MOX-Report No. 67/2020

Multi-Source Geographically Weighted Regression for Regionalized Ground-Motion Models

Caramenti, L.; Menafoglio, A.; Sgobba, S.; Lanzano, G.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Multi-Source Geographically Weighted Regression for Regionalized Ground-Motion Models

Luca Caramenti[1], Alessandra Menafoglio[1*], Sara Sgobba[2], Giovanni Lanzano[2]

[1]MOX, Department of Mathematics, Politecnico di Milano, Italy

[2]INGV Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano, Italy

[*]alessandra.menafoglio@polimi.it

Abstract

This work proposes a novel approach to the calibration of regionalized regression models, with particular reference to ground-motion models (GMMs), which are key for probabilistic seismic hazard analysis and earthquake engineering applications. A novel methodology, named multi-source geographically-weighted regression (MS-GWR), is developed, allowing one to *(i)* estimate regionalized regression models depending on multiple sources of non-stationarity (such as site- and event-dependent non-stationarities in GMMs), and *(ii)* make inference on the significance and stationarity of the regression coefficients. Unlike previous approaches to the problem, the proposed framework is fully non-parametric, the inference being based on a permutation scheme. MS-GWR is here used to calibrate a new regionalized ground-motion model for predicting peak ground acceleration in Italy, based on a large scale database of waveforms and metadata made available by the Italian Institute for Geophysics and Vulcanology (INGV).

Keywords: Geographically-weighted regression; ground motion models; peak-ground acceleration; seismic risk analysis

1 Introduction

Seismic risk analysis and earthquake engineering applications use empirical ground motion models (GMMs) to predict the intensity level of ground shaking caused by an earthquake event at a site. These models quantify the expected median level of a ground motion parameter, i.e. Intensity Measure (IM), along with the associated uncertainty, from a set of independent variables such as the earthquake magnitude and the event-to-site distance. GMMs are traditionally calibrated on global datasets, meaning that seismic records available in different parts of the world – which taken individually would not suffice for a robust calibration – are used together to perform the model calibration, which typically consists of a regression analysis. The obtained relationships are then applied globally, under the hypothesis that the conditional distribution of the ground motion parameter of interest given the magnitude, distance and site conditions, is identical at any site. This assumption however implies a high level of uncertainty associated with the estimated IMs, that reflects the large region-to-region variations observed on ground motion as a consequence of physical peculiarities at smaller scale, such as those related to different source and attenuation properties, as well as to site amplifications. Neglecting such region-specific variations leads not only to a larger variability but also to biased estimates of the IMs at more local scales for individual events and stations.

The current trend in the field of engineering seismology is thus moving towards region-specific GMMs. This is nowadays possible thanks to the increasing availability of seismic records in the majority of the most tectonically active countries. The resulting models provide different median predictions for different locations, instead of a single prediction that roughly averages all the ground motion effects at different scales. Recent studies have indeed focused on the development of new approaches for ground motion regionalization (Stafford (2014), Kotha et al. (2017), Sahakian et al. (2019), Kuehn et al. (2019), Sgobba et al. (2019), Parker et al. (2020), Kuehn and Abrahamson (2020), Kotha et al. (2020), Menafoglio et al. (2020)). The main strategy in this field is providing regional adjustments of the median GMM prediction, assuming that there are repeatable source, path, and site effects, which can be estimated from residual decomposition (where the term "residual" stands for the logarithmic deviation of a data point from the predicted IM; Anderson and Brune (2003), Atik et al. (2010)).

Another approach grounds on the development of GMMs having a single functional form for all sites with coefficients that vary with the geographical location (Landwehr et al. (2016a), Kuehn et al. (2019)). The pioneering work of Landwehr et al. (2016a) used a fully Bayesian approach built on the technique presented by Bussas et al. (2015), to introduce a double spatial non-stationarity of the model coefficients, which were allowed to be constant, or dependent on site- or event-coordinates (without depending on both types of coordinates simultaneously). The methodology – which was applied to build a GMM in California – revealed to be promising in order to improve GMMs accuracy and reduce the associated uncertainty, with a significant impact on hazard and engineering-oriented applications. However, its modeling and computational complexity represents a limitation for its use in the seismological practice.

In our work, we follow the same regionalization strategy adopted by Landwehr et al. (2016a) to introduce a spatial non-stationary GMM for Italy, but we embed the inference on this model in a different methodological framework based on the theory of geographically weighted regression (GWR, Brunsdon et al. (1998)). GWR allows one to model all the regression coefficients of a linear model as varying over space, and estimate them by localizing the model through spatial kernels. Although a generalization of this methodology – named mixed geographically weighted regression (MGWR, Fotheringham et al. (2002)) – allows one to keep some coefficients constant over space, none of the available GWR methods enables one to include *multiple* spatial non-stationarities within the model, i.e., non-stationarities deriving from the presence of multiple spatial indexes in the random process (hereafter called *multi-source* non-stationarity). In fact, even though GWR represents the natural framework to develop spatially variable GMMs, this methodological gap still represents an important limitation to its use, as GMMs need to incorporate both site- and event-coordinates within the model. As a key innovative contribution of this work, we thus further extend the GWR methodology, leading to multi-source GWR (MS-GWR), that allows one to jointly include *(i)* a set of stationary coefficients, and *(ii)* a double spatial non-stationarity within the model. We here propose a computational methodology to estimate the model parameters, as well as to quantify the associated uncertainty. We also develop an inferential framework for hypothesis testing on the model coefficients, based on a permutation approach. These developments enable us to propose a novel approach to build region-specific GMMs, which is here used to calibrate a GMM for the peak-ground acceleration over the entire Italian territory. This model shall be here built upon a large-scale dataset, collecting the seismic measures related with 4784 events recorded in Italy along 40 years.

The remaining part of this work is organized as follows. In Section 2 we recall the seismological background of this work, with particular reference to the state-of-the-art GMM in Italy (ITA18, Lanzano et al. (2019)), and the GMM proposed by Landwehr et al. (2016a); we here also describe the calibration dataset being considered in our study. Section 3 describes the MS-GWR, and the inferential framework we propose for hypothesis testing on the model coefficients. An extensive simulation study assessing the performance of MS-GWR is illustrated in Section 4. Section 5 describes the calibration of the GMM based on MS-GWR for Italy, and its validation. Section 6 eventually concludes the work. Codes and support material for the use of MS-GWR in ground-motion modeling are available on GitHub at github.com/lucaramenti/ms-gwr .

2 Background and data

The proposed methodology aims to extend to a spatial non-stationary framework the ITA18 model (Lanzano et al., 2019), which is the most updated version of the reference GMM for shallow crustal earthquakes in Italy. ITA18 provides the median value and associated uncertainty of a set of intensity measures (IMs), modeled as log-normal random variables. It was calibrated via a maximum likelihood approach, based on a linear model for the logarithmic transformation of the IMs. For ease of exposition, in this work we shall focus on a single IM, which is the peak ground acceleration (PGA) — Figure 1, although an analogous approach can be used on the other IMs considered by Lanzano et al. (2019). PGA is defined as the maximum absolute amplitude of an accelerogram recorded at a site during an earthquake (Douglas, 2003). It is the most commonly used ground motion parameter by engineers, as well as the main parameter considered by design codes to define seismic hazard at a site.

For the scope of the present work, it is relevant to recall the functional form of ITA18 for the PGA. In Lanzano et al. (2019), the PGA is modeled as:

$$\begin{aligned}\log_{10} PGA = a &+ b_1(M_w - M_h)\mathbb{1}_{(M_w \le M_h)} + b_2(M_w - M_h)\mathbb{1}_{(M_w > M_h)} \\ &+ [c_1(M_w - M_{ref}) + c_2)] \log_{10} \sqrt{R_{JB}^2 + h^2} + c_3\sqrt{R_{JB}^2 + h^2} \\ &+ k \left[\log_{10}(\tfrac{V_{S30}}{800})\mathbb{1}_{(V_{S30} \le 1500)} + \log_{10}(\tfrac{1500}{800})\mathbb{1}_{(V_{S30} > 1500)}\right] \\ &+ f_1 SoF_1 + f_2 SoF_2 + \epsilon,\end{aligned} \tag{1}$$

where the explanatory parameters M_w, R_{JB}, V_{S30} and SoF are respectively the event moment magnitude, the Joyner-Boore distance (i.e. a metric that defines the distance from a site to the surface projection of the fault rupture, Joyner and Boore (1981)), the shear wave velocity in the uppermost 30 meters (i.e. a proxy of the site response) and the style of faulting (i.e. a parameter describing the relative movement of the two sides of the fault plane), varying between normal, reverse and strike-slip. M_h, M_{ref} and h are fixed parameters, which have been estimated by a non-linear regression (Lanzano et al., 2019) and are here assumed to be known. Symbols a, b_1, b_2, c_1, c_2, c_3, f_1, f_2 and k denote the regression coefficients, which are the parameters of the model together with the variance σ^2 of the error ϵ. Note that Lanzano et al. (2019) further decomposed the variance of the error term ϵ in components due to event- and site-effects, in a mixed-effect framework. This latter decomposition is not considered further here, as the variability attributable to event- and site-effects shall be here captured through the non-stationarity of the model, as in Landwehr et al. (2016a).

We here aim to regionalize model (1), to allow for spatially varying coefficients in the GMM, similarly as done by Landwehr et al. (2016a) in the formulation of a non-ergodic GMM for California. These authors proposed to model the PGA through the following model – rewritten in log-10 units, for the ease of comparison with model (1)

$$\begin{aligned}\log_{10} PGA = \beta_{-1}(u_e, v_e) &+ \beta_0(u_s, v_s) + \beta_1 M + \beta_2 M^2 \\ &+ [\beta_3(u_e, v_e) + \beta_4 M] \ln \sqrt{R_{JB}^2 + h^2} + \beta_5(u_e, v_e) R_{JB} \\ &+ \beta_6(u_s, v_s) \ln V_{S30} + \beta_7 F_R + \beta_8 F_{NM} + \epsilon,\end{aligned} \tag{2}$$

where (u_e, v_e), (u_s, v_s) denote the event and site coordinates respectively. Landwehr et al. (2016a) calibrated the model for California in a Bayesian setting, by considering a Gaussian process prior over the spatially varying coefficients. In the following, we shall consider a GWR framework instead, as this represents a simpler but fully non-parametric alternative to the approach of Landwehr et al. (2016a). An approach based on GWR is also extendable to the setting of functional data analysis (FDA, Ramsay and Silverman (2005)), as we further discuss in Section 6, which could be used to model the entire response spectrum, as proposed by Menafoglio et al. (2020).

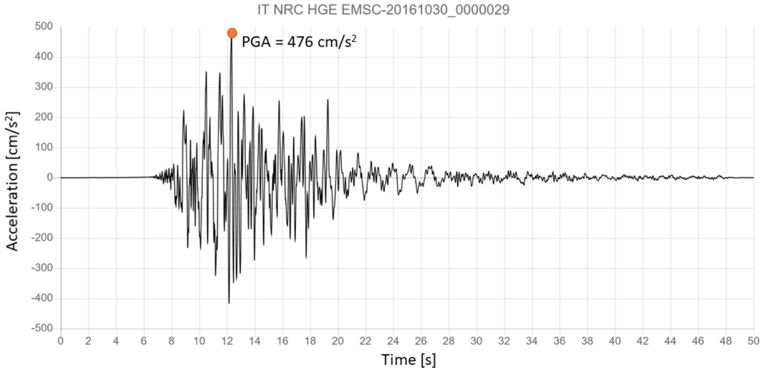


Figure 1: Indication of PGA on the acceleration waveform as recorded by the NRC station during the October 30, 2016 earthquake 06:40:18 UTC event (data are taken from ITACA database).

For consistency reasons, the dataset being considered in this study is substantially the same as the one used for calibration of ITA18, with the only modification consisting in the removal of some worldwide earthquakes, which were introduced in order to better constrain the regression at higher magnitudes. In particular, events occurred in Turkey, Japan, New Zealand, California (USA), Iceland, Iran and Greece are here removed as not relevant to the Italian data, while all the Italian earthquakes are kept in the dataset, along with events located in Slovenia, France and Croatia, which are neighboring countries. The resulting dataset is composed by 4784 observations of 137 events from 925 stations, recorded between 1976 and 2016, with magnitudes ranging from 3.5 to 6.9. The adopted acceleration waveforms and metadata are taken from the Engineering Strong Motion database, ESM - https://esm-db.eu/ (Luzi et al., 2020) and the ITalian ACcelerometric Archive, ITACA - http://itaca.mi.ingv.it/ (D'Amico et al., 2020).

Figure 2 shows the spatial distribution of the stations that recorded the events included in the dataset; here, each event is connected, with lines of the same colour, to all the stations which recorded it.

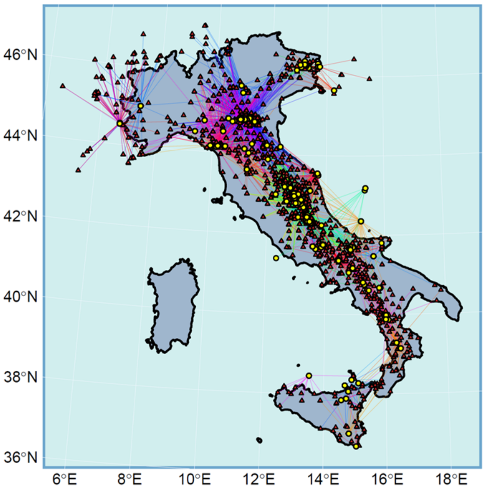


Figure 2: Map of the sampled ray-paths (colored lines) from the events (circles) to the stations (triangles).

3 Multi-source geographically weighted regression

3.1 Geographically weighted regression

Geographically weighted regression (Brunsdon et al., 1998) is a family of statistical methods aimed to estimate a regionalized linear model, characterized by spatially varying (a.k.a. non-stationary) coefficients. In this context, the general form of the regression model is

$$y_i = \sum_j \beta_j(u_i, v_i) x_{ij} + \epsilon_i, i = 1, ..., n, \tag{3}$$

where y_i is the response variable at the i-th site with coordinates (u_i, v_i), x_{ij} is the j-th regressor associated with the i-th unit, $\{\beta_j(u_i, v_i)\}_j$ are the regression coefficients, and $\{\epsilon_i\}_i$ are the i.i.d. random errors. Methods of GWR to estimate model (3) usually consist of localizing the estimation procedure, by calibrating the model in a *neighborhood* of the target site (u_0, v_0). This is typically selected through a spatial kernel K, which is a positive non-increasing function such that *(i)* $K(0) = 0$, and *(ii)* $\lim_{d->\infty} K(d) = 0$. In practice, the spatial kernel allows one to attribute a weight $K(d_{i0})$ to the available data based on their distance d_{i0} from the target site, thus naturally down-weighting data being distant from the target. A widely-used example for K is the Gaussian kernel, which shall also be employed in the following

$$K(d) = \exp\left\{-\frac{d^2}{2h^2}\right\},$$

where h denotes the kernel *bandwidth*.

When a spatial kernel is used, GWR reduces to a weighted least square regression, the weights being determined by the spatial kernel itself. That is, for a target location (u_0, v_0), $\beta_j(u_0, v_0)$, is estimated as

$$\hat{\beta}_j(u_0, v_0) = (X^T W_0 X)^{-1} X^T W_0 Y, \quad i = 1, ..., n, \tag{4}$$

with X the design matrix, Y the vector of observations of the response variable, and W_0 the diagonal matrix of kernel weights $W_{0,ii} = K(d_{i0})$.

To allow for the introduction of spatially stationary regression coefficients in model (3), GWR was lately extended to Mixed GWR (MGWR, Fotheringham et al. (2002)). Here, the general form of the model is

$$y_i = \sum_{j \in C} \beta_j x_{ij} + \sum_{j \in NS} \beta_j(u_i, v_i) x_{ij} + \epsilon_i, i = 1, ..., n, \tag{5}$$

where C denotes the set of spatially stationary terms, and NS the set of spatially non-stationary ones. An estimate of model (5) can be effectively obtained by using a two-steps algorithm, as advocated by Mei (2004). Here, first the constant term is estimated via OLS on an auxiliary regression problem, and then the non-stationary term is fitted by GWR on the residuals from the stationary term. The algorithm is recalled in details in Appendix A. In the following Section, this methodology is generalized to the case of multi-source non-stationarity, to enable the estimate of a model of the kind (2).

3.2 Multi-source GWR: model and estimation algorithm

We now extend GWR to allow for the presence of two sources of spatial non-stationarity, which are here representative of event- and site- effects in the GMM. The general model we aim to estimate takes the form

$$y_i = \sum_{j \in C} \beta_{j_C} x_{ij} + \sum_{j \in E} \beta_{j_E}(u_{e_i}, v_{e_i}) x_{ij} + \sum_{j \in S} \beta_{j_S}(u_{s_i}, v_{s_i}) x_{ij} + \epsilon_i, \quad i = 1, ..., n, \tag{6}$$

where (u_{e_i}, v_{e_i}), (u_{s_i}, v_{s_i}) are the event- and site-coordinates, respectively, of the i-th observation, C is the set of spatially stationary coefficients, E, S are the sets of spatially non-stationary coefficients, depending on event- or site- coordinates respectively, and ϵ_i zero-mean i.i.d. errors with variance σ^2.

In order to formulate the calibration algorithm, we first introduce two auxiliary estimation equations. Denote by $\tilde{y}_i$ and $\tilde{y}_i^{(s)}$ the following partial residuals

$$\tilde{y}_i = y_i - \sum_{j \in C} \beta_{j_C} x_{ij}; \quad \tilde{y}_i^{(s)} = \tilde{y}_i - \sum_{j \in S} \beta_{j_S}(u_{s_i}, v_{s_i}) x_{ij}. \tag{7}$$

The auxiliary estimation equation then reads

$$\tilde{y}_i = \sum_{j \in E} \beta_{j_E}(u_{e_i}, v_{e_i}) x_{ij} + \sum_{j \in S} \beta_{j_S}(u_{s_i}, v_{s_i}) x_{ij} + \epsilon_i \tag{8}$$

$$\tilde{y}_i^{(s)} = \sum_{j \in E} \beta_{j_E}(u_{e_i}, v_{e_i}) x_{ij} + \epsilon_i. \tag{9}$$

For ease of notation, $\beta_{j_E}(u_{e_i}, v_{e_i})$ and $\beta_{j_S}(u_{s_i}, v_{s_i})$ will be denoted hereafter by $\beta_{j_E,i}$ and $\beta_{j_S,i}$ respectively and, moreover, $\beta_{E,i} = (\beta_{1E,i}, ..., \beta_{pE,i})^T$, and $\beta_{S,i} = (\beta_{1S,i}, ..., \beta_{rS,i})^T$.

Note that, one may estimate model (9) for the partial residuals $\tilde{y}_i^{(s)}$ via GWR, as the right term in (9) only depends on a single set of coordinates (u_{e_i}, v_{e_i}). This yields

$$\hat{\beta}_{E,i} = (X_E^T W_{E,i} X_E)^{-1} X_E^T W_{E,i} \tilde{Y}^{(s)} = A_{E,i} \tilde{Y}^{(s)}, \quad i = 1, ..., n, \tag{10}$$

and the following estimate of the partial residuals

$$\hat{\tilde{Y}}^{(s)} = \begin{pmatrix} X_{E,1}\hat{\beta}_{E,1} \\ \vdots \\ X_{E,n}\hat{\beta}_{E,n} \end{pmatrix} = \begin{pmatrix} X_{E,1}(X_E^T W_{E,1} X_E)^{-1} X_E^T W_{E,1} \\ \vdots \\ X_{E,n}(X_E^T W_{E,n} X_E)^{-1} X_E^T W_{E,n} \end{pmatrix} \tilde{Y}^{(s)} = H_E \tilde{Y}^{(s)} \tag{11}$$

where $\tilde{Y}^{(s)} = (\tilde{y}_1^{(s)}, ..., \tilde{y}_n^{(s)})^T$ is the vector of partial residuals, $X_{E,i}$ stands for the i-th row of the design matrix X_E – containing the event-dependent covariates – and $W_{E,i}$ is the weighting matrix associated with the i-th sample unit, and built through the spatial kernel (see Subsection 3.1).

Plugging-in the estimated coefficients in eq. (8), leads to

$$\tilde{Y} - H_E \tilde{Y}^{(s)} = \begin{pmatrix} X_{S,1}\beta_{S,1} \\ \vdots \\ X_{S,n}\beta_{S,n} \end{pmatrix} + \epsilon, \tag{12}$$

with $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$. Replacing the definition of $\tilde{Y}^{(s)}$ (given in eq. (7)) in (12) and rearranging the terms yields

$$(I - H_E)\tilde{Y} = (I - H_E) \begin{pmatrix} X_{S,1}\beta_{S,1} \\ \vdots \\ X_{S,n}\beta_{S,n} \end{pmatrix} + \epsilon. \tag{13}$$

Note that eq. (13) can be interpreted as a regionalized model for a modified response vector ($(I - H_E)\tilde{Y}$) based on a modified regressors ($(I - H_E)X_{S,i}$, $i = 1, ..., n$). Hence, GWR can be applied again, finding the following estimates:

$$\begin{aligned} \hat{\beta}_{S,i} &= \left[X_S^T (I - H_E)^T W_{S,i} (I - H_E) X_S\right]^{-1} X_S^T (I - H_E)^T W_{S,i} (I - H_E)\tilde{Y} \\ &= A_{S,i}\tilde{Y}, \quad i = 1, ..., n \end{aligned} \tag{14}$$

from which we get

$$\begin{pmatrix} X_{S,1}\hat{\beta}_{S,1} \\ \vdots \\ X_{S,n}\hat{\beta}_{S,n} \end{pmatrix} = \begin{pmatrix} X_{S,1}A_{S,1} \\ \vdots \\ X_{S,n}A_{S,n} \end{pmatrix} \tilde{Y} = H_S \tilde{Y}. \tag{15}$$

Replacing all the estimated coefficients in the first auxiliary equation (8), and recalling again the definition of the partial residuals (7), we obtain

$$Y - H_E \tilde{Y}^{(S)} - H_S \tilde{Y} = X_C \beta_C + \epsilon \tag{16}$$

and substituting eq. (9) in this expression we get

$$Y - H_E(\tilde{Y} - H_S \tilde{Y}) - H_S \tilde{Y} = X_C \beta_C + \epsilon. \tag{17}$$

Replacing Equation (8) in Equation (17) we find

$$Y - H_E \left[Y - X_C \beta_C - H_S (Y - X_C \beta_C) \right] - H_S (Y - X_C \beta_C) = X_C \beta_C + \epsilon \tag{18}$$

which leads to

$$(I - H_E + H_E H_S - H_S) Y = (I - H_E + H_E H_S - H_S) X_C \beta_C + \epsilon. \tag{19}$$

Setting $B = (I - H_E + H_E H_S - H_S)$, we can apply OLS to Equation (19), obtaining

$$\hat{\beta}_C = (X_C^T B^T B X_C)^{-1} X_C^T B^T B Y. \tag{20}$$

It is possible to write an explicit formulation of the resulting hat matrix, using all the previous estimates:

$$\begin{aligned} \hat{Y} &= X_C \hat{\beta}_C + \begin{pmatrix} X_{E,1} \hat{\beta}_{E,1} \\ \vdots \\ X_{E,n} \hat{\beta}_{E,n} \end{pmatrix} + \begin{pmatrix} X_{S,1} \hat{\beta}_{S,1} \\ \vdots \\ X_{S,n} \hat{\beta}_{S,n} \end{pmatrix} \\ &= X_C \hat{\beta}_C + H_E \tilde{Y}^{(s)} + H_S \tilde{Y} \\ &= HY \end{aligned} \tag{21}$$

where

$$H = I - B + B X_C (X_C^T B^T B X_C)^{-1} X_C^T B^T B. \tag{22}$$

Summing up, all the coefficients can be estimated in cascade, through the algorithm reported in Figure 3.

Note that the order of estimation has been selected arbitrarily. In fact, using an analogous approach as that here discussed we may obtain six different estimation methods (one for each possible permutation of the estimation order). The following subsections shall assume that the estimation order is set as in Fig. 3, although they could be restated analogously with any other permuted order. In Section 4, we shall investigated by simulation the effect of the estimation order on the quality of the resulting estimates.

3.3 Parameter estimation accuracy and prediction uncertainty

In order to quantify the error made in estimating coefficients, set

$$A_C = (X_C^T B^T B X_C)^{-1} X_C^T B^T B,$$

Initialization: Define H_E, H_S as in eq. (11) and (15), and set $B = I - H_E + H_E H_S - H_S$.

Estimation steps:

- Estimate β_C as $\hat{\beta}_C = (X_C^T B^T B X_C)^{-1} X_C^T B^T BY$;
- Evaluate the estimated partial residuals $\hat{\tilde{Y}} = Y - X_C \hat{\beta}_C$;
- Estimate $\beta_{S,i}$ as $\hat{\beta}_{S,i} = A_{S,i} \hat{\tilde{Y}}, \quad i = 1, ..., n$;
- Evaluate the estimated partial residuals $\hat{\tilde{Y}}^{(S)} = \hat{\tilde{Y}} - H_S \hat{\tilde{Y}} = (I - H_S) \hat{\tilde{Y}}$;
- Estimate $\beta_{E,i}$ as $\hat{\beta}_{E,i} = A_{E,i} \hat{\tilde{Y}}^{(S)}, \quad i = 1, ..., n$.

Figure 3: Estimation algorithm of Multi-source GWR

and let e_k be a column vector whose k^{th} element is one, and the other elements are null. Then, denoting by $\hat{\beta}_{k,i}$ the estimate of the coefficient vector $\hat{\beta}_C$, $\hat{\beta}_{E,i}$, or $\hat{\beta}_{S,i}$ for $k = 1, 2, 3$ respectively, one has

$$\hat{\beta}_{k,i} = e_k^T \begin{pmatrix} \hat{\beta}_C \\ \hat{\beta}_{E,i} \\ \hat{\beta}_{S,i} \end{pmatrix} = e_k^T \begin{pmatrix} A_C Y \\ A_{E,i} \tilde{Y}^{(S)} \\ A_{S,i} \tilde{Y} \end{pmatrix} = e_k^T Q_i Y, \quad i = 1, ..., n \tag{23}$$

where

$$Q_i = \begin{pmatrix} A_C \\ A_{E,i}(I - H_S)(I - X_C A_C) \\ A_{S,i}(I - X_C A_C) \end{pmatrix}, \quad i = 1, ..., n. \tag{24}$$

Denoting by $\hat{\beta}_{\cdot,i}$ all the regression coefficients in location i, and noting that $\hat{\beta}_{\cdot,i} = Q_i Y$, one may compute the standard error of the estimator of the coefficients at location i as

$$Var(\hat{\beta}_{\cdot,i}) = \sigma^2 Q_i Q_i^T. \tag{25}$$

Using the unbiased estimate of σ^2 given by $\hat{\sigma}^2 = \frac{RSS}{\delta_1}$, where $\delta_1 = tr\{(I - H)^T (I - H)\}$ are the effective degrees of freedom of the estimator (Leung et al., 2000), we finally get

$$\widehat{Var}(\hat{\beta}_{\cdot,i}) = \frac{RSS}{\delta_1} Q_i Q_i^T, \quad i = 1, ..., n. \tag{26}$$

The variance (26) can also be used to provide an estimate of the prediction uncertainty at a target site (u_{S0}, v_{S0}) and for a target event at (u_{E0}, v_{E0}), as

$$\widehat{Var}(\hat{y}_0) = \widehat{Var}(x_0^T \hat{\beta}_{\cdot,0}) = S_0 \hat{\sigma}^2, \tag{27}$$

where $S_0 = x_0^T Q_0 Q_0^T x_0$, and Q_0 is defined analogously as in (24), but with the weight matrices $W_{S,0}$, $W_{E,0}$ (see eq. (14), (10)) computed through spatial kernel centred in (u_{S0}, v_{S0}) and (u_{E0}, v_{E0}) respectively. This result is fully analogous to the prediction uncertainty obtained for GWR by Fotheringham et al. (2002).

Notice that in seismological applications uncertainty is commonly split into two components, namely *aleatory* variability and *epistemic* uncertainty. Aleatory variability is intended as the natural randomness in a process, while epistemic uncertainty is defined as the uncertainty in the model of the process, caused by limited data and knowledge (Al Atik et al., 2010). One way to reduce aleatory variability is to identify those components of ground motion variability that are not completely random and to transfer them to the quantification of the epistemic uncertainty, for instance introducing spatially varying coefficients. For instance, consider a simple linear model

$$y = \beta_0 + \beta_1 x + \epsilon_1 \tag{28}$$

and its (single-source) non-stationary counterpart

$$y = \beta_0 + \beta_1(u, v)x + \epsilon_2 \tag{29}$$

where $Var(\epsilon_1) = \sigma_1^2$ and $Var(\epsilon_2) = \sigma_2^2$. The aleatory variability of the models (28) and (29) is represented by σ_1 and σ_2 respectively, while the epistemic uncertainty includes also the variability associated with the estimation of the model coefficients, resp. $\{\beta_0, \beta_1\}$, and $\{\beta_0, \beta_1(u, v)\}$. Introducing spatial non-stationarity in model (28) –yielding model (29)– may allow us to remove repeatable effects from σ_1, leading to $\sigma_2 < \sigma_1$, but also to an increased uncertainty related to parameter estimation. The advantage of transferring repeatable effects to epistemic uncertainty is that, unlike aleatory variability, it can be reduced introducing new data or knowledge. In fact, aleatory variability of a stationary linear model is constant over space and can be estimated using the variance of the error, while epistemic uncertainty for MS-GWR varies over space and can be partially quantified by estimating the statistical variability in the median predictions using eq. (27). This point shall be further explored in Section 5, and will be part of the comparative study between the proposed non-stationary GMM and the model of Lanzano et al. (2019).

3.4 Permutational inference for MS-GWR

In order to carry out inferential tests on regression parameters without relying on the normality assumption over residuals, we here develop a set of permutation tests, following the Freedman and Lane permutation scheme (Freedman and Lane, 1983). Its distinctive trait is that the permutations are carried out, under the null hypothesis, over the model residuals. Notice that this is an approximate test, since it is based on empirical residuals.

The general idea is that, if the null hypothesis being tested holds, the derived datasets should be equivalent to the original one: a small reported significance level indicates an unusual dataset under the null assumption.

Consider the test

$$\begin{aligned} &H_0 : \text{a given coefficient, other than the intercept, is constant} \\ &H_1 : \text{all coefficients, except for the intercept, vary over space} \end{aligned} \tag{30}$$

As test statistic consider

$$T = \frac{RSS_{H_0} - RSS_{H_1}}{RSS_{H_1}} = \frac{Y^T[R_{H_0} - R_{H_1}]Y}{Y^T R_{H_1} Y}, \tag{31}$$

where $R_{H_i} = (I - H_{H_i})^T(I - H_{H_i})$, $i = 0, 1$. The statistic (31) has already been used in GWR and MGWR literature for testing analogous assumptions on simpler models (Mei et al. (2016), Leung et al. (2000), Mei et al. (2006)), and compare, on a relative scale, the residuals of the models under H_0 and under H_1. To perform the test, we propose a permutation procedure which consists of the following steps

1. Find the optimal bandwidths, under H_0, for the spatial kernels involved in the computation of $W_{E,i}, W_{S,i}$ – appearing in (10) and (14);
2. Calibrate the models under H_0 and H_1 with the bandwidths found at Step 1.; compute the statistic T and the residuals under H_0, $\hat{\epsilon}_{H_0} = (I - H_{H_0})Y$;
3. Permute the residuals $\hat{\epsilon}_{H_0}$, obtaining $\hat{\epsilon}^{*b}$;
4. Build $Y^{*b} = H_{H_0}Y + \hat{\epsilon}^{*b}$;
5. Recalibrate both models under H_0 and H_1 using Y^{*b}, always with the same bandwidths, and compute T_i^{*b};
6. Repeat Steps 3. to 5. for B times;
7. Estimate the distribution of T^* from the replicates $\{T^{*b}\}_{b=1,\dots,B}$ and compare it with T, computing the *p-value* of the test as

   $$p = \frac{1}{b}\sum_{b=1}^{B} \mathbb{1}_{(T^{*b} > T)},$$

   the symbol $\mathbb{1}$ denoting the indicator function.

Finding the optimal bandwidths at Step 1. is not strictly necessary, the crucial part is calibrating the model under H_0 and H_1 with the same bandwidths, since this is the only way to obtain comparable values, thus a meaningful *p-value*. Moreover, we remark that one needs not to recompute the hat matrices for the recalibration at Step 5., as T_i^{*b} can be computed as

$$T^{*b} = \frac{(Y^{*b})^T[R_{H_0} - R_{H_1}]Y_i^{*b}}{(Y^{*b})^T R_{H_1} Y^{*b}}. \tag{32}$$

Note that considering T as test statistic yields a computational procedure which is much more efficient than that obtained by considering any test statistic based on the coefficients themselves, as in the latter case the hat matrices would need to be recomputed at any iteration.

Although formulated so far for testing on a single coefficient, the proposed test is very general, and can be used for testing the joint stationarity of multiple coefficients, or the single-source stationarity. This can be achieved by properly setting H_0 and H_1, and by consistently interpreting eq. (31). Moreover, to evaluate whether some explanatory variables in the stationary part of the model are significant or not, one may set a null model such that the coefficients corresponding to these explanatory variables are all zero.

Summing up, a possible approach, inspired by the bootstrap procedure proposed by Mei et al. (2016), is the following.

1. Test one at a time, exploiting also *a priori* knowledge if possible, the non-stationarity of the coefficients;
2. Test simultaneously the coefficients identified as not-significant at Step 1., considering them as spatially stationary under H_0;
3. Test singularly whether a stationary coefficient is significant or not, setting it to zero under H_0 and comparing two spatially varying models;
4. Test simultaneously the coefficients identified as not-significant at Step 3., considering them as null under H_0.

In the following Section 4 we illustrate an extensive simulation study which assess the performances of the proposed inferential procedure, in terms of level and power of the tests.

4 MS-GWR: a simulation study

4.1 Assessment of the estimation procedure

In this section, we explore through simulation the performances of MS-GWR, with particular regard to the estimation and prediction accuracy when changing *(i)* the order of estimation in the algorithm of Section 3 (Figure 3) and *(ii)* the bandwidth of the spatial kernels involved in the GWR estimates.

Data generation We consider the following three models

$$y_i = \beta_{0C} + \beta_{1C} x_{C,i} + \beta_{1E,i} x_{E,i} + \beta_{1S,i} x_{S,i} + \epsilon_i, \quad i = 1, ..., n \tag{33}$$

$$y_i = \beta_{0E,i} + \beta_{1C} x_{C,i} + \beta_{1E,i} x_{E,i} + \beta_{1S,i} x_{S,i} + \epsilon_i, \quad i = 1, ..., n \tag{34}$$

$$y_i = \beta_{0S,i} + \beta_{1C} x_{C,i} + \beta_{1E,i} x_{E,i} + \beta_{1S,i} x_{S,i} + \epsilon_i, \quad i = 1, ..., n \tag{35}$$

where x_{C_i}, x_{E_i} and x_{S_i} are the constant, event-dependent and site-dependent covariates, respectively, which are drawn from $\mathcal{N}(3, 25)$, $\mathcal{N}(-4, 25)$ and $\mathcal{N}(4, 25)$

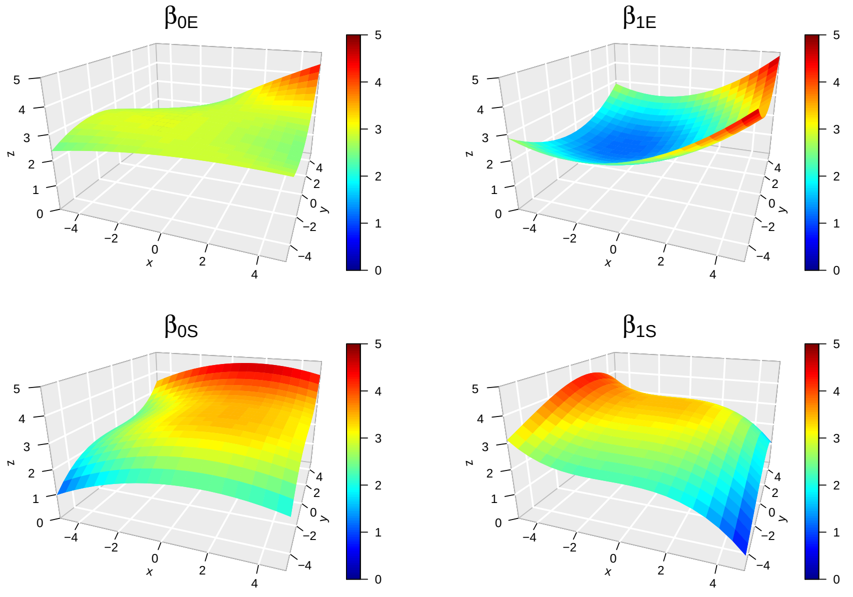


Figure 4: MS-GWR, True spatially varying coefficients

respectively, independently for any location i. The stationary coefficients are set to

$$\beta_C = (\beta_{0C}, \beta_{1C})^T = (8, 4)^T,$$

whereas the coefficient surfaces, shown in Figure 4, are formulated in the following way:

$$\begin{cases} \beta_{0E}(u_e, v_e) = -\frac{1}{100}(u_e+2)^2 + \frac{1}{100}(v_e+2)^2(u_e-1)+3 \\ \beta_{1E}(u_e, v_e) = \frac{1}{20}(u_e+1)^2 + \frac{1}{20}v_e^2 + \frac{1}{10}(u_e-2)+1.5 \\ \beta_{0S}(u_s, v_s) = \frac{1}{100}(u_s-2)^3 + \frac{1}{100}v_s^3 - \frac{1}{100}(u_s-1)^3+3.5 \\ \beta_{1S}(u_s, v_s) = -\frac{1}{100}(u_s+0.5)^3 - \frac{1}{100}(v_s-0.5)^3 + \frac{7}{1000}(v_s-2.5)^3+3.5 \end{cases} \tag{36}$$

Data are generated in a four-dimensional coordinate space, each observation being associated with a coordinate vector (u_e, v_e, u_s, v_s) in $\mathbb{R}^4$. However, since no regression coefficient depends on both site- and event-location, they are modelled as polynomial surfaces, each depending on pairs of coordinates in $\mathbb{R}^2$. For this reason, two spatial grids $\mathcal{G}_1$ and $\mathcal{G}_2$ are generated, both with values ranging in $[-5, 5]$ with step 0.5, resulting in two grids made of $n_g = 441$ elements each. The joint grid $\mathcal{G} = \mathcal{G}_1 \times \mathcal{G}_2$ in the four-dimensional space, resulting from the combination of $\mathcal{G}_1$ and $\mathcal{G}_2$, thus contains $n_g^2 = 194481$ points.

For the i-th grid point, with $i = 1, ..., n_g^2$, the value of ϵ_i is drawn from a normal distribution with zero mean and variance $\sigma^2 = 4$, independently on the other grid points, eventually computing the response y_i by applying either model (33) or (34) or (35). Note that we include the intercept in one component only, otherwise identifiability issues arise.

To perform the simulations hereafter shown, we build the training set by randomly and uniformly sampling $N = 50$ locations within the grid $\mathcal{G}$. The remaining $(n_g^2 - N)$ observations are used as test set. The model is thus estimated on the training set by following the method proposed in Section 3. The performance of the estimation procedure is assessed based on

(i) the accuracy in the estimate of the coefficients, quantified as

$$ERR_\beta = \left(n_g^2 \cdot \|\beta_C - \hat{\beta}_C\|^2 + \sum_{i=1}^{n_g^2} \|\beta_{E,i} - \hat{\beta}_{E,i}\|^2 + \sum_{i=1}^{n_g^2} \|\beta_{S,i} - \hat{\beta}_{S,i}\|^2 \right)^{1/2} ,$$

with β_C being the vector of stationary coefficients, and $\beta_{E,i}, \beta_{S,i}$ the vector of non-stationary coefficients at location $(u_{ei}, v_{ei}, u_{si}, v_{si})$ in $\mathcal{G}$;

(ii) prediction accuracy on the test set

$$ERR_{std} = \sum_{i \in test} \frac{y_i - \hat{y}_i}{\hat{\sigma}}.$$

For the sake of simplicity, in the following we always consider the same bandwidth for the kernels associated with components depending on site- and event-locations (i.e., those leading to define W_E and W_S in (10) and (14) respectively).

To filter out the dependence of the results on the sampled configuration of the data, we repeat the simulation for $M = 100$ replicates of the random training set –each obtained as described above– keeping fixed the value of the parameters described before, and study the distribution of the errors ERR_β and ERR_{std} across repetitions.

Dependence on the estimation order As far as notation is concerned, all different permutations have been named after the estimation order, which has to be read from right to left. In particular, C stands for spatially stationary, S for site-dependent and E for event-dependent. When model (33) is considered, we can see in Figure 5 that SEC and ESC seem to work significantly better than all the others, especially for low and medium bandwidth values, while this difference is not evident in case of larger bandwidths. This is related to the fact that the wider the bandwidth, the closer we get to a global model in which the estimation order is not relevant.

To better understand the level of accuracy attained on each coefficient of the model, we represent in Figure 6 the decomposition of the error ERR_β in the three terms appearing in (4.1). From Figure 6 it can be seen that the main difference of accuracy is in the estimation of the intercept. Indeed the error on spatially varying coefficients does not show a significant discrepancy

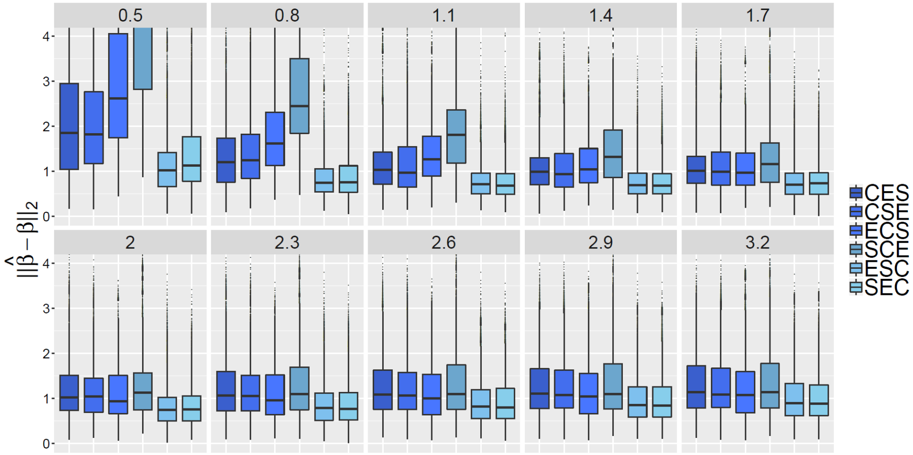


Figure 5: Simulation results for MS-GWR. Total error for different bandwidths when estimating model (33) (with stationary β_0) over $M = 100$ replicates of the simulations. Titles of the panels refer to the bandwidth used for the spatial kernels –the same bandwidth being used for both event- and site-coordinates.

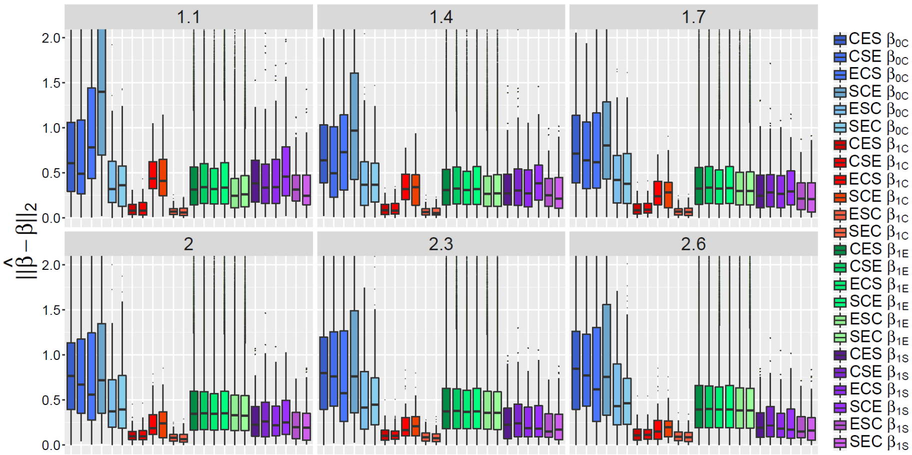


Figure 6: Simulation results for MS-GWR. Decomposed error for different bandwidths when estimating model (33) (with stationary β_0) over $M = 100$ replicates of the simulations. Titles of the panels refer to the bandwidth used for the spatial kernels –the same bandwidth being used for both event- and site-coordinates.

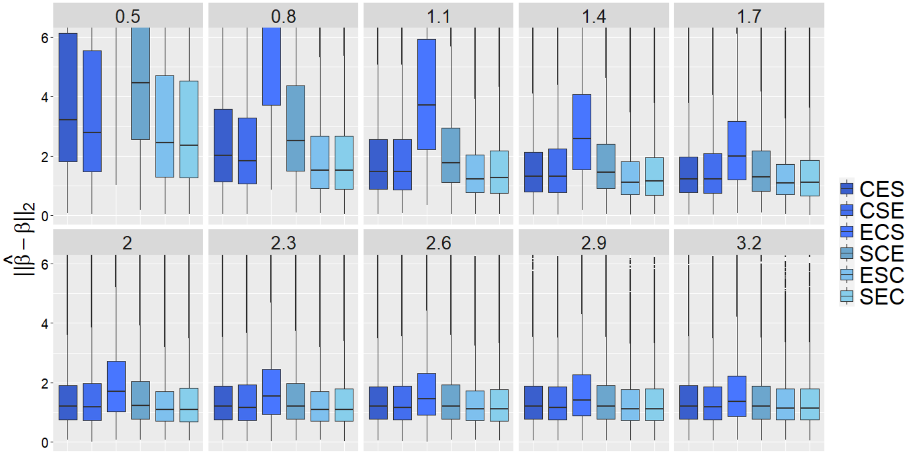


Figure 7: Simulation results for MS-GWR. Total error for different bandwidths when estimating model (34) over $M = 100$ replicates of the simulations. Titles of the panels refer to the bandwidth used for the spatial kernels –the same bandwidth being used for both event- and site-coordinates.

when changing the estimation order, while the estimate of the intercept is much preciser when the stationary component is estimated first.

If we consider a spatially varying intercept (models (34) and (35)), it can be noted in Figure 7 that there is no big dissimilarity between the quality of the estimates associated with different estimation orders; in fact, the error made in estimating the intercept is rather big with respect to all the other components, as shown in 8. Moreover, simulations suggest that, misspecifying the model for the intercept – by considering it as stationary when it is not – leads anyway to improved results in terms of quality in the estimation of remaining coefficients, as it can be seen in Figure 9. Thus in the application of MS-GWR shown in Section 5 we shall always consider the intercept as constant, and focus on the permutations which estimate the stationary component first, namely ESC and SEC.

Further analyses –which are not illustrated in details here for brevity– have shown that changing the magnitude of regression coefficients has no relevant impact on the relative error which is made in the coefficient estimation, while increasing the variance of the data generally leads to worse coefficient estimates. This is coherent with the fact that the lower the variance, the more difficult the estimation of regression coefficient becomes.

Bandwidth selection As far as the bandwidth selection is concerned, there is a bias-variance trade-off which is worth mentioning. Indeed, if the kernel includes points that are too far away, the variance will be low but the bias high, otherwise if the kernel only covers the closest points, the bias will be

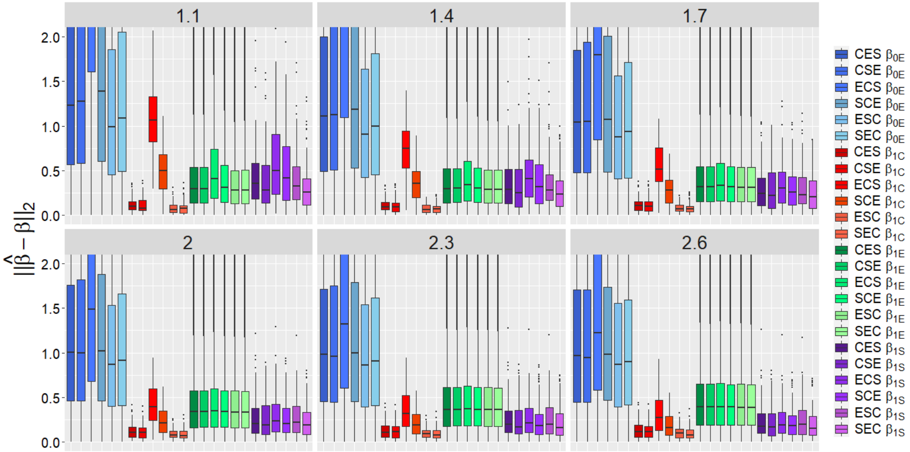


Figure 8: Simulation results for MS-GWR. Decomposed error for different bandwidths when estimating model (34) (with stationary β_0) over $M = 100$ replicates of the simulations. Titles of the panels refer to the bandwidth used for the spatial kernels –the same bandwidth being used for both event- and site-coordinates.

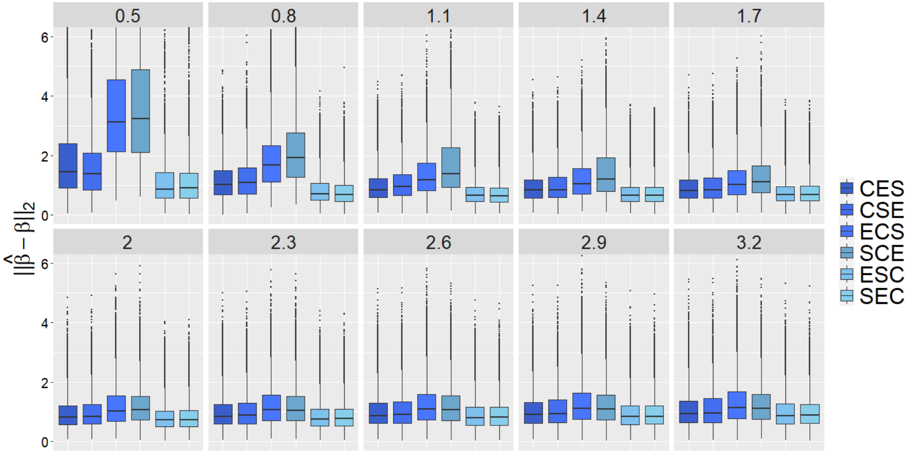


Figure 9: Simulation results for MS-GWR. Total error for different bandwidths when estimating model (34) over $M = 100$ replicates of the simulations, misspecifying the model and considering it as (33). Titles of the panels refer to the bandwidth used for the spatial kernels –the same bandwidth being used for both event- and site-coordinates.

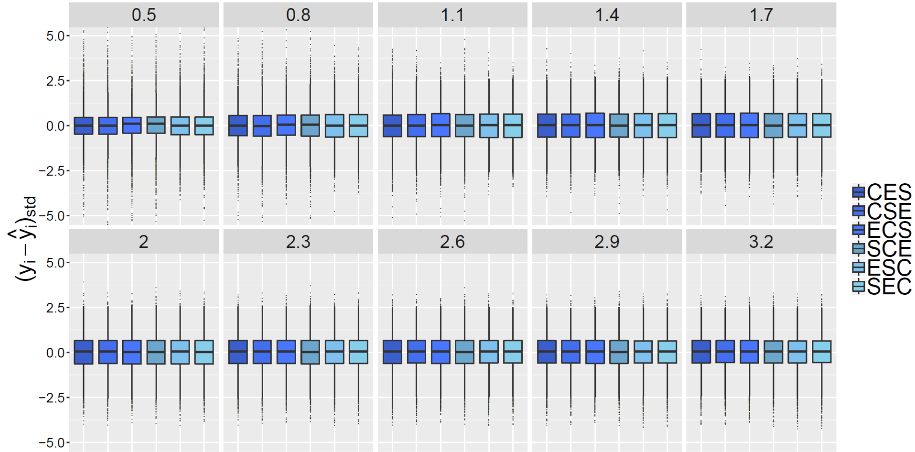


Figure 10: Simulation results for MS-GWR. Standardized residuals for different bandwidths when estimating model (33) (with stationary β_0) over $M = 100$ replicates of the simulation. Titles of the panels refer to the bandwidth used for the spatial kernels –the same bandwidth being used for both event- and site-coordinates.

low but the variance high. In fact, the larger the bandwidth, the higher is the number of data which are influential in the regression for location i, resulting in estimates closer to the ones that would be obtained by weighting all the data in the same way. On the other hand, if a small bandwidth is selected, only close data are attributed a high weight, leading to estimates associated with large standard errors, especially in the case of a small sample size. In this work, the optimal bandwidth will be selected by cross-validation, aiming to balance this bias-variance trade-off.

Prediction error Considering finally the prediction accuracy rather then the coefficients estimates, simulations show that there is no significant difference between the standardized error ERR_{std} when permuting the order of estimation, regardless of the chosen bandwidth. For instance, Figure 10 show the standard error estimated for model (33), when using different orders of estimation and different bandwidths –similar results being obtained for models (34) and (35) (not shown). This result suggests that the order of estimation has no relevant impact on the quality of point estimates of the response variable, but only on the coefficients. Thus, if we are interested in prediction only, any order of estimation can be selected. However, if, in addition to this, one aims to give an interpretation of the coefficients –as in our application to GMMs– the algorithms SEC and ESC appear to be the most suitable.

4.2 Permutational inference

We now assess the performance of the inferential procedure proposed in Section 3. To design the simulation study, we take inspiration from the simulation setting used by Mei et al. (2016) for assessing the performances of the bootstrap test for the constant coefficients of MGWR. We consider a spatial region made of two identical squares, with coordinates ranging from 0 to m with step equal to 1, where $m = 15$ or $m = 20$. The considered sample sizes are $N = 256$ and $N = 441$, respectively. As a model, the following equation is considered, for $i = 1, ..., N$,

$$y_i = \beta_{0C} + \beta_1(u_{e_i}, v_{e_i})x_{i1} + \beta_2(u_{s_i}, v_{s_i})x_{i2} + \beta_3(u_{e_i}, v_{e_i})x_{i3} + \beta_4(u_{s_i}, v_{s_i})x_{i4} + \epsilon_i, \tag{37}$$

where

$$\begin{cases} \beta_1(u_e, v_e) = 4\sin(\frac{\pi}{10}u_e) \\ \beta_2(u_s, v_s) = \frac{4}{625}u_s v_s(10 - u_s)(10 - v_s) \\ \beta_3(u_e, v_e) = 1 + 4c\left[1 + e^{-(u_e-5)}\right]^{-1} \\ \beta_4(u_s, v_s) = 0.5 + 2c\left[1 + \frac{1}{125}(v_s - 5)^3\right] \end{cases} \tag{38}$$

with c being a constant which will be designated different values to evaluate the power of the test. The model errors ϵ_i are independent and identically distributed random variables with zero mean and variance $\sigma^2 = 1$. Two different models for ϵ_i will be tested, namely $\epsilon_i \sim \mathcal{N}(0, 1)$ and $\epsilon_i \sim U(-\sqrt{3}, \sqrt{3})$. The covariates x_{ij}, $i = 1, ..., N$, $j = 1, ..., 4$, are generated as follows. Let Z_1, Z_2, Z_3 and Z_4 be independent random variables, distributed as $U(0, 1)$; the covariates are obtained as

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} 1 & \gamma & 0 & 0 \\ \gamma & 1 & \tau & 0 \\ 0 & \tau & 1 & \delta \\ 0 & 0 & \delta & 1 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{pmatrix}. \tag{39}$$

Parameters γ, δ and τ in (39) assume value 0.27 or to 0.5 one at a time, in order to introduce collinearity, with a correlation coefficient ρ equal to 0.5 or 0.8, respectively, between two covariates. The test we are going to consider is

$$\begin{aligned} &H_0 : \beta_3(u_e, v_e) = \beta_3 \text{ and } \beta_4(u_s, v_s) = \beta_4 \\ &H_1 : \text{all coefficients, except for the intercept, vary over space in their coordinate system.} \end{aligned} \tag{40}$$

Finally, the optimal bandwidths are selected minimizing over h the generalized cross-validation criterion (GCV, Mei (2004)), defined as

$$GCV(h) = \sum_{i=1}^{n} \frac{[y_i - \hat{y}_i(h)]^2}{[1 - H_{ii}(h)]^2} \tag{41}$$

where $H_{ii}(h)$ is the i^{th} diagonal element of the resulting hat matrix and $\hat{y}_i(h)$ are the estimated values. In the following, we shall always consider the SEC order of estimation, analogous results being obtained with ESC order (not shown).

		Error distribution	
Collinearity	n	$\mathcal{N}(0,1)$	$U(-\sqrt{3},\sqrt{3})$
Independent	256	0.052	0.048
	441	0.054	0.048
$\rho_{X_1X_2}=0.5$	256	0.038	0.040
	441	0.042	0.042
$\rho_{X_1X_2}=0.8$	256	0.042	0.042
	441	0.046	0.044
$\rho_{X_2X_3}=0.5$	256	0.034	0.038
	441	0.046	0.042
$\rho_{X_2X_3}=0.8$	256	0.040	0.042
	441	0.046	0.044
$\rho_{X_3X_4}=0.5$	256	0.042	0.040
	441	0.044	0.042
$\rho_{X_3X_4}=0.8$	256	0.040	0.044
	441	0.042	0.046

Table 1: Rejection rates of the permutation test under the null hypothesis ($c = 0$).

At first we set $c = 0$ and compute the rejection rate at a level $\alpha = 0.05$, running $M = 500$ replications, each with $B = 1000$ permutations. As we can see in Table 1, the rejection rates are reasonably close to the significance level, and the method seems to be robust to collinearity and different error distributions, as far as detecting stationary coefficients is concerned.

Then we set $c \neq 0$, with values ranging in $[0.1, 0.7]$ with step 0.1. In this case, the alternative hypothesis is true. We run $M = 500$ replications, each with $B = 1000$ permutations, with a significance level of $\alpha = 0.05$, focusing on independent covariates and on highly correlated covariates only. As we can see in Figure 11, the power increases both with increasing sample size and with increasing constant c. Considering different error distributions, no significant difference can be observed. As far as collinearity is concerned, there seems to be a slight loss in power, even though the capacity of non stationarity detection is not compromised.

5 Case study

5.1 Model calibration

In this section we calibrate via MS-GWR a non-ergodic GMM to describe the PGA, extending eq. (1) (Lanzano et al., 2019) to a spatially varying formulation inspired by the model (2) of Landwehr et al. (2016a). However, unlike Landwehr et al. (2016a), we here consider a stationary intercept, consistent with the results of the simulation study presented in Section 4. The model for the PGA we aim

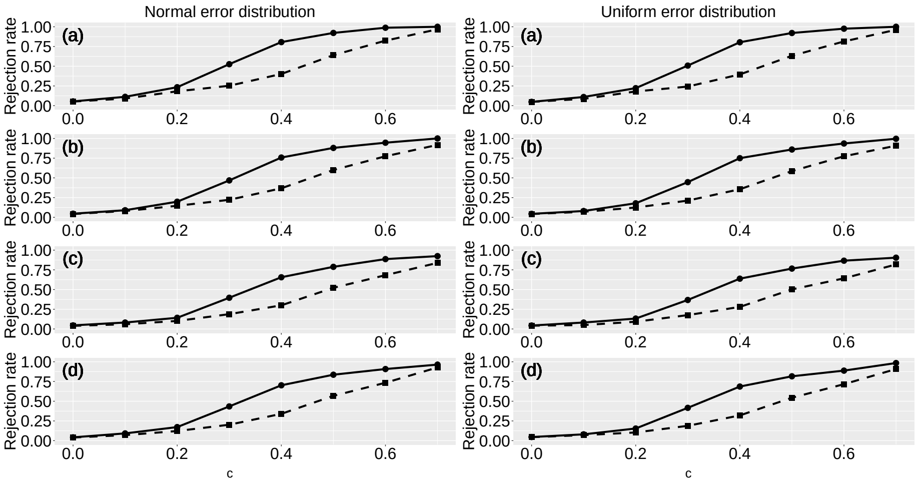


Figure 11: Power functions of the permutation test, under significance level $\alpha = 0.05$. The first to the fourth rows are, respectively, for (a) mutually independent variables, (b) $\rho_{X_1 X_2} = 0.8$, (c) $\rho_{X_2 X_3} = 0.8$ and (d) $\rho_{X_3 X_4} = 0.8$. Solid line for $n = 441$, dashed line for $n = 256$.

to estimate is

$$\begin{aligned} \log_{10} PGA = {} & a + b_1(M_w - M_h)\mathbb{1}_{(M_w \leq M_h)} + b_2(M_w - M_h)\mathbb{1}_{(M_w > M_h)} \\ & + [c_1(M_w - M_{ref}) + c_2(u_e, v_e)] \log_{10} \sqrt{R_{JB}^2 + h^2} + c_3(u_e, v_e)\sqrt{R_{JB}^2 + h^2} \\ & + k(u_s, v_s)\left[\log_{10}\left(\tfrac{V_{S30}}{800}\right)\mathbb{1}_{(V_{S30} \leq 1500)} + \log_{10}\left(\tfrac{1500}{800}\right)\mathbb{1}_{(V_{S30} > 1500)}\right] \\ & + f_1 SoF_1 + f_2 SoF_2 + \epsilon. \end{aligned} \tag{42}$$

Notice that M_h, M_{ref} and h are fixed parameters; M_w, R_{JB} and V_{S30} represent the covariates and a, b_1, b_2, c_1, c_2, c_3, f_1, f_2 and k the regression coefficients.

Bandwidth selection The calibration is carried out for grid with step equal to 10 km, covering the whole Italian territory, except for Sardinia, which is non-seismic; as a result, the considered grid is made of 2760 grid cells.

We carry out the whole calibration using SEC and then we select the best between ESC and SEC, by comparing their generalized cross-validation criterion values (GCV) found with the same bandwidths. More in details, we first select the optimal bandwidths for model (42) using the SEC order, finding $bw_E = 25$ km and $bw_S = 75$ km, and then carry out all the following tests using the same bandwidths. The reason for this choice is that selecting the optimal bandwidths bw_E and bw_S is the computationally heaviest step in the whole calibration, which is thus applied once, on the model we are most likely to use based on the prior knowledge on the GMM.

Null hypothesis	*p-value* PGA
$H_0 : b_{1,i} = b_1$	0.988
$H_0 : b_{2,i} = b_2$	0.116
$H_0 : c_{1,i} = c_1$	0.156
$H_0 : c_{2,i} = c_2$	**0.047**
$H_0 : c_{3,i} = c_3$	**0.012**
$H_0 : f_{1,i} = f_1$	0.972
$H_0 : f_{2,i} = f_2$	**0.033**
$H_0 : k_i = k$	**0.087**

Table 2: Permutation tests for stationary coefficients (1000 permutations). P-values lower than 10% are highlighted in bold.

Null hypothesis	*p-value* PGA
$H_0 : a = 0$	**0.000**
$H_0 : b_1 = 0$	**0.000**
$H_0 : b_2 = 0$	0.151
$H_0 : c_1 = 0$	**0.000**
$H_0 : f_1 = 0$	**0.031**
$H_0 : f_2 = 0$	0.117

Table 3: Permutation tests for null coefficients (1000 permutations). P-values lower than 10% are highlighted in bold.

Model selection Having fixed the bandwidths, we verify whether introducing spatial non-stationarity leads to improved results with respect to a stationary approach. A joint test for the stationarity of the coefficients shows a strong evidence of non-stationarity (*p-value*=0.000).

By analogy with Landwehr et al. (2016b) we expect that c_2 and c_3 –controlling geometric divergence and anelastic attenuation, respectively– shall depend on event-location. Moreover, we expect that k –which characterizes the soil under the station– shall depend on site-coordinates. A joint test on the non-stationarity of c_2, c_3 and k (H_0: all the coefficients are stationary; H_1: all the coefficients except c_2, c_3, k are stationary) shows evidence (level 10%) of their non-stationarity (*p-value*=0.078). These coefficients are hereafter considered as non-stationary, consistent with Landwehr et al. (2016b).

For the sake of completeness, Table 2 reports the results of hypothesis testing on the stationarity of the coefficients, when these tests are carried out one at a time. Note that one would reject (at the same level 10%) the null hypothesis for f_2. However, a non-stationary f_2 would hinder the physical interpretation to the model; in the following, f_2 is thus considered as constant.

Looking at how influential stationary covariates are, we refer to the results reported in Table 3. Here, one can see that b_2 and f_2 seem not to be significant at level 10%, consistently with the results obtained in the calibration of ITA18. Despite their limited impact on model predictions, these covariates were kept in ITA18, and shall be included in our model as well, to ease the comparison. This choice is also supported by the joint test on these coefficients, according to which they are jointly significant at level 10% (*p-value*=0.048).

Finally, the comparison of GCV values obtained for ESC or SEC –on the final model, estimated using the same bandwidths– leads to the selection of SEC over ESC (GCV=450.8 for SEC, GCV=498.6 for ESC).

	a	b_1	b_2	c_1	f_1	f_2
MS-GWR	3.5502	0.2354	-0.0513	0.2654	0.0510	0.0394
	(0.0454)	(0.0326)	(0.0372)	(0.0188)	(0.0205)	(0.0240)
ITA18	3.4210	0.1940	-0.0220	0.2871	0.0860	0.0105
	(0.0459)	(0.0332)	(0.0411)	(0.0104)	(0.0359)	(0.0344)

Table 4: Point estimate of the stationary coefficients; standard deviations are reported between brackets.

5.2 Interpretations

In Table 4 the estimated *stationary* coefficients are reported together with their standard deviation, and compared with the ones obtained in ITA18. No evident discrepancy between the two models is observed in this stationary part.

Figure 12 displays the spatial representation of the non-stationary coefficients, each referred to the corresponding domain of variation (i.e., event-coordinates (u_e, v_e) for c_2, c_3 and site-coordinates (u_s, v_s) for k). The site-dependent estimate varies much more smoothly than the event-dependent ones. This is likely to be due to the different density of events with respect to stations and to the differing bandwidths that have been previously selected.

As far as c_2 and c_3 are concerned, one can see that they behave in a complementary way, higher values of geometrical spreading being associated with lower values of anelastic attenuation and vice versa.

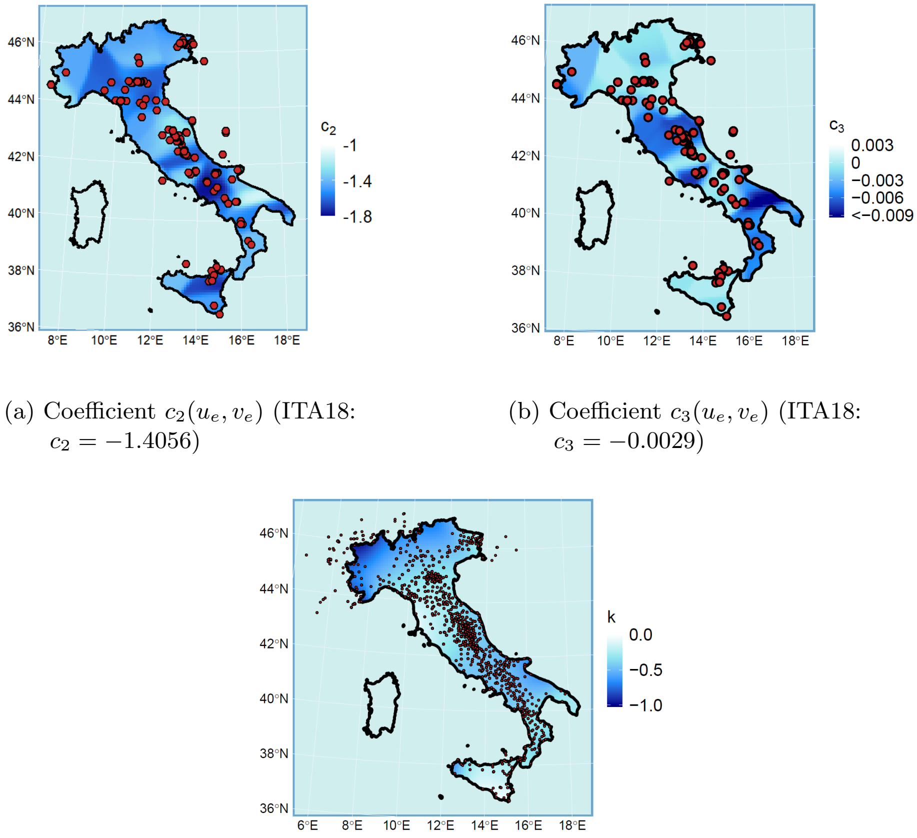
Notice that, in the calibration of ITA18, c_3 was set to 0 when positive, since it would lead to an enhancement of the spectral amplitudes, which is not physically meaningful in general. Nevertheless, in the calibration using MS-GWR, positive values of c_3 have been kept, since this phenomenon can be observed in the Po Plain, where we may have reflection effects, for both long and short periods, due to Moho discontinuity, which marks the transition in composition between the Earth's rocky outer crust and the more plastic mantle (Lanzano et al., 2016).

5.3 Residuals and uncertainty

Focusing on the residuals, they do not show relevant patterns (see Figure 13), thus indicating that the model succeeds in capturing the effects of the input variables.

We now compare the uncertainty of our model and associated predictions with those of ITA18. Recall that RSS/δ_1, with δ_1 the effective degrees of freedom, is an unbiased estimate of the variance of the error σ^2 (see Section 3). On this basis, we obtain an estimated standard deviation of $\hat{\sigma} = 0.3001$ against a standard deviation for ITA18 of $\hat{\sigma}_{ITA18} = 0.3362$. Introducing spatial non-stationarity thus leads to a moderate reduction of the aleatory variability.

While the aleatory variability of the model is constant over space, epistemic uncertainty for MS-GWR is spatially varying. The joint effect of both



(a) Coefficient $c_2(u_e, v_e)$ (ITA18: $c_2 = -1.4056$)

(b) Coefficient $c_3(u_e, v_e)$ (ITA18: $c_3 = -0.0029$)

(c) Coefficient $k(u_s, v_s)$(ITA18: $k = -0.3946$)

Figure 12: Maps of non-stationary coefficients, estimated via MS-GWR

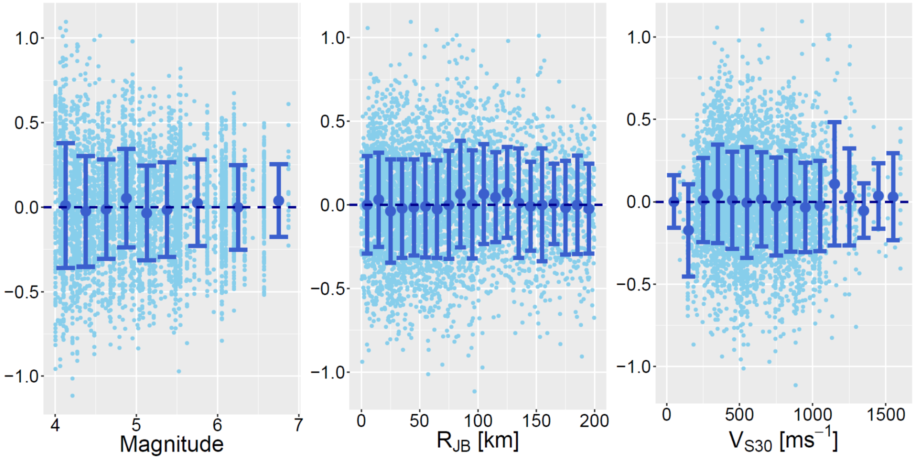


Figure 13: Residuals of model (42) estimated via MS-GWR

variabilities can be assessed by evaluating the statistical variability in the median predictions. To evaluate its spatial variation, we set the input variables to M_w=5, V_{S30}=300$\frac{m}{s}$, SoF=NF and R_{JB}=10km and predict the response for PGA. For the sake of simplicity and following Landwehr et al. (2016b), the same event-station coordinates are considered (i.e., $(u_e, v_e) = (u_s, v_s)$).

Graphical inspection of Figure 14 suggests that the lowest values of predictive uncertainty are located in areas with a very high density of data, both of stations and events. On the other hand, its highest values are observed in areas characterized by a lack of data, e.g., in the region of the Alps, Apulia and Sicily. These values are generally higher than the epistemic uncertainty related to ITA18, coherently with the transferral of repeatable effect from aleatory variability to epistemic uncertainty.

5.4 Model validation

In order to validate the model, we carry out a 10-fold cross-validation, splitting the dataset completely at random in 10 folds $F_1, ..., F_{10}$ and comparing the mean squared error, defined as

$$MSE_{10-fold} = \frac{1}{10}\sum_{j=1}^{10}\frac{\sum_{i\in F_j}(y_i - \hat{y}_{-j})^2}{N_j}, \tag{43}$$

where $\hat{y}_{-j}$ is the predicted value using the model calibrated using all folds except for F_j. Results show that MS-GWR leads to improved results ($MSE_{10-fold}$ = 0.09252) with respect to ITA18 ($MSE_{10-fold}$ = 0.11996), supporting the introduction of spatially varying coefficients.

The resulting GMM is also tested to independent events (i.e., events outside the calibration dataset). We thus consider the following events

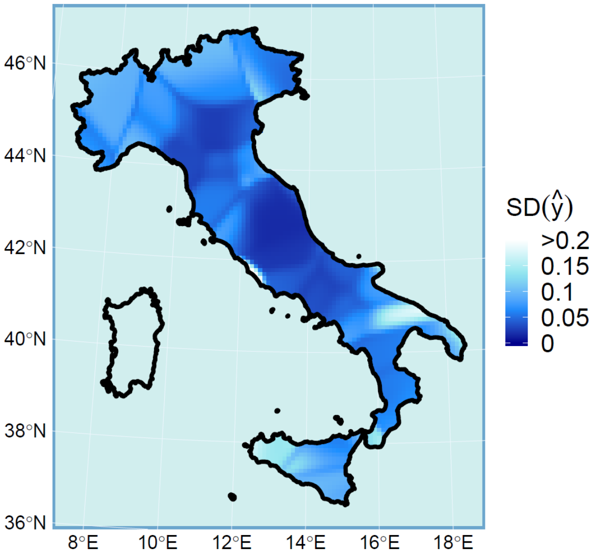


Figure 14: PGA, standard deviation of $\hat{y}$ (in $\log_{10}$ units) for the input M_w=5, V_{S30}=300$\frac{m}{s}$, SoF=NF and R_{JB}=10km

- Muccia (MC), ID EMSC-20180410_0000011, 10.04.2018, M_w=4.6, 174 observations;
- Termoli (CB), ID EMSC-20180816_0000090, 16.08.2018, M_w=5.1, 167 observations;
- Barletta (BT), ID EMSC-20190521_0000022, 21.05.2019, M_w=4.0, 51 observations;
- Siracusa (SR), ID IT-1990-0003, 13.12.1990, M_w=5.6, 7 observations.

The adopted records for testing are taken from the Italian Accelerometric Archive (ITACA) (Pacor et al., 2011), available at website http://itaca.mi.ingv.it. Figure 15 show the standardized residuals of the four selected seismic events. One can note that MS-GWR leads to equivalent or better results than ITA18; in particular, the best results are observed in areas which are densely sampled in the calibration dataset.

6 Discussion and conclusion

In this work, we proposed a novel approach to calibrate regionalized regression models accounting for multiple spatial non-stationarities, with a particular focus on non-stationary ground motion models depending on site- and event-effects. The proposed approach is of general validity, and could be potentially applied

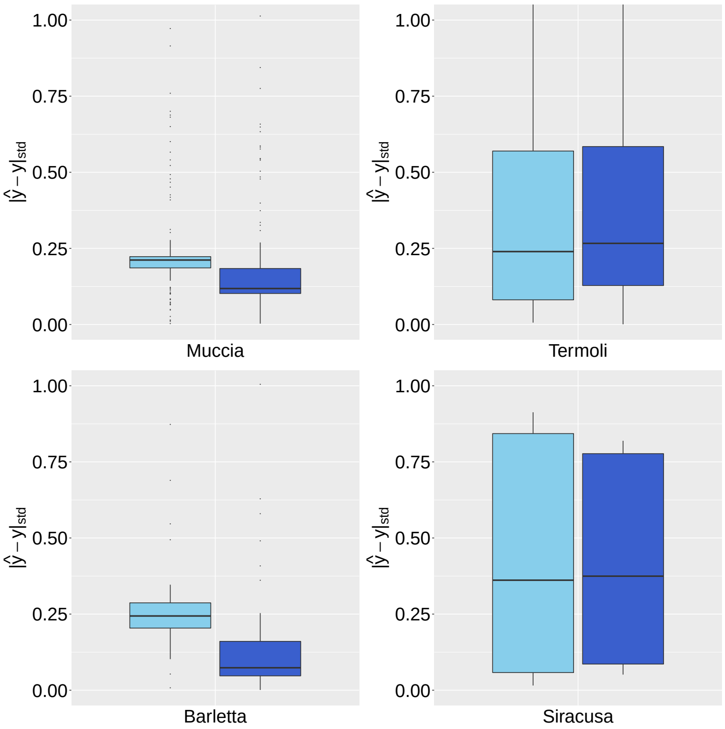


Figure 15: Standardized residuals of independent events: ITA18 (left light box) vs MS-GWR (right dark box)

in varied environmental and industrial settings, ranging from climatology to the oil and gas industry. In the field of seismology, the approach represents an alternative to the Bayesian methodology described by Landwehr et al. (2016a), presenting the significant advantage of being simpler and fully non-parametric. From the application viewpoint, the proposed approach allowed us to regionalize the state-of-the-art model for PGA in Italy (Lanzano et al., 2018), making explicit the non-stationary relation between the response variable (PGA) and the predictors. The extensive validation study performed in Section 5 allows us to conclude that the proposed model exhibits (a) a good capability to capture the main physical aspects related to the source, site and path terms; (b) a model uncertainty which is generally higher for the Italian regions where data are sparse (Western Sicily, Southern Apulia) and lower where data are densely sampled (Central Italy); (c) a lower aleatory variability, as a consequence of the regionalization process through spatially varying predictions, which necessarily reflects on a larger epistemic uncertainty; and (d) a decrease in the overall prediction error (both in cross-validation and on independent events) with respect to the state-of-the-art stationary model (ITA18, Lanzano et al. (2018)).

The results here presented thus appear very promising, and classify the methodology as a good candidate for the regionalization of global ground motion models when enough sampling coverage is available. This opens important perspectives for the computation of site-specific Probabilistic Seismic Hazard Analysis (PSHA), as well as for the development of shaking scenarios in loss prediction and emergency planning purposes.

Grounding on the theory of geographically weighted regression (GWR), the approach here proposed is also prone to be extended to more complex settings as functional data analysis (FDA, Ramsay and Silverman (2005)) and object oriented data analysis (OODA, Marron and Alonso (2014)). Such extension could potentially allow one to consider functional intensity measures, such as the spectral acceleration $SA(T)$ as a function of the period of oscillation T (of which PGA is a point evaluation at $T = 0$). A pioneering study in this direction was recently proposed by Menafoglio et al. (2020), who presented a functional simulation setting for these types of data, based on the residuals of the GMM ITA18. The development of functional GMMs is seen by the authors as a powerful perspective of research, which could lead to breakthrough advances in engineering seismology, and could naturally stem from the research proposed in this work.

A Mixed Geographically Weighted Regression

In this Appendix we show a short overview over *mixed geographically weighted regression* (MGWR), a generalization of geographically weighted regression (GWR), in which some coefficients are constant over space and the others present spatial non-stationarity. In particular, we will show the estimation algorithm proposed by Mei (2004), which is computationally less intensive than the original iterative estimation (Fotheringham et al., 2002).

The starting formulation of the model is

$$y_i = \sum_{j \in C} \beta_j x_{ij} + \sum_{j \in NS} \beta_j(u_i, v_i) x_{ij} + \epsilon_i, \quad i = 1, ..., n. \tag{A.1}$$

which can be rewritten as:

$$\tilde{y_i} = y_i - \sum_{j \in C} \beta_j x_{ij} = \sum_{j \in NS} \beta_j(u_i, v_i) x_{ij} + \epsilon_i \quad i = 1, ..., n. \tag{A.2}$$

The general idea is now to apply at first standard GWR to eq. (A.2), in order to estimate spatially non-stationary coefficients, and then to apply OLS to obtain an estimate of the stationary coefficients. Applying this procedure, the following three-step estimation algorithm is obtained:

- Estimate β_C as $\hat{\beta}_C = [X_C^T(I - H_{NS})^T(I - H_{NS})X_C]^{-1}X_C^T(I - H_{NS})^T(I - H_{NS})Y$;
- Evaluate the estimated partial residuals $\tilde{Y} = Y - X_C\hat{\beta}_C$;
- Estimate $\beta_{NS,i}$ as $\hat{\beta}_{NS,i} = (X_{NS}^T W_i X_{NS})^{-1} X_{NS}^T W_i \tilde{Y}, \quad i = 1, ..., n$

where W_i is the weighting matrix associated with the i-th sample unit and

$$H_{NS} = \begin{pmatrix} X_{NS,1}(X_{NS}^T W_1 X_{NS})^{-1} X_{NS}^T W_1 \\ \vdots \\ X_{NS,n}(X_{NS}^T W_n X_{NS})^{-1} X_{NS}^T W_n \end{pmatrix}. \tag{A.3}$$

The explicit formulation of the resulting hat matrix finally is:

$$H = H_{NS} + (I - H_{NS})X_C \left[X_C^T(I - H_{NS})^T(I - H_{NS})X_C\right]^{-1} X_C^T(I - H_{NS})^T(I - H_{NS}) \tag{A.4}$$

References

L. Al Atik, N. Abrahamson, J. Bommer, F. Scherbaum, F. Cotton, and N. Kuehn. The Variability of Ground-Motion Prediction Models and Its Components. *Seismological research letters*, 81(5), 08 2010.

N. Anderson and J. Brune. Probabilistic seismic hazard analysis without the ergodic assumption. *Seismol. Res. Lett.*, pages 19–28, 2003.

A. L. Atik, N. Abrahamson, Bommer J. J., Scherbaum F., Cotton F., and Kuehn N. The variability of ground-motion prediction models and its components. *Seismol. Res. Lett.*, pages 794–801, 2010.

C. Brunsdon, A. S. Fotheringham, and M. Charlton. Geographically Weighted Regression - Modelling spatial non-stationarity. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443, 1998.

M. Bussas, C. Sawade, T. Scheffer, and N. Landwehr. Varying-coefficient models with isotropic Gaussian process priors. *arXiv e-prints*, page arXiv:1508.07192, August 2015.

M. C. D'Amico, C. Felicetta, E. Russo, S. Sgobba, G. Lanzano, F. Pacor, and L. Luzi. ITalian ACcelerometric Archive (ITACA), version 3.1, 2020. URL `http://itaca.mi.ingv.it/`.

J. Douglas. Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. *Earth-Science Reviews*, 61(1):43 – 104, 2003.

S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically Weighted Regression - the analysis of spatially varying relationships.* John Wiley & Sons Ltd, 2002.

D. Freedman and D. Lane. A Nonstochastic Interpretation of Reported Significance Levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983.

W. B. Joyner and D. M. Boore. Peak horizontal acceleration and velocity from strong-motion records including records from the 1979 imperial valley, California, earthquake. *Bulletin of the Seismological Society of America*, 71(6): 2011–2038, 1981.

S. R. Kotha, D. Bindi, and F. Cotton. From ergodic to region- and site-specific probabilistic seismic hazard assessment: Method development and application at european and middle eastern sites. *Earthquake Spectra*, 33(4):1433–1453, 2017.

S. R. Kotha, G. Weatherill, D. Bindi, and F. Cotton. A regionally-adaptable ground-motion model for shallow crustal earthquakes in Europe. *Bulletin of Earthquake Engineering*, pages 1–35, may 2020.

N. Kuehn, S. Kotha, and N. Landwehr. A Non-ergodic GMPE for Europe and the Middle East with Spatially Varying Coefficients. In *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, page 11166, 2019.

N. M. Kuehn and N. A. Abrahamson. Spatial correlations of ground motion for non-ergodic seismic hazard analysis. *Earthquake Engineering & Structural Dynamics*, 49(1):4–23, 2020.

N. M. Kuehn, N. A. Abrahamson, and M. A. Walling. Incorporating Nonergodic Path Effects into the NGA-West2 Ground-Motion Prediction Equations. *Bulletin of the Seismological Society of America*, 109(2):575–585, 2019.

N. Landwehr, N. M. Kuehn, T. Scheffer, and N. Abrahamson. A nonergodic ground-motion model for California with spatially varying coefficients. *Bulletin of the Seismological Society of America*, 106(6):2574–2583, 2016a.

N. Landwehr, N. M. Kuehn, T. Scheffer, and N. Abrahamson. A Nonergodic Ground-Motion Model for California with Spatially Varying Coefficients. *Bulletin of the Seismological Society of America*, 106(6):2574–2583, 2016b.

G. Lanzano, M. D'Amico, C. Felicetta, R. Puglia, L. Luzi, F. Pacor, and D. Bindi. Ground-motion prediction equations for region-specific probabilistic seismic-hazard analysis. *Bulletin of the Seismological Society of America*, 106(1):73–92, 2016.

G. Lanzano, S. Sgobba, L. Luzi, R. Puglia, F. Pacor, C. Felicetta, M. D'Amico, F. Cotton, and D. Bindi. The pan-European Engineering Strong Motion (ESM) flatfile: compilation criteria and data statistics. Bulletin of Earthquake Engineering. *Bulletin of Earthquake Engineering*, 17(2):561–582, 2018.

G. Lanzano, F. Pacor L. Luzi, C. Felicetta, R. Puglia, S. Sgobba, and M. D'Amico. A Revised Ground-Motion Prediction Model for Shallow Crustal Earthquakes in Italy. *Bulletin of the Seismological Society of America*, 109 (2):525–540, 2019.

Y. Leung, C. L. Mei, and W. X. Zhang. Statistical Tests for Spatial Nonstationary Based on the Geographically Weighted Regression Model. *Environment and Planning A*, 32:9–32, 02 2000.

L. Luzi, G. Lanzano, C. Felicetta, M. C. D'Amico, E. Russo, S. Sgobba, and F.and ORFEUS Working Group 5 Pacor. Engineering strong motion database (esm), 2020. URL `https://esm-db.eu`.

J. S. Marron and A. M. Alonso. Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753, 2014.

C. L. Mei. Geographically Weighted Regression Technique for Spatial Data Analysis. 01 2004.

C. L. Mei, N. Wang, and W. X. Zhang. Testing the importance of the explanatory variables in a mixed geographically weighted regression model. *Environment and Planning A*, 38:587–598, 02 2006.

C. L. Mei, M. Xu, and N. Wang. A bootstrap test for constant coefficients in geographically weighted regression models. *International Journal of Geographical Information Science*, 30(8):1622–1643, 2016.

A. Menafoglio, S. Sgobba, G. Lanzano, and F. Pacor. Simulation of seismic ground motion fields via object-oriented spatial statistics with an application in Northern Italy. *Stochastic Environmental Research and Risk Assessment*, 34(10):1607–1627, 2020.

F. Pacor, R. Paolucci, L. Luzi, F. Sabetta, , A. Spinelli, A. Gorini, M. Nicoletti, S. Marcucci, L. Filippi, and M. Dolce. Overview of the Italian strong motion database ITACA 1.0. Bulletin of Earthquake Engineering. *Bulletin of Earthquake Engineering*, 9(6):1723–1739, 2011.

G. A. Parker, A. S. Baltay, J. Rekoske, and E. M. Thompson. Repeatable source, path, and site effects from the 2019 m 7.1 ridgecrest earthquake sequence. *Bulletin of the Seismological Society of America*, pages 1–19, 2020.

J. Ramsay and B. Silverman. *Functional data analysis.* Springer, New York, second edition, 2005.

V. J. Sahakian, A. Baltay, T. C. Hanks, J. Buehler, F. L. Vernon, D. Kilb, and N. A. Abrahamson. Ground motion residuals, path effects, and crustal properties: A pilot study in southern california. *Journal of Geophysical Research: Solid Earth*, 124(6):5738–5753, 2019.

S. Sgobba, G. Lanzano, F. Pacor, R. Puglia, M. D'Amico, C. Felicetta, and L. Luzi. Spatial correlation model of systematic site and path effects for ground-motion fields in northern italy. *Bulletin of the Seismological Society of America*, 109(4):1419—-1434, 2019.

P.J. Stafford. Crossed and nested mixed-effects approaches for enhanced model development and removal of the ergodic assumption in empirical ground-motion models. *Bulletin of the Seismological Society of America*, 104(2): 702–719, 2014.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

66/2020 Didkovsky, O.; Ivanov, V.; Papini, M.; Longoni, L.; Menafoglio, A.
A comparison between machine learning and functional geostatistics approaches for data-driven analyses of solid transport in a pre-Alpine stream

65/2020 Di Gregorio, S.; Vergara, C.; Montino Pelagi, G.; Baggiano, A.; Zunino, P.; Guglielmo, M.; Fu
Prediction of myocardial blow flow under stress conditions by means of a computational model

64/2020 Fiz, F.; Viganò, L.; Gennaro, N.; Costa, G.; La Bella, L.; Boichuk A.; Cavinato, L.; Sollini, M.;
Radiomics of Liver Metastases: A Systematic Review

63/2020 Tuveri, M.; Milani, E.; Marchegiani, G.; Landoni, L.; Torresani, E.; Capelli, P.; Sperandio, N.;
HEMODYNAMICS AND REMODELING OF THE PORTAL CONFLUENCE IN PATIENTS WITH CANCER OF THE PANCREATIC HEAD: A PILOT STUDY

62/2020 Massi, M. C.; Ieva, F.
Representation Learning Methods for EEG Cross-Subject Channel Selection and Trial Classification

61/2020 Pozzi, S.; Redaelli, A.; Vergara, C.; Votta, E.; Zunino, P.
Mathematical and numerical modeling of atherosclerotic plaque progression based on fluid-structure interaction

60/2020 Lupo Pasini, M; Perotto, S.
Hierarchical model reduction driven by a Proper Orthogonal Decomposition for parametrized advection-diffusion-reaction problems

59/2020 Massi, M.C.; Franco, N.R; Ieva, F.; Manzoni, A.; Paganoni, A.M.; Zunino, P.
High-Order Interaction Learning via Targeted Pattern Search

58/2020 Beraha, M.; Pegoraro, M.; Peli, R.; Guglielmi, A
Spatially dependent mixture models via the Logistic Multivariate CAR prior

57/2020 Regazzoni, F.; Quarteroni, A.
An oscillation-free fully partitioned scheme for the numerical modeling of cardiac active mechanics