, $h = \|\mathbf{h}\|$, and assuming the semivariogram to be invariant through earthquakes, Eqn. 3 becomes Eqn. 4 (see Cressie, 1993, for more details).

$$\gamma\left(h\right) = \frac{1}{2} Var\left[\varepsilon\left(\mathbf{u}+h\right) - \varepsilon\left(\mathbf{u}\right)\right] = Var\left(\varepsilon\right) \cdot \left[1 - \rho\left(h\right)\right] \tag{4}$$

In Eqn. 4 $\rho(h)$ denotes the correlation coefficient between $\varepsilon(\mathbf{u}+h)$ and $\varepsilon(\mathbf{u})$. Therefore, for an isotropic *homogenous* random field, it is possible to model the covariance structure through the estimation of the semivariogram, using suitable functions that satisfy some conditions in order to guarantee a *valid* correlation structures (Goovartes, 1997).

The steps needed to model spatial correlation are:

(1) to compute the empirical semivariogram;
(2) to choose a model among the family of valid functions;
(3) to estimate the semivariogram parameters by fitting the model on the empirical data.

These steps are discussed briefly in the following sections and applied to the dataset described.

3.1.1 Empirical semivariogram estimation

The classical estimator of the semivariogram is the *method-of-moments* estimator proposed by Matheron (1962), which is defined as follows.:

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

where $N(\mathbf{h})$ denotes the set of pairs of sites separated by a similar lag $\mathbf{h}$ and $|N(\mathbf{h})|$ is the cardinal of $N(\mathbf{h})$. (Note that in Eqn. 5 and Eqn. 6 the subscripts are not reported for simplicity.)

If the data locations are not regularly spaced, as always happens for earthquake records, the variogram can be computed defining *tolerance* bins around **h.** If it is assumed that the random field is isotropic, only $|N(h)|$ matters for the data binning and semivariogram computation.

The selection of distance bins has important effects on the empirical semivariogram and it is similar to the definition of classes in histograms. If the lag size is too large, correlation at short distances may be masked. If the lag size is too small, empty bins or small size samples within bins may not lead to representative averages. A rule of thumb is to make the maximum bin size as half of the maximum distance between pairs in the dataset, and set the number of bins so that there are more than 30 pairs per bin (Journel and Huijbregts, 1978).

The *method-of-moments* estimator is unbiased; however it is <<badly affected by atypical observations>> (Cressie, 1993). Therefore, Cressie and Hawkins (1980) propose a more robust estimator (in statistical sense, that is, less sensitive to outliers), Eqn. 6.

$$\hat{\gamma}(h) = \frac{1}{2}\left\{\left[\frac{1}{\left|N(h)\right|}\sum_{N(h)}\left[\varepsilon(\mathbf{u}+h)-\varepsilon(\mathbf{u})\right]^2\right]^4 \Bigg/ \left(0.457 + \frac{0.494}{\left|N(h)\right|}\right)\right\} \tag{6}$$

3.1.2 Model identification and fit

The interpretation of experimental semivariograms consists in the identification of a model among the family of functions able to capture and emulate its trend. The three basic stationary and isotropic models are: exponential, spherical, and Gaussian. In particular, the exponential model takes the following form:

$$\gamma(h) = c_0 + c_e \cdot \left[1 - e^{(-3h/b)}\right] \tag{7}$$

where c_0 is defined *nugget*, c_e is the *sill*, and b is the *practical range*. The sill of a semivariogram equals the population variance of the random field and the *range* is defined as the inter-site distance at which $\gamma(h)$ equals the sill. For exponential and Gaussian model, the sill is asymptotic and it is possible to define a *practical range* as the separation distance at which γ(h) equals 95% of the sill. From Eqn. 5 it should be $\gamma(0)=0$, then c_o is the limit value of $\gamma(h)$ when $h \to 0$; i.e., the *nugget* effect (or micro-scale process) and is due to variations at distances smaller than the sampling interval and measurement errors, which cause a discontinuity at the origin (Matheron, 1962).

Several goodness-of-fit criteria for finding the best parametric model have been proposed in geostatistical literature. However, several studies in the earthquake field fit a model “by eye” or via a *trial and error* approach (e.g., Wang and Takada, 2005; Goda and Hong, 2009). In order to give more importance to model the semivariogram structure well at short site-to-site distances, as suggested in Jayaram and Baker (2009), experimental semivariograms are fitted manually in this work.

4. RESULTS AND DISCUSSION

To compute the empirical estimation of the semivariogram, characterization of recorded data residuals is needed. Instead of considering intra-event residuals it is possible to use standardized intra-event residuals, computed for a single earthquake as follow:

$$\varepsilon_p * = \frac{\varepsilon_p}{\sigma_p} \tag{8}$$

where σ_p is the standard deviation of the intra-event residual at the site *p*, then Eqn. 4 becomes Eqn. 9.

$$\hat{\gamma}(h) = 1 - \hat{\rho}(h) \tag{9}$$

Eqn. 9, in which the superscript represents estimated quantities, shows that the use of standardized residuals let not to estimate the sill because the sill should equal to one. This is valid if the standardization has been done with the true variance. With real data, the sample variance and the standard deviation provided by the GMPE can be used to obtain standardized residuals[2]. The variance provided by the GMPE is used in this work.
In the following, to compute the semivariogram it is also assumed that the random field is isotropic. This hypothesis should be investigated through an estimation of directional semivariograms; i.e., also depending on the direction of the inter-site distance. Nonetheless, past research has shown that the hypothesis of isotropic random field is reasonable (Wang and Takada, 2005; Jayaram and Baker, 2009) also within specific earthquakes.
Another important issue concerns how many data are used in the estimation of empirical semivariograms. Correlation models available in literature are based on dense observations of single event earthquake records, but for ESD dataset no dense observations of single earthquake were available. In this case, the requirement for a relatively large number of data to estimate the empirical semivariogram is coped with the option of using records from multiple events and regions. This means to assume the same isotropic semivariogram with the same parameters (i.e. the same spatial covariance structure for all events), and may be a strong assumption. In fact, since modern GMPEs are usually characterized by a variance that is not constant over the events because it is function of different explanatory variables, this could be in contrast with a constant covariance structure assumption; this issue has not been investigated in this work, but it should be taken into account especially when a GMPE incorporating a covariance model has to be estimated.
Standardized intra-event residuals enable to have a constant variance among data, which is a feature of a homogenous random field. Considering the isotropic hypothesis reasonable and a common model for different earthquakes, each of which characterized by n_j records, the experimental semivariogram becomes that of Eqn. 10 and Eqn. 11 (which states that individual events are treated separately and then used to fit a unique model).

$$\hat{\gamma}(h) = \frac{1}{2\left|N(h)\right|} \sum_{N(h)} \left[\varepsilon^*_{pj} - \varepsilon^*_{qj} \right]^2 \tag{10}$$

$$N(h) = \left\{ \left(j, \varepsilon^*_{pj}, \varepsilon^*_{qj} \right) : \left\| \varepsilon^*_{pj} - \varepsilon^*_{qj} \right\| = h; \quad p, q = 1, \ldots n_j; \; j = 1, \ldots k \right\} \tag{11}$$

These hypotheses avoid the possibility of estimating the effect, if any, of *near-source directivity* on the correlation. This seems acceptable given the results of other researchers (Jayaram and Baker, 2009).

[2] It is usual to use the sample variance as an estimator of the sill for the experimental semivariogram, but this may be improper in some circumstances, see Barnes (1991) for a discussion.

4.1 Proposed models

To compute the experimental semivariogram, the first step is to determine the *tolerance* region because the data are not regularly spaced. The bin width must to be chosen in order to have a reasonable number of sites pairs in the bins (at least 30) and a stable trend of semivariogram. In this case the experimental semivariograms are obtained using a width of 4km. Both estimators (classical and robust) were used without finding any considerable differences in the shape of the fitted variogram. To determine the model and model's parameter a manual fitting was preferred. Of the three basic models, the exponential model was found the best providing the best fit at the small separation distances.
The estimation first and the fitting then were performed for PGA and both: (i) the largest horizontal component (coherently with the GMPE used); (ii) and the vertical component. Assuming that there is no nugget effect, the only parameter to estimate is the range b that results equal to 12 km for horizontal component and 18 km for vertical component, as shown in Fig. 4.

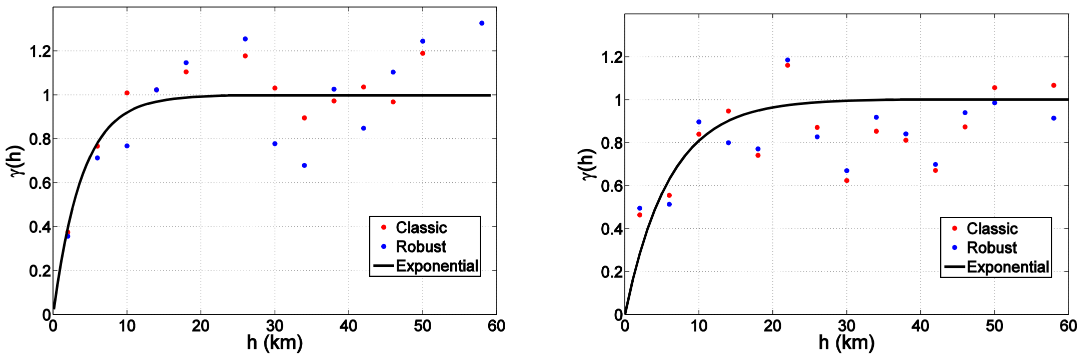

Figure 4. PGA semivariogram model for the largest horizontal PGA component (left), and the vertical component (right).

5. REGIONAL HAZARD CRITERIA

Developed correlation models can be used, for example, to obtain the exceedance probability of ground motion peak accelerations in a region and in a time interval of interest. The hazard integral of Eqn. 12 provides the annual rate of exceedance of a ground motion intensity measure or IM (e.g., PGA) jointly in a region. In Eqn. 12 $f_{M,R}(m,r)$ is the joint distribution of magnitude and distance referred to a particular seismic source, ν is the rate of occurrence of earthquakes on it, and $P\left[IM_1 > im^*,........,IM_n > im^* \mid M,R\right]$ is the conditional probability that *im** is exceeded at the *n* sites in which the region is discretized.

$$\lambda = \nu \int_M \int_R P\left[IM_1 > im^*,........,IM_n > im^* \mid M,R\right] f_{M,R}(m,r)\, dm\, dr \qquad (12)$$

As an example, using as source the Paganica fault on which L'Aquila (central Italy) 2009 was originated, the PGA hazard was computed for a characteristic earthquake of M_w 6.3 and occurrence rate of $\nu = 1/750$ (Pace *et al.*, 2006) considering as exceedance region different areas (A*) between 2.5% to 25% of a zone of 2500 km^2 around the fault.
To compute the hazard curve, the conditional joint distribution has been estimated simulating spatially correlated random fields of horizontal peak ground acceleration using the Ambraseys *et al.* (2005a) GMPE under the assumption that all the sites have the same local site conditions (rock).

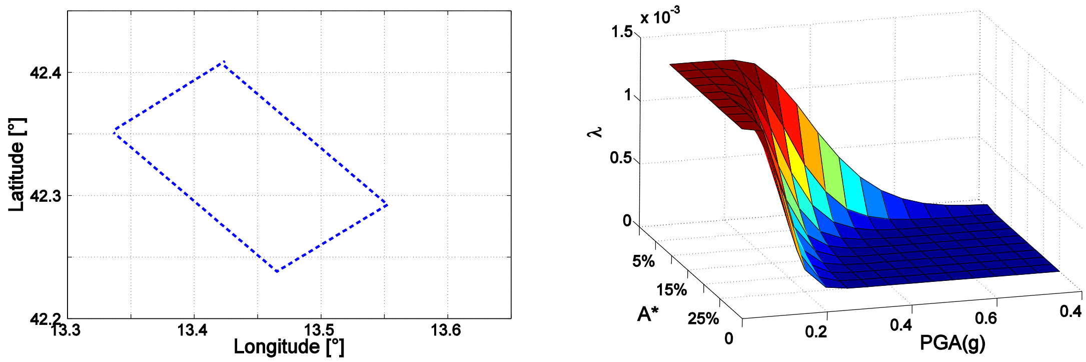

Figure 5. Paganica fault's projection (left), and the *regional* hazard surface (right).

6. CONCLUSIONS

This paper focuses on the semi-empirical estimation of spatial correlation of peak ground acceleration using the European Strong-motion Database (ESD). Geostatistical tools, in particular semivariograms, have been used to quantify the spatial variability of intra-event residuals. The main aspects related to the estimation of a spatial correlation model have been considered and discussed.
The correlation model has been estimated under the hypotheses of isotropic and homogenous random fields. Standardized intra-event residuals from multiple earthquakes have been used to compute experimental semivariograms that are a function of the inter-site distance. The advantage of using an unique model fitted with data from multiple earthquakes is to have the possibility to work with a large number of observations; nevertheless, the assumption of assuming a common correlation model should be further investigated.
Exponential correlation functions have been identified and *practical ranges* (the inter-site distance at which the correlation is practically zero) of 12 km and 18 km have been estimated for horizontal and vertical components of PGA, respectively.
The proposed methodology provide the basis to obtain the joint distribution of IMs that can be used in hazard applications for seismic risk assessment of distributed systems.

ACKNOWLEDGEMENTS

This work was financially supported by AMRA scarl (http://www.amra.unina.it) under the frame of the SYNER-G project (seventh framework programme of the European Community for research, technological development and demonstration activities ; project contract no. 244061).

REFERENCES

Albert P.S and Mc Shane L.M. (1995) A Generalizaed Esimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data. Biometrics **51**, 627-638

Ambraseys N. N., Douglas J., Sarma S. K. and Smit P.M. (2005a). Equations for the Estimation of Strong Ground Motions from Shallow Crustal Earthquakes Using Data from Europe and the Middle East: Horizontal Peak Ground Acceleration and Spectral Acceleration. *Bulletin of Earthquake Engineering* **3:1**, 1-53.

Ambraseys N. N., Douglas J., Sarma S. K. and Smit P.M. (2005b). Equations for the Estimation of Strong Ground Motions from Shallow Crustal Earthquake Using Data from Europe and the Middle East: Vertical Peak Ground Acceleration and Spectral Acceleration. *Bulletin of Earthquake Engineering* **3:1**, 55-73.

Ambraseys, N. N., Smit, P., Sigbjornsson, R., Suhadolc, P. and Margaris, B. (2002). Internet-Site for European Strong-Motion Data, European Commission, Research-Directorate General, Environment and Climate Programme.

Barnes R. J. (1991). The Variogram Sill and the Sample Variance. *Mathematical Geology,* **23:4**, 673-678.

Bazzurro P., Park J., Tothong P. and Jayaram N. (2009). Effects of spatial correlation of ground motion parameters for multi-site seismic risk assessment: Collaborative Research with Stanford University and

AIR. *Technical report*, report for U.S. Geological Survey National Earthquake Hazards reduction program (NEHRP) external Research program Awards 07HQGR0031. Available from: http://earthquake.usgs.gov/research7external/reports/07HQGR0031.pdf

Boore D. M., Gibbs J. F., Joyner W. B., Tinsley J. C. and Ponti D. J. (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.

Cressie N. (1993). Statistical for Spatial Data. Revised edition. Wiley, New York, USA.

Cressie N. and Hawkins D.M.(1980). Robust estimation of variogram. *Mathematical Geology* **12:2**, 115-125.

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

Goovartes P. (1997). Geostatistics for Natural Resources Evaluation. Oxford University Press, Oxford, New York.

Jayaram N. and Baker J.W. (2008). Statistical Tests of the Joint Distribution of Spectral Acceleration Values. *Bulletin of the Seismological Society of America* **98:5**, 2231-2243.

Jayaram N. and Baker J.W. (2009). Correlation model for spatially distributed ground-motion intensities. *Earthquake Engineering and Structural Dynamics* **38:15**, 1687–1708.

Journel and Huijbregts, (1978). Mining Geostatistics. Academic Press, London.

Malhotra P. (2008). Seismic Design Loads from Site-Specific and Aggregate Hazard Analyses. *Bulletin of the Seismological Society of America* **98:4**, 1849-1862.

Matheron G. (1962). *Traité de géostatistique appliquée.* Editions Technip, Paris, France.

Pace B., Perruzza L., La Vecchia G. and Boncio P. (2006). Layered Seismogenic Source Model and Probabilistic Seismic-Hazard Analyses in Central Italy. *Bulletin of the Seismological Society of America* **96:1**, 107-132.

Wang M. and Takada T. (2005). Macrospatial correlation model of seismic ground motions. *Earthquake Spectra* **21:4**, 1137-1156.