Received: 20 October 2020 | Revised: 12 October 2021 | Accepted: 20 October 2021

DOI: 10.1002/asmb.2654

RESEARCH ARTICLE

WILEY

High leverage detection in general functional regression models with spatially correlated errors

Elvira Romano[1] | Ramón Giraldo[2] | Jorge Mateu[3] | Andrea Diana[1]

[1]Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy

[2]Department of Statistics, Science Faculty, Universidad Nacional de Colombia, Bogotà, Colombia

[3]Department of Mathematics, University Jaume I, Castellòn, Spain

Correspondence
Elvira Romano, Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy.
Email: elvira.romano@unicampania.it

Funding information
Universita degli Studi della Campania Luigi Vanvitelli

Abstract

The presence of curves that deviate markedly from the core of a set of curves can greatly affect inference and forecasting in a functional regression model. Thus their detection is key to increase the accuracy of the required estimates. This work introduces the concepts of high leverage in general functional regression models with independent and spatially correlated errors. The projection matrix, also known as Hat matrix, plays a crucial role in classical model diagnosis, since it provides a measure of leverage. We propose a generalisation of the projection matrix in both the functional and the spatial functional frameworks under two settings, when the response variable is a scalar, and when it is a function itself, the so-called total model. Commonly used influence measures are also proposed as functions of the generalised functional leverages and residuals. An application of the proposed procedures for investigating the effect of outliers on the relationship between transformation of the banking industry and the size of cooperative banks in Italy over a period of 14 years is presented.

KEYWORDS

functional data, functional regression model, geostatistics, hat matrix, high leverage, outlier, spatial dependence structure

1 | INTRODUCTION

In recent years, with the progress of high-performance technologies, there has been a considerable need for statistical models describing the evolution of phenomena in time and in space-time able to understand the simultaneous behaviour of several variables. In a plethora of cases, the raw observations can be viewed as a (continuous) curve, and thus be treated as functional data. References 1-3, as key standard references in this field, propose a large number of statistical models that require the specification of the dependence structure amongst variables. This dependence can be considered purely functional or either spatial-functional, depending on the aim of the analysis and on the application itself. Two such examples are the following. Assume we want to model the influence of climate on biodiversity, as in Reference 4, or to predict temperature and precipitation under climate change scenarios, as in Reference 5. These two examples show the need for tools able of handle the complexity in the relation amongst functional variables, the former by taking into account the underlying spatial dependence, and the latter without considering such dependence.

In both cases, the functional regression model[6] and its generalisation for spatially dependent functional data[7] serve this purpose. However, functional regression models come into trouble when there are curves that largely deviate from the other ones. This happens when there are curves that are contaminated in some way (e.g., an error in the experimental procedure), or they may represent a rare case.[8]

Although the literature on functional regression is extensively investigated (a good review can be found in Reference 9), very few recent contributions deal with the problem of outliers for robust functional regression. References 10-12 investigate the theoretical properties of robust estimators for the regression coefficients but differ in the estimation process.

In particular, Reference 10 proposes a Bayesian method in the context of robust functional mixed models (R-FMM) to perform robust functional regression. Reference 11 develops a robust version of the splines-based estimation method in Reference 13. Finally, Reference 12 presents a robust procedure by using outlier-resistant loss functions in the functional linear regression problem, and computes these robust estimates by using an iteratively reweighted penalised least-squares algorithm. Few other works deal with the problem of influential observations in estimation and prediction of the functional linear model with scalar response; see, for example, References 14-16. References 14 and 15 propose a functional version of the Cook's distance in the case where the predictors are real or functional and the responses are functional. Reference 16 introduces a set of statistics that seem to be useful in detecting which observations have strong influence and a smoothed bootstrap-based method to estimate the quantiles of the influence measures.

These methods do not deal with the problem of leverage detection when the curves are spatially correlated. We indeed focus our attention on the detection of spatially referenced leverage curves as potential outliers by generalising the criterion used in the general regression model[17] for both pointwise and total functional regression models.

In functional data analysis, we can observe two different types of outliers: shape and magnitude outliers. Shape outliers may be defined as those curves that exhibit a different shape from the rest of the sample.[18] Shape outliers are often masked, and thus they are difficult to be detected. On the other hand, magnitude curves are such that show a much larger magnitude than the rest of the curves, and are easier to identify.[19] According to our procedure, high leverage curves are potential outliers that may influence the estimation of the functional regression coefficients, and may also cause the standard errors of the regression coefficients be much smaller than they should be if these curves were excluded. A potential outlier could be a magnitude or shape outlier since both can produce more variability in the model estimation with a consequent minor standard error in the estimation of the coefficients. This definition implies that one should be able to order functional observations according to some measure of their high influence.

In this article, we focus on the following contribution. We first generalise the notion of leverage values for the Hat matrix in the more classical regression model to the context of pointwise and total functional regression models. Then we extend the leverage concept to the case of spatially correlated curves. In particular, we define the projection matrix for both models in the functional framework by considering a criteria to search for observations that typically show a high leverage when the functional observations are spatially correlated curve[20,21] and extend the proposed leverage detection criterion to this case.

The article is organised as follows. Section 2 proposes leverage detection for the case of functional total regression models for independent and spatially correlated errors. Section 3 shows an intensive simulation study, and a real data analysis comes in Section 4. The article ends with some conclusions and a discussion in Section 5.

2 | LEVERAGE IN LINEAR REGRESSION MODELS

The aim of this section is to introduce a criteria for *leverage* detection in functional total regression models for independent and spatially correlated errors. As well-known in the literature,[22] the least-squares projection matrix, also called the Hat matrix, plays a key role for leverage detection. Before presenting our proposal, we review the basics of leverage detection in the regression model.

2.1 | Classical multiple regression model

In matrix notation, the classical multiple regression model is written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

with $\mathbf{Y} \sim \text{NMV}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} = \sigma^2\mathbf{I})$ a multivariate Gaussian random variable, and $\boldsymbol{\epsilon}$ a residual Gaussian variable. In this context, the Hat matrix is defined as follows

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{2}$$

Many authors suggest as a rule of thumb that a point with $h_{ii} > 2\overline{h}$ ($\overline{h}$ is the average of the h_{ii} values) is a high leverage point[23,24] (for more details see the supplementary material).

Under heteroscedasticity $\boldsymbol{\Sigma} \neq \sigma^2 I$ is not diagonal, and the Hat matrix is defined as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}. \tag{3}$$

H in (3) is not symmetric but the other properties still hold. When $\mathbf{Y} \sim \text{NMV}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is not diagonal (correlated data) using (3) is not appropriate. Several authors have proposed alternatives for identifying leverages under this scenario. Let $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}$. According to Reference 25, the ith observation is a high leverage if $h^*_{ii} = \frac{h_{ii}}{\sum_{i=1}^n h_{ii}} > \frac{2}{n}$, where $h_{ii} = \frac{p_{ii}}{\sigma^{ii}}$, with p_{ii} and σ^{ii} the main diagonal elements of matrices $\boldsymbol{P}$ and $\boldsymbol{\Sigma}^{-1}$, respectively.

2.2 | Functional regression models with scalar response with independent and spatially correlated errors

The set $(X(t), Y)_D = \{(X_s(t), Y_s),\ s \in D \subset \mathbb{R}^d,\ t \in [a,\ b] \subset \mathbb{R}\}$ is called a bivariate spatial functional-scalar stochastic process or a bivariate functional-scalar random field, if for each $s \in D$, $(X_s(t), Y_s)$ is a paired of functional and scalar variables. A subset of this process is defined by n paired functional and scalar variables $((X_{s_1}(t), Y_{s_1}), \dots, (X_{s_n}(t), Y_{s_n}))$ with $(s_1, \dots, s_n)$ a n-tuple of sites in D. Assume we have a realisation of a bivariate functional-scalar random field $(X(t), Y)_D = \{(X_{s_i}(t), Y_{s_i}), i = 1, \dots, n\}$. A functional linear model with scalar response is given by the expression

$$Y_{s_i} = \alpha + \int_T X_{s_i}(t)\beta(t)dt + \epsilon_{s_i}, \quad i = 1, \dots, n, \tag{4}$$

with α an overall intercept parameter, and ϵ_{s_i} observations of an error or residual term. It is usually assumed that the sampling trajectories $X_{s_i}(t)$ and $\beta(t)$ are square integrable functions in a Hilbert space and they are generated by basis functions $\boldsymbol{\psi}(t) = (\psi_1(t), \dots, \psi_k(t))^T$, and $\boldsymbol{\theta}(t) = (\theta_1(t), \dots, \theta_k(t))^T$, respectively.[1] Thus, we can write the model as follows

$$Y_{s_i} = \alpha + \int_T \boldsymbol{X}^T_{\boldsymbol{s_i}}\boldsymbol{\psi}(t)\boldsymbol{\theta}^T(t)\boldsymbol{b}dt + \epsilon_{s_i}. \tag{5}$$

In matrix notation (for more details see the supplementary material), the model is equivalent to

$$\mathbf{Y} = \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{6}$$

with $\mathbf{Y} \sim \text{NMV}(\mathbf{Z}\boldsymbol{b}, \sigma^2\mathbf{I})$ a multivariate Gaussian random variable and $\mathbf{Z}$ is related to the representation in terms of basis functions of the functional data. In this case the Hat matrix is defined as follows

$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T. \tag{7}$$

Then, following the criterion by Reference 24, the ith observation $(X_{s_i}(t), Y_{s_i})$ is a *leverage* if h_{ii}, the ith element in the principal diagonal of the Hat matrix, exceeds $2\overline{h}$.

When curves are spatially correlated, which means that the errors are spatially correlated, $\mathbf{Y} \sim \text{NMV}(\mathbf{Z}\boldsymbol{b}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ the spatial covariance matrix. Defining $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}$, the ith observation is a high leverage if $h^*_{ii} = \frac{h_{ii}}{\sum_{i=1}^n h_{ii}} > \frac{2}{n}$, where $h_{ii} = \frac{p_{ii}}{\sigma^{ii}}$, with p_{ii} and σ^{ii} the main diagonal elements of matrices $\boldsymbol{P}$ and $\boldsymbol{\Sigma}^{-1}$, respectively.

2.3 | Functional regression models with functional response with independent and spatially correlated errors

We now consider the case of a more complex model which is known as the total model. It is specified by a dependent georeferenced functional variable and an independent functional variable. We focus on two different situations, when the errors are independent and when they are spatially correlated.

Let $(X(t), Y(\tau))_D = \{(X_s(t), Y_s(\tau),\ s \in D \subset \mathbb{R}^d,\ t \in [a,\ b] \subset \mathbb{R},\ \tau \in [c,\ d] \subset \mathbb{R}\}$ be a bivariate spatial functional stochastic process or a bivariate functional random field. For each $s \in D$, $(X_s(t), Y_s(\tau))$ is a paired of functional variables.

Again, a subset of this process is defined by n paired functional variables $((X_{s_1}(t), Y_{s_1}(\tau)), \dots, (X_{s_n}(t), Y_{s_n}(\tau)))$ with $(s_1, \dots, s_n)$ a n-tuple of sites in D. Assume we have a realisation of a bivariate functional random field $(X(t), Y(\tau))_D = \{(X_{s_i}(t), Y_{s_i}(\tau)), i = 1, \dots, n\}$.

Then a functional linear model is given by the expression (see Reference 1)

$$Y_{s_i}(\tau) = \alpha(\tau) + \int_T X_{s_i}(t)\beta(t,\ \tau)dt + \epsilon_{s_i}(\tau), \quad i = 1, \dots, n, \tag{8}$$

with $\alpha(\tau)$ an overall intercept parameter. It is usually assumed that the sampling trajectories $Y_{s_i}(\tau),\ X_{s_i}(t)$ and $\beta(t,\ \tau)$ are square integrable functions in a Hilbert space and they are generated by function basis $\boldsymbol{\phi}(\tau) = (\phi_1(\tau), \dots, \phi_k(\tau))^T$ and $\boldsymbol{\psi}(t) = (\psi_1(t), \dots, \psi_k(t))^T$, respectively.[1] Thus, we can write the model as follows

$$\boldsymbol{Y}_{s_i}^T\boldsymbol{\phi}(\tau) = \boldsymbol{a}^T\boldsymbol{\phi}(\tau) + \left(\int_T \boldsymbol{X}_{s_i}^T\boldsymbol{\psi}(t)\boldsymbol{\psi}^T(t)\boldsymbol{B}dt\right)\boldsymbol{\phi}(\tau) + \boldsymbol{e}_{s_i}^T\boldsymbol{\phi}(\tau). \tag{9}$$

In matrix notation (for more details see the supplementary material), the model is equivalent to

$$\boldsymbol{Y} = \boldsymbol{Z\mathcal{B}} + \boldsymbol{E}, \tag{10}$$

with $\boldsymbol{Y} \sim \mathrm{NMV}(\boldsymbol{Z\mathcal{B}}, \sigma^2\mathbf{I})$ a multivariate Gaussian random variable and $\mathbf{Z}$ is related to representation in basis functions. In this case the Hat matrix is defined as follows

$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T. \tag{11}$$

Following the criterion by Reference 24, the ith curve $(X_{s_i}(t), Y_{s_i}(\tau))$ is a *leverage* if h_{ii}, the ith element in the principal diagonal of the Hat matrix, exceeds $2\overline{h}$.

When curves are spatially correlated, which means that the error term is spatially correlated, $\mathbf{Y} \sim \mathrm{NMV}(\boldsymbol{Z\mathcal{B}}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ the spatial covariance matrix. Again, as in Section 2.2, defining $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}$, the ith observation is a high leverage if $h_{ii}^* = \frac{h_{ii}}{\sum_{i=1}^n h_{ii}} > \frac{2}{n}$, where $h_{ii} = \frac{p_{ii}}{\sigma^{ii}}$, with p_{ii} and σ^{ii} the main diagonal elements of matrices $\boldsymbol{P}$ and $\boldsymbol{\Sigma}^{-1}$, respectively. The h_{ii} measures the distance of the ith curve to the centre of the functional space. Indeed, large h_{ii} elements reveal observations that are potentially influential. The spatial correlation structure we consider is enclosed in the matrix $\boldsymbol{\Sigma}$ which can encode different types of spatial correlation structures such as Exponential or Gaussian models.

The procedures for identifying leverages given above depend on the matrix $\boldsymbol{\Sigma}$ which is unknown in practice. Then a feasible generalised least squares estimator[26] can be used replacing Σ by $\hat{\Sigma}$. An illustration of this methodology in the context of experimental design for spatially correlated functional data is shown in Reference 27.

According to this rule, leverages only take into account the extremeness of the curves in terms of shape and magnitude, but we note that a high leverage curve may or may not be an influential one. A curve that has high leverage is different from a curve that has high influence on the regression analysis. In particular, a leverage curve identifies a potential curve that has high leverage and thus strong influence on the regression analysis. Finally, curves which are isolated in the functional space will have high leverage, that is, they will have a large h_{ii}^*. Thus these can be thought of as magnitude outliers in the functional space. Therefore, the smaller the leverage is, the better the prediction will be. All the same results can be obtained for the functional regression concurrent model. For simplicity, it is not included in this section because it is a particular case of the total model in Equation (8), when $t = \tau$.

3 | SIMULATION STUDY

In this section, we study the performance of the proposed method to detect high leverages in functional regression models. We simulate two different cases, one related to i.i.d. errors and one related to spatially correlated errors. Our main aim is to show how the elements h_{ii} of the matrices H in Equations (7) and (11) and the corresponding h_{ii}^* identify leverages in functional regression models both with scalar (FRMSR) and functional response (FRMFR). The models with spatial dependence have been analysed when the error structure is assumed known a priori (scenario A) and when an estimation procedure is defined for the error structure (scenario B).

Assume the models given in (4) and in (8). We have generated $n = 100$ curves for a given set of parameters and variables, considering both i.i.d. and spatially correlated errors, and have added four curves distant from the rest of observations in terms of magnitude or shape. For each model, we have simulated data according to the following scheme in R.[28]

We set *basis functions*, *scalar and functional parameters*, *functional covariates*, *response values*, *errors of the model* as follows:

- *Basis functions*: For both models, $X_{s_i}(t)$, $\beta(t)$ $\beta(t,\tau)$, $Y_{s_i}(\tau)$, $\alpha(\tau)$ and $\epsilon_{s_i}(\tau)$ are expanded in terms of the same B-spline basis functions $\boldsymbol{\phi}(t) = (\phi_1(t), \dots, \phi_k(t))^T$, with $k = 7$, and $\tau$and $t \in [0, 1]$.
- *Scalar and functional parameters*: In the FRMSR model, we consider the scalar and the functional parameters $\alpha = \beta_0 = 10$, $\beta(t) = \sum_{j=1}^{k} b_j\phi_j(t) = \boldsymbol{\phi}^T(t)\mathbf{b}$, with $\mathbf{b} = (30, 26, 28, 30, 28, 26, 30)^T$. In the FRMFR, we set $\alpha(\tau) = \beta_0(\tau) = (28.96,\ 22.65, 23.72, 25.72, 29.08, 22.01, 28.98)^T$, and

$$\beta(t,\tau) = \sum_{j=1}^{k}\sum_{h=1}^{k}\phi_h(\tau)b_{j,h}\phi_j(t) = \boldsymbol{\phi}^T(t)\mathbf{B}\boldsymbol{\phi}^T(\tau),$$

  where $\mathbf{B} \sim \text{NMV}\left(\boldsymbol{\mu}_{k\times 1}, \mathbf{I}_{k\times k}\right)$ is a matrix of B-spline basis generated with $\boldsymbol{\mu}_{k\times 1} = (10, 11, 12, \dots, 16)^T$.
- *Functional covariates*: For both models, this covariate is defined as $X_{s_i}(t) = \sum_{j=1}^{k} X_{s_ij}\phi_j(t)$, with $\phi_j(t)$ defined as above. The coefficients $X_{s_ij}$with $i = 1, \dots, n$ and $j = 1, \dots, k$ are generated from $\mathbf{X} \sim \text{NMV}\left(\boldsymbol{\mu}_{k\times 1}, \mathbf{I}_{k\times k}\right)$, where $\boldsymbol{\mu}_{k\times 1} = (10, 11, 12, \dots, 16)^T$. To add some extreme functions (high leverage curves) and mix them with the rest of curves, the curves $X_{s_1}(t)$ and $X_{s_{100}}(t)$ are replaced by $X^l_{s_1}(t)$ and $X^l_{s_{100}}(t)$, where $X^l_{s_1}(t) = \sum_{j=1}^{k} X^l_{s_1j}\phi_j(t)$, and $X^l_{s_{100}}(t) = \sum_{j=1}^{k} X^l_{s_{100}j}\phi_j(t)$, with $X^l_{s_1j} = X_{s_1j} + a_1$, and $X^l_{s_{100}j} = X_{s_{100}j} - a_1$, and $a_1 \sim \text{N}(2.8, 1)$, a Gaussian random variable. In addition, the curves $X_{s_2}(t)$ and $X_{s_{99}}(t)$are replaced by:
  - $X^l_{s_2}(t) = \sum_{j=1}^{k} X^l_{s_2j}\phi_j(t)$, with $X^l_{s_2j} = (10, 7.5, 14.9, 9.9, 18.8, 10.4, 16.9)^T$.
  - $X^l_{s_{99}}(t) = \sum_{j=1}^{k} X^l_{s_{99}j}\phi_j(t)$, with $X^l_{s_{99}j} = (10, 12.7, 11, 11.4, 17.7, 12.7, 15.1)^T$.

  These curves are obtained by simulating a set of sinusoidal curves with a different shape from the rest of date.
- *Response values*: We use model (4) for the FRMSR, and model (8) for FRMFR to generate the values of the response Y_{s_i} and $Y_{s_i}(\tau)$ for $i = 1, \dots, n$, considering independence and spatial dependence structures.
- *Errors of the model*: $\boldsymbol{\epsilon}_D = (\epsilon_{s_1}, \dots, \epsilon_{s_n})^T \sim \text{NMV}(0, \boldsymbol{\Sigma})$ for the FRMSR and $\boldsymbol{\epsilon}_D(\tau) = (\epsilon_{s_1}(\tau), \dots, \epsilon_{s_n}(\tau)) \sim \text{NMV}(0, \boldsymbol{\Sigma})$for the FRMFR are considered, under two cases:
  - Case 1, $\boldsymbol{\Sigma} = \sigma^2_\epsilon \mathbf{I}_{n\times n}$ when the errors are independent.
  - Case 2, $\boldsymbol{\Sigma} = (\sigma_{qr})_{n\times n}$ when the errors are spatially correlated.

  In the second case, $\sigma_{qr} = \sigma^2_\epsilon \exp(\frac{-h}{\phi})$, with $h = ||s_q - s_r||$, and $s_q = (x_q, y_q)$ and $s_r = (x_r, y_r)$, $q, r = 1, \dots, n$, the coordinates of the points over a regular grid for which we observe our functional sample. We take $\sigma_\epsilon = 10$ and $\phi = 8$, which corresponds approximately to the 75% of the maximum distance in the grid, and therefore this value is indicative of a strong spatial dependence.

Throughout the simulation study, we have assumed known the type of basis functions (a B-spline basis) and we have chosen the optimal number of such basis functions ($k = 7$) by using a cross validation criteria.[1] In practice, this is an additional step of the statistical analysis.

In Case 1 with independent errors, for the simulated data, the parameters β_0 and $\beta(t)$ for the model FRMSR are estimated by

$$\hat{\beta}(t) = \sum_{j=1}^{k}\hat{b}_j\phi_j(t) = \boldsymbol{\phi}^T(t)\hat{\mathbf{b}}, \tag{12}$$

for $t \in [0, 1]$, and with $\hat{\mathbf{b}}$ obtained by using some regularisation method.[1] For the simulated data, the parameters $\beta_0(\tau)$ and $\beta(t,\tau)$ for the model FRMFS are estimated by

$$\hat{\beta}(t,\tau) = \sum_{j=1}^{k}\sum_{h=1}^{k}\phi_h(\tau)\hat{b}_{j,h}\phi_j(t) = \boldsymbol{\phi}^T(t)\hat{\mathbf{B}}\boldsymbol{\phi}^T(\tau), \tag{13}$$

for $t \in [0, 1]$, $\tau \in [0, 1]$ and with $\hat{\mathbf{B}}$ obtained by using some regularisation method.[1]

The hat matrix **H** was constructed as in (7) and (11) where $\boldsymbol{\Sigma} = \boldsymbol{I}$, respectively, for the model FRMSR and FRMFR. For the high leverage detection, the rule we have used is $h_{ii} > 2\overline{h}$.

Under Case 2, $\boldsymbol{\Sigma}$ was used to describe the spatial variability. However, $\boldsymbol{\Sigma}$ is unknown in practice so two scenarios have been considered. In scenario A, the error structure is assumed known a priori, the analysis is performed under the hypothesis $\boldsymbol{\Sigma} = \sigma_\epsilon^2 \exp(\frac{-h}{\phi})$. In scenario B, $\boldsymbol{\Sigma}$ is unknown and the variogram is used to obtain an approximation of the estimation of the spatial structure.[29] The procedure that we consider to estimate Σ comes from using the model FRMFR for the estimate of the error without spatial dependence. Then the functional residuals of this model are used to estimate the matrix $\hat{\boldsymbol{\Sigma}}$ by using the trace-variogram function.[30]

Let $(\epsilon_{s_i}(\tau), \epsilon_{s_j}(\tau))$ be the random errors of the model in sites $s_i,\ s_j \in D$, with D the spatial domain. The trace-variogram function between these random functions is defined as

$$\gamma_{s_i,s_j}(h) = \frac{1}{2}\mathbb{E}\left[\int_T (\epsilon_{s_i}(\tau) - \epsilon_{s_j}(\tau))^2 dt\right]. \tag{14}$$

Given the residuals $e_{s_1}(t), e_{s_2}(t), \dots, e_{s_n}(t)$ of the model assuming independent errors, we use the method of moments to estimate $\gamma_{s_i,s_j}(h)$ in 14 given by

$$\hat{\gamma}_{s_i,s_j}(h) = \frac{1}{2|N(h)|}\sum_{s_i,s_j \in N(h)} \int_T (e_{s_i}(\tau) - e_{s_j}(\tau))^2 dt, \tag{15}$$

where $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$, and $|N(h)|$ is the number of distinct elements in $N(h)$. Once we have estimated the trace-semivariogram for a list of M values h_m, we fit a parametric model[31] (spherical, Gaussian, exponential, or Matérn, for instance) to the points $(h_m, \hat{\gamma}(h_m))$, $m = 1, \dots, M$, as if they were obtained in the classic one-dimensional geostatistical setting. Given $\hat{\gamma}_{s_i,s_j}(h)$ we estimate the elements of the covariance matrix by $\hat{\boldsymbol{\Sigma}}_{i,j} = \hat{\sigma}^2 - \hat{\gamma}_{s_i,s_j}(h)$, with $\hat{\sigma}^2$ also obtained from the estimation of the trace-variogram function (the tracevariogram function value when $h \to \infty$).

Thus $\boldsymbol{\Sigma} \neq \boldsymbol{I}$ and the matrix **P** was constructed under both models FRMSR and FRMFR. The rule we defined for the high leverage detection curve is $h_{ii}^* = \frac{h_{ii}}{\sum_{i=1}^n h_{ii}} > \frac{2}{n}$, where $h_{ii} = \frac{p_{ii}}{\sigma^{ii}}$, with p_{ii} and σ^{ii} the main diagonal elements of matrices $\boldsymbol{P}$ and $\boldsymbol{\Sigma}^{-1}$, respectively.

3.1 | Simulation results

Table 1 shows a summary of results for both models FRMSR and FRMFR with i.i.d. errors and spatially correlated errors. We repeated the procedure $p = 200$ times and obtained the mean values of average residual, number of leverage curves (NLC), and the percentage of times in which all the four outliers are identified (POI, percentage outlier identified).

In both cases and scenarios, the h_{ii} and h_{ii}^* values correspond to the contaminated pairs $(X_{s_i}(t), Y_{s_i})$ $(X_{s_i}(t), Y_{s_i}(\tau))$ for $i = 1, 2, 99{,}100$. All the others h_{ii} values are clearly below the threshold or just slightly higher.

Table 2 shows the estimated coefficients $\hat{b}_j^1$ for Case 1, and $\hat{b}_j^a$, $\hat{b}_j^b$ for Case 2 for scenarios A and B. The results show good performances. We repeated the estimation procedure $p = 200$ times, and obtained the average values as

$$\hat{b}_j^l = \frac{1}{p}\sum_{r=1}^{p} \hat{b}_{r,j}^l, \tag{16}$$

with $j = 0, \dots, k$ and $l = 1, a, b$. We note that these estimated means are in all cases very close to the b_j values which is an indication that $\hat{\beta}(t)$ and $\hat{\beta}_c(t)$ (scenarios A and B) are unbiased estimators of the functional parameter $\beta(t)$.

In both cases, we use the role $h_{ii} > 2\overline{h}$, where $\overline{h} = \frac{(\sum_{i=1}^n h_{ii})}{n}$. In Case 1, we now that $\frac{(\sum_{i=1}^n h_{ii})}{n} = \frac{k+1}{n} = \frac{tr(\mathbf{H})}{n}$ for the properties of **H** matrix. Thus the role depends on k (the dimension of basis function) and for k tending to infinity, there are no high leverage curves because the threshold grows up to infinity. The values of h_{ii} are all positive and satisfy the property $\sum h_{i,i} = rank(H) < \infty$ for the each value of k. In Case 2, the relation $\frac{(\sum_{i=1}^n h_{ii})}{n} = \frac{k+1}{n} = \frac{tr(\mathbf{H})}{n}$ is not true. We thus define $\frac{h_{ii}}{\sum_{i=1}^n h_{ii}} > \frac{2}{n}$ that is independent from k and dependent on n.

TABLE 1 Performances of each model FRMSR and FRMFR on 200 simulations for two scenarios in the two cases

Model	Case	Scenario	AResiduals	NLC	POI
FRMSR	1	$\boldsymbol{\Sigma} = \sigma_\epsilon^2 \mathbf{I}_{n\times n}$ (A)	2.88×10^{-13}	5	97%
		$\boldsymbol{\Sigma} \sim \hat{\boldsymbol{\Sigma}}$ (B)	0.2	6	92.5%
	2	$\boldsymbol{\Sigma} = \sigma_\epsilon^2 \exp(\frac{-h}{\phi})$ (A)	0.08	6	83.5%
		$\boldsymbol{\Sigma} \sim \hat{\boldsymbol{\Sigma}}$ (B)	0.1	6	81%
FRMFR	1	$\boldsymbol{\Sigma} = \sigma_\epsilon^2 \mathbf{I}_{n\times n}$ (A)	15	5	94%
		$\boldsymbol{\Sigma} \sim \hat{\boldsymbol{\Sigma}}$ (B)	2	6	74.6%
	2	$\boldsymbol{\Sigma} = \sigma_\epsilon^2 \exp(\frac{-h}{\phi})$ (A)	20	6	73.6%
		$\boldsymbol{\Sigma} \sim \hat{\boldsymbol{\Sigma}}$ (B)	3.66	6	72.1%

Note: Mean of average residuals (AResiduals), Average of number high leverage curves (NLC), and average of percentage of times in which all the added outliers are identified (POI) are shown.

TABLE 2 Theoretical coefficients $b_j, j = 0, \dots, k$ (used to define the functional parameter β_0 and $\beta(t)$) and their estimated values $\hat{b}_j^l, j = 0, \dots, 7$ for both cases, non-correlated data ($l = 1$) and spatially correlated data scenario A ($l = a$) and scenario B ($l = b$)

$\boldsymbol{b_j}$	$\hat{\boldsymbol{b}}_{\boldsymbol{j}}^{\boldsymbol{1}}$	$\hat{\boldsymbol{b}}_{\boldsymbol{j}}^{\boldsymbol{a}}$	$\hat{\boldsymbol{b}}_{\boldsymbol{j}}^{\boldsymbol{b}}$
10	10.0	10.2	10.1
30	24.5	29.5	29.5
26	34.2	26.1	26.1
28	21.9	27.9	27.9
30	33.6	30.2	30.2
28	25.4	27.8	27.0
26	25.2	26.0	26.0
30	32.5	30.3	30.3

Note: $\hat{b}_j^1$, $\hat{b}_j^a$ and $\hat{b}_j^b$ correspond to the estimated mean values (based on 200 simulations) of $\hat{b}_j$ values.

4 | APPLICATION

In the last years, the relationship between the development of a country's financial system with the country's economic development has been the object of a rich empirical and theoretical literature. In Italy, where deep regional economic differences exist, several studies have investigated the local (regional) aspects of the relationship between finance and growth. The present study aims at investigating the effect of outliers on the relationship between transformation of the banking industry over the period 2000–2014 at a provincial scale and the size of cooperative banks on the territory.

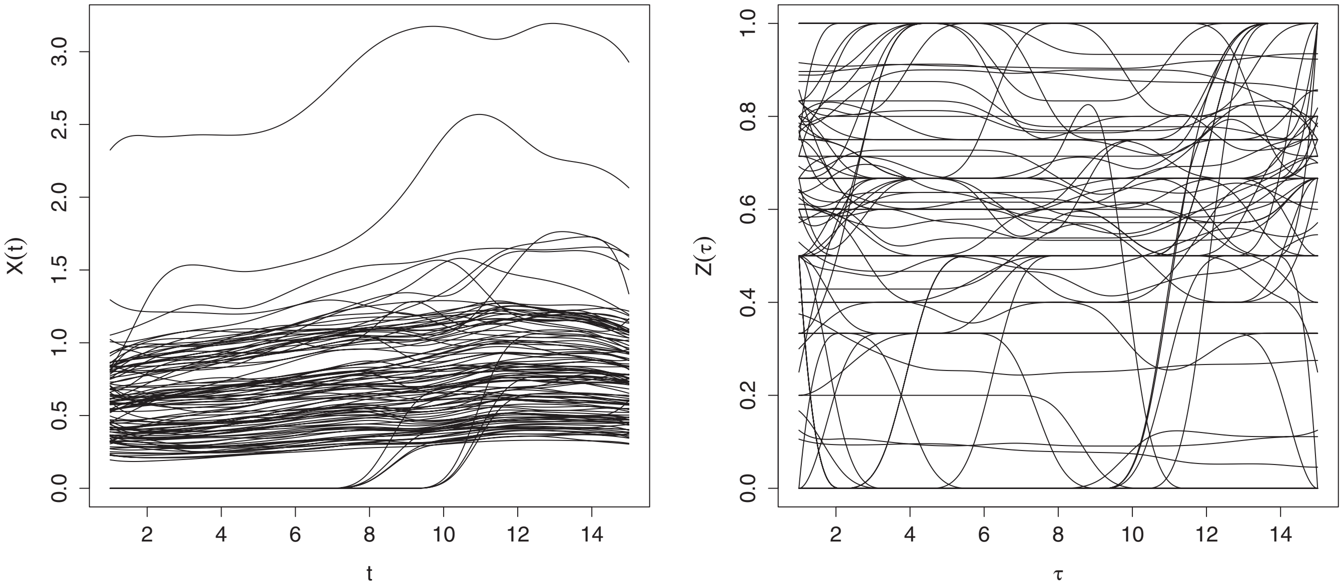

FIGURE 1 *Left*: Curves of $X_{s_i}(t)$ over t. $X_{s_i}(t)$ represents a functional indicator of financial development (measured by the ratio of total loans to non-financial firms over the total added value at time t). *Right*: $Z_{s_i}(\tau)$ is the per cent of small banks belonging to the BCC in the time. Both are observed on a monthly grid over 14 years

We thus considered data from ISTAT on the size classes of Italian firms. In particular, we analysed 112 Italian provinces and three variables per province, Y_{s_i}, $X_{s_i}(t)$ and $Z_{s_i}(\tau)$, with $i = 1, \dots, 112$, where Y_{s_i} is the log of added-value per capita in county s_i, $X_{s_i}(t)$ is a functional indicator of financial development (measured by the ratio of total loans to non-financial firms over the total added-value at time t), and $Z_{s_i}(\tau)$ is the per cent of small banks belonging to the Banche di Credito Cooperativo (credit cooperatives; henceforth BCC) at time τ. The functional regression of Y_{s_i} on $X_{s_i}(t)$ helps to capture how the added-value per capita is affected by changes in the structure of the banking system at a county level. In addition, the functional regression of $X_{s_i}(t)$ on $Z_{s_i}(\tau)$describes the positive impact of the presence of BCCs. Figure 1 shows both $X_{s_i}(t)$ and $Z_{s_i}(\tau)$.

Leverage curves in both functional models detect particular anomalous behaviours. The $X_{s_i}(t)$ functions are spatially correlated, as noted by the estimation of the variogram function in Figure 2. This means that the curves show a local correlation and thus dependence structure which brings light to the fact that the distribution of the financial development in Italy generally shows diverse characteristics at different spatial scales, and our model is able to take this into account. We indeed allow for the interaction between one of the main variables of interest (financial structure indicators) and the variables that reflect financial development. We note the following from our functional fitted models. First, the financial structure activity and efficiency have an effect on the spatial economic growth, however the size and the number of financial structures does not. Second, the positive impact of higher BBC development relative to banking sector development is reverted if the county financial structure is unbalanced. The robustness of the fitted models is checked by exploring the existence of leverage curves, and we use our approach to detect tentative leverage curves that might affect the sensitivity of our results.

Leverage curves correspond to those $X_{s_i}(t)$, $i = 1, \dots, 112$ in Figure 1 that markedly deviate from the others and have an influential effect on the statistical model. Indeed, a curve may be judged influential, and thus a leverage curve, if important features of the analysis are altered substantially changing the functional relationship. Although we have evidence of existence of a spatial structure, we have also considered the case of independent observations for comparison purposes.

Figure 3 reports the functional estimation of $\beta(t)$ for the 15 temporal instants (years) considered under the independent and the spatially correlated cases for the scalar functional model of Y_{s_i} on $X_{s_i}(t)$.

Figures 4 and 5 show the h^*_{ii} values, respectively, in the independent and the spatially correlated cases, with the corresponding upper limits to identify the leverage curves for the scalar functional model. We detect 15 leverages for the spatially correlated case, and 14 for the independent case. Twelve curves are common leverages for both cases which correspond to the following provinces Ancona, Bergamo, Firenze, Lodi, Milano, Monza-Brianza, Olbia-Tempio, Padova, Roma, Siena, Sondrio, Verona. Then we have three different provinces (Lucca, Macerata and Pisa) that are considered leverages

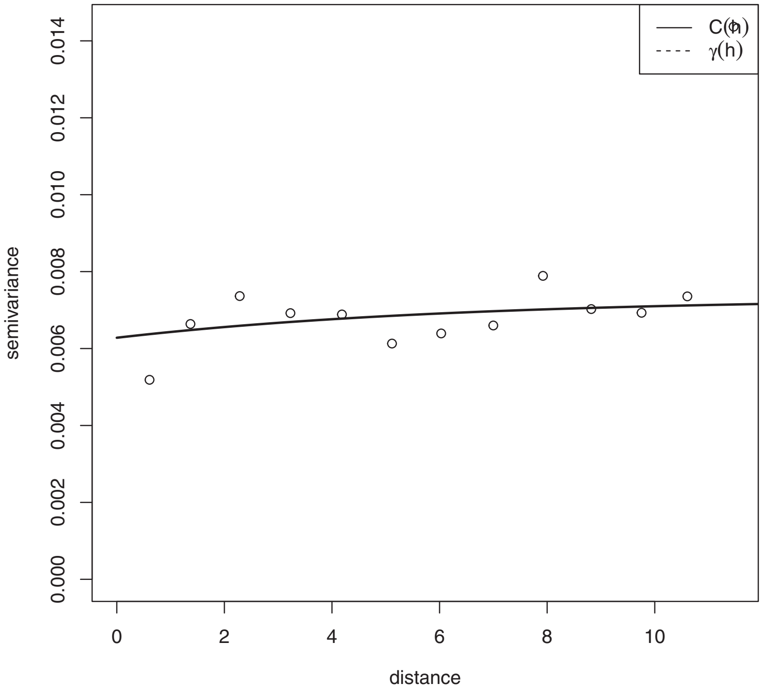

FIGURE 2 Estimated variogram functions for $X_{s_i}(t)$

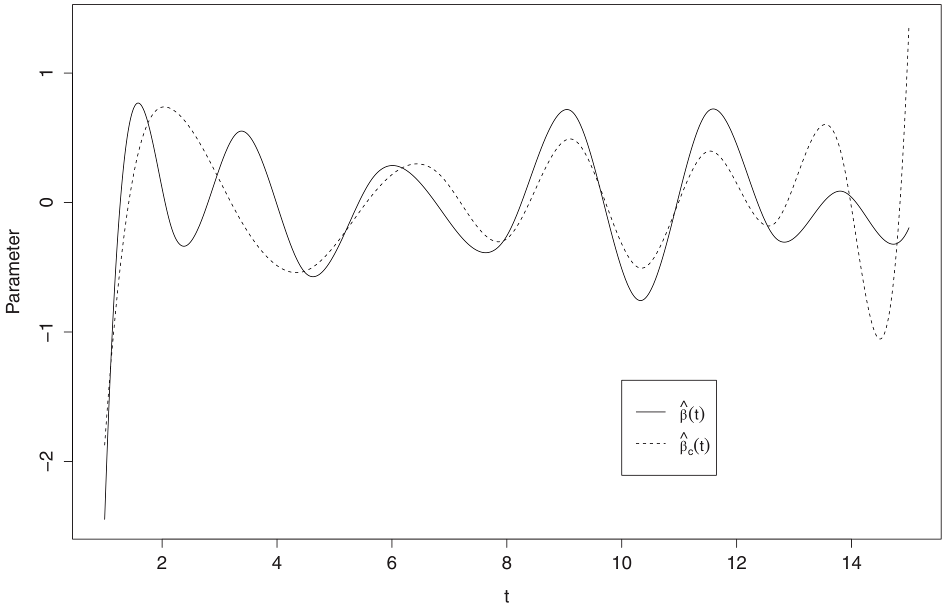

FIGURE 3 Estimated functional parameter β for the Italian financial data under the independent observations case, and under the spatially correlated situation for the scalar functional model

only under the spatially correlated case. These provinces are indeed close together and show correlation in space. The model under the assumption of independence cannot detect them as leverages, but when we take into account the spatial correlation these are highlighted as such outlying curves. These results show the good performance of our method in detecting leverage curves. From an economical point of view, we can conclude that the selected leverage provinces correspond to places on which the impact of the bank transformations has had a negative effect. According to several conducted studies on this topic, there is a positive relationship to banks' increased ability to geographically diversify their risks (subsequent to de-regulation) and, as a consequence, their greater willingness to supply credit to innovative firms.[32] According to this observation, the underlined three provinces present in this case an anomalous behaviour.

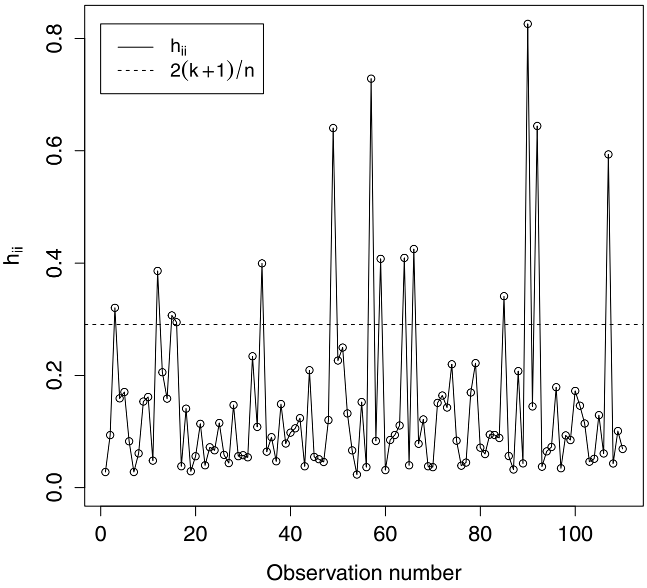

FIGURE 4 Leverages detection for the Italian financial data when the spatial correlation is not considered for the scalar functional model

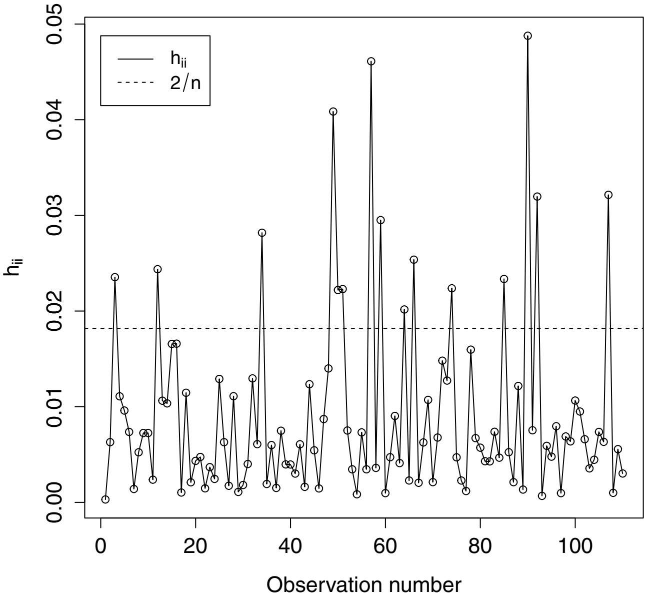

FIGURE 5 Leverages detection for the Italian financial data when the spatial correlation is considered for the scalar functional model

As mentioned above, Italy's small banks have faced a significant transformation of their competitive and regulatory environment over time, in line with what has happened in other European countries. This transformation has led to large-scale variability and this is reflected in what we have studied by the subsequent red total regression model with spatially correlated errors.

We used a total model to fit a functional regression between $X_{s_i}(t)$ and $Z_{s_i}(\tau)$. Indeed, this functional regression describes the positive impact of the presence of BCCs on financial development. Figure 6 shows the surface of the functional coefficients that point to this influence of the presence of BBC banks in Italy on the financial development along time. We note that this function is extremely variable. This is mainly due to the presence of curves describing major level of lending that can be considered leverage curves (see Figure 7) and these correspond to the provinces of Alessandria, Cagliari, Lecco, Livorno, Lodi, Lucca, Massa Carrara, Matera, Messina, Monza-Brianza, Pavia, Pescara, Piacenza, Pisa, Reggio Calabria, Savona, Siracusa, Terni, Verbano-Cusano-Ossola.

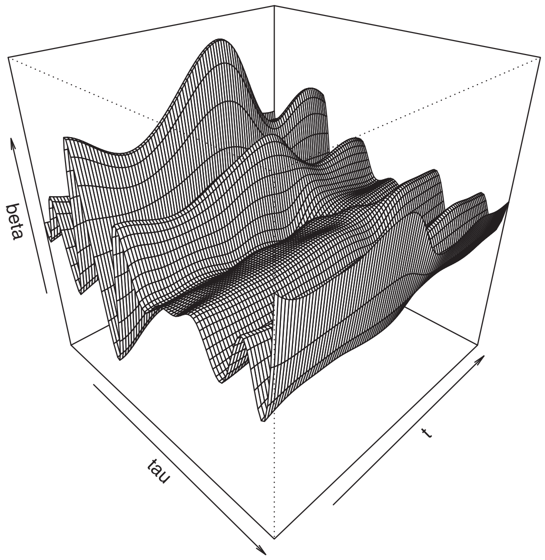

FIGURE 6 Surface of the functional coefficients $\beta(t,\ \tau)$ in the total model, with spatial dependence, between $X_{s_i}(t)$ and $Z_{s_i}(\tau)$

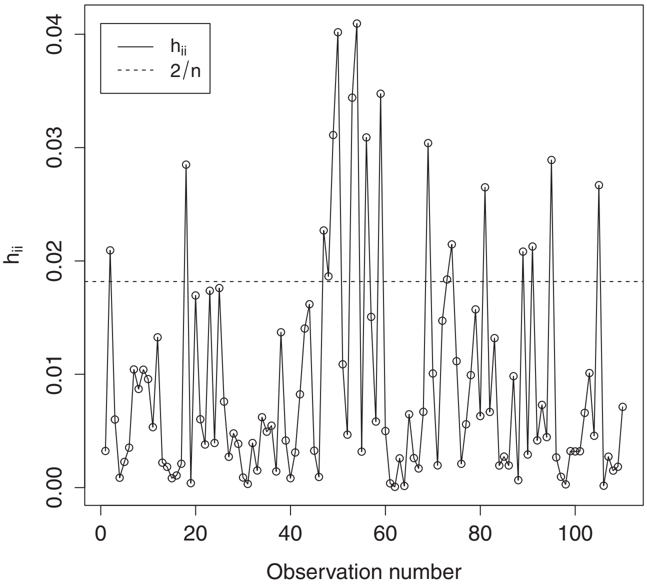

FIGURE 7 Leverages for the functional total model, with spatial dependence

These selected countries show a behaviour which is far more oscillated than the rest of the counties in terms of size and financial developments. It means that in high-income counties, financial structure in terms of size and activity are associated with more economic development. These counties have more influence in the model estimation and this is reflected in the oscillate behaviour of the surface $\beta(t, \tau)$.

5 | DISCUSSION

A usual way to have more accurate predictions in a functional regression context is by first detecting the presence of possible leverage curves, and then deciding if they come from typing errors or they carry outlying but important information for the problem at hand. We consider the problem of leverage curve detection under the presence of spatially correlated errors. It is the case that a small subset of curves could have an influence on the model coefficients. In an extreme case, the parameter estimates could depend on the influential subset of curves much more than on the majority of the curves. A statistical model has to be a representative of all the sample curves and not only of just a few. Thus our aim is to find these influential curves and assess their impact on the model by considering how they affect the general functional regression model, in particular when the curves are spatially correlated.

We thus generalise the concept of projection matrix and some of the useful classical criteria for selecting high leverage curves to the functional framework to detect leverages considering the more usual case of independent observations and, additionally, considering the case where the curves are spatially correlated. Our proposed approach, to the best of our knowledge, is the first to provide a measure of 'leverages' for curves in presence of spatially correlated errors in a general functional regression framework. As it has been shown by simulations and a real data analysis, our approach is effective in rightly leveraging curves, even in the presence of substantial masking.

An extension of the presented procedure to the geographically weighted regression model will be object of further research.

ACKNOWLEDGEMENT

Open Access Funding provided by Universita degli Studi della Campania Luigi Vanvitelli within the CRUI-CARE Agreement.

DATA AVAILABILITY STATEMENT

https://www.istat.it

ORCID

Elvira Romano https://orcid.org/0000-0001-8998-7099

REFERENCES

1. Ramsay JO, Silverman BW. *Functional Data Analysis*. Springer Series in Statistics 2nd Ed. New York, NY: Springer; 2005. https://link.springer.com/book/10.1007/b98888
2. Ramsay JO, Silverman BW. *Applied Functional Data Analysis: Methods and Case Studies*. Springer New-York: Springer; 2002. https://link.springer.com/book/10.1007/b98886
3. Ferraty F, Vieu P. Non-parametric functional data analysis. *Theory and Practice*. Springer Series in Statistics New-York: Springer; 2006. https://link.springer.com/book/10.1007/0-387-36620-2
4. Bel L, Bar-Hen A, Cheddadi R, Petit R. Spatio-temporal functional regression on paleoecological data. *J Appl Stat*. 2010;38:695-704. https://doi.org/10.1080/02664760903563650
5. Ghumman AR, Ateeq-ur-Rauf AU, Haider H, Shafiquzamman M. Functional data analysis of models for predicting temperature and precipitation under climate change scenarios. *Water & Climate Change*. 2020;11(4):1748-1765. https://doi.org/10.2166/wcc.2019.172
6. Cardot H, Sarda P. Linear regression models for functional data. In: Sperlich S, Härdle W, Aydınlı G, (eds). *The Art of Semiparametrics*. Contributions to Statistics Physica-Verlag HD; 2006. https://link.springer.com/chapter/10.1007/3-7908-1701-5_4
7. Muller HG, Stadtmüller U. Generalized functional linear models. *Ann Stat*. 2005;33(2):774-805.
8. Gervini D. Detecting and handling outlying trajectories in irregulary sampled functional datasets. *Ann Appl Stat*. 2009;3(4):1758-1775.
9. Paganoni A, Sangalli L. Functional regression models: some directions of future research. *Stat Model*. 2017;17:1-6.
10. Zhu H, Brown P, Morris J. Robust adaptive functional regression in functional mixed model framework. *J Am Stat Assoc*. 2011;106:1167-1179.
11. Moronna R, Yohai V. Robust functional linear regression based on splines. *Computational Statistics and Data Analysis*. 2013;65:46-55. https://doi.org/10.1016/j.csda.2011.11.014
12. Shin H, Lee S. An RKHS approach to robust functional linear regression. *Statistica Sinica*. 2016;26:255-272. http://dx.doi.org/10.5705/ss.202014.0063
13. Crambes C, Delsol L, Laksaci A. Robust nonparametric estimation for functional data. In: Niang SD, Ferraty F, (eds). *Functional and Operatorial Statistics*; Contribution to Statistics Physica-Verlag HD; 2008:109-116. https://doi.org/10.1007/978-3-7908-2062-1_18
14. Shen Q, Xu H. Diagnostics for linear models with functional responses. *Technometrics*. 2007;49(1):26-33. https://www.jstor.org/stable/25471272
15. Chiou JM, Muller HG. Diagnostics for functional regression via residual processes. *Computational Statistics and Data Analysis*. 2007;51:4849-4863. https://doi.org/10.1016/j.csda.2006.07.042
16. Febrero-Bande M, Galeano P, Gonzalez-Manteiga W. Measures of influence for the functional linear model with scalar response. *Journal of Multivariate Data Analysis*. 2010;101:327-339. https://doi.org/10.1016/j.jmva.2008.12.011
17. Shi L, Chen G. Influence measures for general linear models with correlated errors. *Am Stat*. 2009;63(1):40-42.

18. Arribas-Gil A, Romo J. Shape outlier detection and visualization for functional data. *Biometrics*. 2014;15:603-619. https://pubmed.ncbi.nlm.nih.gov/24622037/
19. Romano E. Clustering and outlier detection. New proposals and open problems. *Atto della XLIV Riunione Scientifica della Societá Italiana di Statistica*. Vol 25-26. Universita di la Calabria; 2008:189-196.
20. Delicado P, Giraldo R, Comas C, Mateu J. Statistics for spatial functional data: some recent contributions. *Environmetrics*. 2010;21:224-239. https://doi.org/10.1002/env.1003
21. Mateu J, Romano E. Advances in spatial functional statistics. *Stoch Environ Res Risk Assess*. 2017;31:1-6. https://doi.org/10.1007/s00477-016-1346-z
22. Kutner MH, Nachtsheim C, Neter J, Li W. *Applied Linear Statistical Models*; Operations and decision sciences. Fifth. Boston: McGraw-Hill/Irwin; 2005.
23. Renche A, Schaalje B. *Linear Models in Statistics*. John Wiley Sons; 2008. https://www.wiley.com/en-us/9780471754985
24. Belsley D, Kuh E, Welsch R. *Regression Diagnostics: Identifying Influential Data, and Sources of Collinearity*. John Wiley & Sons; 2004. https://www.wiley.com/en-us/9780471754985
25. Ho L, Valliant R. Survey weighted hat matrix and leverages. *Surv Methodol*. 2009;35(1):15-24. https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10881-eng.pdf
26. Fomby T, Hill R, Johnson S. *Advanced Econometric Methods*. Springer Science & Business Media; 2012.
27. Aristizabal J, Giraldo R, Mateu J. Analysis of variance for spatially correlated functional data: application to brain data. *Spat Stat*. 2019;32:100381.
28. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2005. http://www.R-project.org
29. Schabenberger O, Gotway CA. *Statistical Methods for Spatial Data Analysis*. Taylor & Francis Group; 2005.
30. Giraldo R, Delicado P, Mateu J. Continuous time-varying kriging for spatial prediction of functional data: an environmental application. *J Agric Biol Environ Stat*. 2010;48:66-82.
31. Cressie N. *Statistics for Spatial Data*. John Wiley and Sons; 1993.
32. Butzbach OKE Italian banking regulation and the legal obstacles to corporate governance convergence; Vol. 1, 2019.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Romano E, Giraldo R, Mateu J, Diana A. High leverage detection in general functional regression models with spatially correlated errors. *Appl Stochastic Models Bus Ind*. 2021;1-13. doi: 10.1002/asmb.2654