CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva and Carolina Marchant

A free open-access journal indexed by

INDEXED BY
EMERGING SOURCES CITATION
(ISI WEB OF SCIENCE)
CLARIVATE ANALYTICS

Web of Science Group

Scopus®

NON SOLUS
ELSEVIER

Volume 13 Number 1
April 2022
ISSN: 0718-7912 (print)
ISSN: 0718-7920 (online)

Published by the
Chilean Statistical Society
SOCHE
SOCIEDAD CHILENA DE ESTADÍSTICA

Aims

The Chilean Journal of Statistics (ChJS) is an official publication of the Chilean Statistical Society (`www.soche.cl`). The ChJS takes the place of *Revista de la Sociedad Chilena de Estadística*, which was published from 1984 to 2000.

The ChJS covers a broad range of topics in statistics, as well as in artificial intelligence, big data, data science, and machine learning, focused mainly on research articles. However, review, survey, and teaching papers, as well as material for statistical discussion, could be also published exceptionally. Each paper published in the ChJS must consider, in addition to its theoretical and/or methodological novelty, simulations for validating its novel theoretical and/or methodological proposal, as well as an illustration/application with real data.

The ChJS editorial board plans to publish one volume per year, with two issues in each volume. On some occasions, certain events or topics may be published in one or more special issues prepared by a guest editor.

Editors-in-Chief

Editors

Contents

Carolina Marchant and Víctor Leiva
Chilean Journal of Statistics: Thirty eight years generating quality knowledge 1

María Dueñas and Ramón Giraldo
Multivariate spatial prediction based on Andrews curves and functional geostatistics 3

Ernesto San Martín and Eduardo Alarcón-Bustamante
Dissecting Chilean surveys: The case of missing outcomes 17

Abdeldjalil Slama
A Bayesian detection of structural changes in autoregressive time series models 47

Christophe Chesneau, Muhammed Rasheed Irshad,
Damodaran Santhamani Shibu, Soman Latha Nitin, and Radhakumari Maya
On the Topp-Leone log-normal distribution: Properties, modeling, and applications in astronomical and cancer data 67

Lucas D. Ribeiro-Reis, Gauss M. Cordeiro, and José J. de Santana e Silvas
The Mc-Donald Chen distribution: A new bimodal distribution with properties and applications 91

Emilio Gómez-Déniz, Enrique Calderín-Ojeda, and José María Sarabia
The arctan family of distributions: New results with applications 113

Chilean Journal of Statistics
Volume 13, Number 1 (April 2022), Pages 3–16
DOI: 10.32372/ChJS.13-01-01

Spatial Statistics
Research Paper

Multivariate spatial prediction based on Andrews curves and functional geostatistics

María Dueñas[1] and Ramón Giraldo[1,*]

[1]Department of Statistics, Universidad Nacional de Colombia, Bogotá, Colombia

(Received: 03 March 2022 · Accepted in final form: 06 April 2022)

Abstract

There are two usual ways for modeling the realizations of multivariate random fields: Applying kriging individually on each variable or using cokriging, which considers the spatial cross-dependence between the variables. It has been shown that the second way, in general, allows a prediction variance reduction. The use of cokriging may be limited in practice when the number of variables increases because estimating the linear model of coregionalization (the cross-dependence between the variables) becomes complex. This work explores ordinary kriging for functional data based on Andrews curves as an alternative to the classical multivariate approach. Employing a simulation study, we compare the predictor proposed with kriging and cokriging. The methodology is applied to an environmental dataset.

Keywords: Andrews curves · Cokriging · Functional data · Geostatistics · Kriging

Mathematics Subject Classification: Primary 60G10 · Secondary 60G25.

1. Introduction

In many fields of applied science, it is required to simultaneously model data of several variables. Several statistical tools have been adapted to deal challenging multivariate problems. Among other areas, regression analysis (Bilodeau and Brenner, 1999), ANOVA (Smith et al., 1962), longitudinal data (Verbeke et al., 2014), and generalized linear models (Fahrmeir et al., 1994) have been tailored to this challenge. When the number of characteristics increases, the modeling becomes more complex. Also, the analysis of multivariate data is a big problem if there are inherent temporal and spatial dependence structures. One example is the multivariate spatial statistics (Gelfand et al., 2010), where it is necessary to consider auto and cross-correlations. The problem is solved using cokriging (assuming stationarity)(Giraldo et al., 2021). An advantage of this method is that it does not require that the variables are measured at the same sites. Its use has demonstrated to reduce uncertainty concerning ordinary kriging (spatial prediction of each variable separately). In its simplest form, cokriging assumes that the joint spatial correlation of the multivariate random field is generated from combinations of basic spatial covariance models and coregionalization matrices. If there are p variables, it is then necessary to estimate $p(p+1)/2$ variograms (including simple and cross-variograms). This makes this technique difficult to implement when p increases.

[*] Corresponding author. Email: rgiraldoh@unal.edu.co

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

Andrews curves (Andrews, 1972) are generally utilized in multivariate analysis to detect outliers (Embrechts et al., 1986), carry out clustering (Moustafa, 2011) and discriminant analysis. In this work, we propose its usage in multivariate geostatistics (Genton and Kleiber, 2015) as a tool for solving the high dimensionality problem. When the number of variables increases, it is not easy to estimate the coregionalization model and, therefore, to make predictions using cokriging. Employing Andrews curves combined with functional geostatistics (Giraldo et al., 2011) can simplify the problem because it only requires to fit a single variogram model. Once an Andrews curve is predicted on an unsampled site, implicitly all the variables of the multivariate random field of interest are predicted too.

Classical tools for spatial data analysis can be extended to functional data. Particularly in geostatistics, several alternatives for this purpose have been proposed. Ordinary, residual, and universal kriging for functional data (Mateu and Giraldo, 2022) are some approaches to solve the problem of spatial prediction when we have a realization of a functional random field (when a curve or, in general, a function is recorded at several sites of a region with spatial continuity). Here we propose an alternative for carrying out spatial prediction in multivariate geostatistics using ordinary kriging for functional data (Giraldo et al., 2011) based on Andrews curves. This alternative does not require to estimate a linear coregionalization model (Wackernagel, 2003), and consequently reducing the complexity of the problem.

The work is organized as follows. Section 2 gives a review on Andrews curves, multivariate geostatistics, and functional geostatistics. Section 3 presents the methodology proposed. An illustration with simulated data and an application to real data are shown in Section 4. The article ends with some conclusions, limitations and ideas for further research in Section 5.

2. Background

In this section, we present a short overview about Andrews curves (Andrews, 1972; Moustafa, 2011), multivariate geostatistics (Wackernagel, 2003), and ordinary kriging for functional data (Giraldo et al., 2011).

2.1 Andrews curves

A statistical multivariate analysis is considered when we have data of a p-dimensional random vector ($p > 1$). Given a realization of size n of a random vector $X = (X_1, \ldots, X_p)^\top$, we obtain the data matrix

$$\boldsymbol{x} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}. \tag{2.1}$$

The underlying idea of Andrews curves is that each multivariate data point (observation) can be represented by a curve using a Fourier interpolation function where the coefficients are the observation's components (Moustafa, 2011). Andrews curves are used as a descriptive tool for summarizing a multivariate data set as represented in Matrix given in expression (2.1) or employed to identify atypical values or clustering the individuals (Moustafa, 2011). These are built as linear combinations of the observations (Andrews, 1972). Specifically, for all i, for $i = 1, \ldots, n$, the i-th Andrews curve is given by

$$x_i(t) = \frac{1}{\sqrt{2\pi}} x_{i1} + \sin(t)x_{i2} + \cos(t)x_{i3} + \sin(2t)x_{i4} + \cdots, \tag{2.2}$$

with $t \in [-\pi, \pi]$. The order of the variables plays an important role in obtaining the curve: when there are many variables, the last ones have a low contribution to the shape of the curve. For this reason, they are usually ordered previously according to the amount of information that each of them provides. Generally, for this, a principal component analysis is initially carried out.

2.2 Multivariate geostatistics

This subsection is based on Giraldo et al. (2017). Let $\{\boldsymbol{X}(s) = (X_1(s), \ldots, X_m(s)) : s \in D\}$ be a multivariate spatial process defined over a domain $D \subset \mathbb{R}^2$. Assume $\boldsymbol{X}(s) = \boldsymbol{\mu}(s) + \boldsymbol{\epsilon}(s)$ is a stationary process with $\boldsymbol{\mu}(s)$ the mean vector and $\boldsymbol{\epsilon}(s)$ a stationary noise process with $\mathrm{E}(\boldsymbol{\epsilon}(s)) = \mathbf{0}$. We use the following notation: (i) $2\gamma_{lq}(s_i, s_j) = \mathrm{V}(X_l(s_i) - X_q(s_j))$, for $l, q = 1, \ldots, m$, $i, j = 1, \ldots, n$; (ii) $\boldsymbol{\gamma}_{lk}^\top = (\gamma_{lk}(s_1, s_0), \ldots, \gamma_{lk}(s_n, s_0))$; and (iii)

$$\boldsymbol{\Gamma}_{lq} = \begin{pmatrix} \gamma_{lq}(s_1, s_1) & \cdots & \gamma_{lq}(s_1, s_n) \\ \vdots & \ddots & \vdots \\ \gamma_{lq}(s_n, s_1) & \cdots & \gamma_{lq}(s_n, s_n) \end{pmatrix}.$$

The cokriging predictor of the random variable $X_k(s_0)$ based on the realization $\boldsymbol{X}(s_i)$, for $i = 1, \ldots, n$, is defined as

$$\widehat{X}_k(s_0) = \sum_{j=1}^{m} \lambda_{1j}^k X_j(s_1) + \cdots + \sum_{j=1}^{m} \lambda_{nj}^k X_j(s_n) = \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_{ij}^k X_j(s_i). \tag{2.3}$$

The predictor given in Equation (2.3) is unbiased if $\sum_{i=1}^n \lambda_{ik}^k = 1$ and $\sum_{i=1}^n \lambda_{ij}^k = 0$ for $j \neq k$, $j = 1, \ldots, m$. Using the Lagrange method to minimize the mean squared prediction error, $\mathrm{E}(\widehat{X}_k(s_0) - X_k(s_0))^2$, subject to the unbiasedness constraints gives the cokriging system of equations, which in matrix notation can be expressed by $\boldsymbol{C}\boldsymbol{\lambda}^k = \boldsymbol{c}^k$, with

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \cdots & \boldsymbol{\Gamma}_{1k} & \cdots & \boldsymbol{\Gamma}_{1m} & \mathbf{1} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots \\ \boldsymbol{\Gamma}_{k1} & & \boldsymbol{\Gamma}_{kk} & & \boldsymbol{\Gamma}_{km} & \mathbf{0} & & \mathbf{1} & & \mathbf{0} \\ \vdots & & & \ddots & \vdots & \vdots & & & \ddots & \vdots \\ \boldsymbol{\Gamma}_{m1} & \cdots & \boldsymbol{\Gamma}_{m2} & \cdots & \boldsymbol{\Gamma}_{mm} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{1} \\ \mathbf{1}^\top & \cdots & \mathbf{0}^\top & \cdots & \mathbf{0}^\top & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots & \vdots & \ddots & & & \vdots \\ \mathbf{0}^\top & & \mathbf{1}^\top & & \mathbf{0}^\top & 0 & & 0 & & 0 \\ \vdots & & \vdots & \ddots & \vdots & \vdots & & & \ddots & \vdots \\ \mathbf{0}^\top & \cdots & \mathbf{0}^\top & \cdots & \mathbf{1}^\top & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{Z} \\ \boldsymbol{Z}^\top & \mathbf{0}^* \end{pmatrix},$$

$$\boldsymbol{\lambda}^k = \begin{pmatrix} \boldsymbol{\lambda}_1^k \\ \vdots \\ \boldsymbol{\lambda}_k^k \\ \vdots \\ \boldsymbol{\lambda}_m^k \\ \delta_1 \\ \vdots \\ \delta_k \\ \vdots \\ \delta_m \end{pmatrix}, \boldsymbol{c}^k = \begin{pmatrix} \boldsymbol{\gamma}_1^k \\ \vdots \\ \boldsymbol{\gamma}_k^k \\ \vdots \\ \boldsymbol{\gamma}_m^k \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix},$$

where $(\mathbf{\Gamma}_{ij})_{(n\times n)}$, $\mathbf{1} = (1,\ldots,1)^\top_{(n\times 1)}$, $\mathbf{0} = (0,\ldots,0)^\top_{(n\times 1)}$, $(\mathbf{\Gamma})_{(m\times n)\times(n\times m)}$, $(\boldsymbol{Z})_{(n\times m)\times m}$, $(\mathbf{0}^*)_{(m\times m)}$, $\boldsymbol{\lambda}_j^k = (\lambda_{1j}^k,\ldots,\lambda_{nj}^k)^\top$, and $\boldsymbol{\gamma}_j^k = (\gamma_{1j}^k,\ldots,\gamma_{nj}^k)^\top$, for all $i,j = 1,\ldots,m$. Cokriging could be used for predicting simultaneously all m variables instead of predicting a variable, one at a time.

2.3 Functional geostatistics

Let $\{X_t(s), t \in \mathbb{R}, s \in D \subseteq \mathbb{R}^2\}$ be a second-order stationary and isotropic functional random field (Giraldo et al., 2011) whose realizations are functions defined in the real interval T with $X_t(s) \in L_2(T)$ the space of square integrable functions. From the stationarity conditions and taking $h = \|s_i - s_j\|$ we have

- $\mathrm{E}(X_t(s)) = \mu_t$.
- $\mathrm{V}(X_t(s)) = \sigma_t^2$.
- $\mathrm{C}(X_t(s_i), X_t(s_j)) = C(||s_i - s_j||; t) = C(h; t)$.
- $\frac{1}{2}\mathrm{V}(X_t(s_i) - X_t(s_j)) = \gamma(||s_i - s_j||; t) = \gamma(h; t)$.

The ordinary kriging predictor of the function on a site s_0 is defined as (Giraldo et al., 2011)

$$\widehat{X}_t(s_0) = \sum_{i=1}^{n} \lambda_i X_t(s_i),\ \lambda_1,\ldots,\lambda_n \in \mathbb{R}. \tag{2.4}$$

Optimal λ in Equation (2.4) that guarantee $\mathrm{E}(\widehat{X}_t(s_0)) = X_t(s_0)$ are obtained by solving the system

$$\begin{pmatrix} \int_T \gamma\left(||s_1 - s_1||, t\right) \mathrm{d}t & \cdots & \int_T \gamma\left(||s_1 - s_n||, t\right) \mathrm{d}t & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \int_T \gamma\left(||s_1 - s_n||, t\right) \mathrm{d}t & \cdots & \int_T \gamma\left(||s_n - s_n||, t\right) \mathrm{d}t & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \nu \end{pmatrix} = \begin{pmatrix} \int_T \gamma\left(||s_0 - s_1||, t\right) \mathrm{d}t \\ \vdots \\ \int_T \gamma\left(||s_0 - s_n||, t\right) \mathrm{d}t \\ 1 \end{pmatrix}. \tag{2.5}$$

The function $\gamma(h) = \int_T \gamma\left(h, t\right) \mathrm{d}t = (1/2)\mathrm{E}(\int_T (X_t(s_i) - X_t(s_j))^2 \mathrm{d}t)$ is called the trace-variogram. A review on its estimation based on the observed data is provided in Giraldo et al. (2011). Note that ν in Equation (2.5) is the Lagrange multiplier used to consider the unbiasedness constraint.

3. Multivariate geostatistics based on Andrews curves

We show how ordinary kriging based on Andrews curves is an alternative to perform multivariate spatial prediction. We assume isotropy and that all variables are recorded in the same sites.

3.1 From multivariate to functional Kriging

Let $\{\boldsymbol{X}(s), s \in D \subset \mathbb{R}^d\}$ be a p-dimensional random field and $[\boldsymbol{X}(s_1), \boldsymbol{X}(s_2), ..., \boldsymbol{X}(s_n)]^\top$ a sample of the process with

$$\boldsymbol{X}(s_i) = \begin{bmatrix} X_1(s_i) \\ X_2(s_i) \\ \vdots \\ X_p(s_i) \end{bmatrix}, \quad i = 1,\ldots,n.$$

Suppose we want to predict the random field at a site s_0. Employing Andrews curves given in Equation (2.2), the sample of the multivariate random field can be used to define a sample of a functional random field of Andrews curves $\{X_t(s), s \in D \subset \mathbb{R}^d, t \in [-\pi, \pi] \subset \mathbb{R}\}$ with the transformation

$$X_t(s_i) = \sum_{k=1}^{p} X_k(s_i)\phi_k(t), \tag{3.6}$$

with $\phi_k(t)$ the k-th coefficient of a Fourier series as defined in Equation (2.2). Likewise, from the multivariate observed sample of the random process $[\boldsymbol{x}(s_1), \boldsymbol{x}(s_2), ..., \boldsymbol{x}(s_n)]$, we have that

$$x_t(s_i) = \sum_{k=1}^{p} x_k(s_i)\phi_k(t). \tag{3.7}$$

Assuming that the curves defined in Equation (3.6) are a sample of a functional random field, we can use functional geostatistical methods (Giraldo et al., 2011, 2017) for carrying spatial prediction of all variables. Particularly, using ordinary kriging for functional data given in Equation (2.4) and taking as input the observed curves in Equation (3.7) we can predict the Andrews curves on unsampled sites. Note that the coefficients in Equation (2.2) are known and correspond to the data recorded from the p variables in the n sites $s_1, \ldots, s_n$.

3.2 Functional random field of Andrews curves

Assume the multivariate random field of interest is second order stationary. Consequently, we have the following properties for the random field of Andrews curves $\{X_t(s), s \in D \subset \mathbb{R}^d, t \in [-\pi, \pi] \subset \mathbb{R}\}$:

(i)

$$\begin{aligned} \mu(t) = \mathrm{E}\Big[X_t(s)\Big] &= \mathrm{E}\left[\sum_{k=1}^{p} X_k(s)\phi_k(t)\right] \\ &= \sum_{k=1}^{p} \mathrm{E}\left[X_k(s)\phi_k(t)\right] \\ &= \sum_{k=1}^{p} \phi_k(t)\mathrm{E}\left[X_k(s_i)\right] \\ &= \sum_{k=1}^{p} \phi_k(t)\mu_k, \end{aligned}$$

with $\mu_k = \mathrm{E}\left[X_k(s_i)\right]$ being the mean of the k-th random field.

(ii)

$$\begin{aligned} \mathrm{V}\left[X_t(s)\right] &= \sigma_t^2 \\ &= \mathrm{V}\left[\sum_{k=1}^{p} X_k(s)\phi_k(t)\right] \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p} \phi_k(t)\phi_l(t)\mathrm{Cov}\left[X_k(s), X_l(s)\right] \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p} \phi_k(t)\phi_l(t)C_{kl}(0), \end{aligned}$$

with $C_{kl}(0)$ being the covariance between the variables k and l.

(iii)

$$\begin{aligned} \mathrm{C}\Big[X_t(s_i), X_t(s_j)\Big] &= \mathrm{C}(h,t) \\ &= \mathrm{C}\left[\sum_{k=1}^{p} X_k(s_i)\phi_k(t), \sum_{k=1}^{k} X_k(s_j)\phi_k(t)\right] \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p} \phi_k(t)\phi_l(t)\mathrm{C}\left[X_k(s_i), X_l(s_j)\right] \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p} \phi_k(t)\phi_l(t)\mathrm{C}_{kl}(||s_i - s_j||) \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p} \phi_k(t)\phi_l(t)\mathrm{C}_{kl}(h). \end{aligned}$$

Note that the functional covariance depends only on the distance between sites s_i and s_j.

3.3 Spatial prediction of Andrews curves

Let $X_t(s_i)$, for $i = 1, \ldots, n$, be the sample of a functional random field of Andrews curves. Then the ordinary kriging predictor of an Andrews curve on a site s_0 is given by

$$\begin{aligned} \widehat{X}_t(s_0) &= \sum_{i=1}^{n} \lambda_i X_t(s_i) \\ &= \sum_{i=1}^{n} \lambda_i \sum_{k=1}^{p} X_k(s_i)\phi_k(t) \\ &= \sum_{k=1}^{p}\sum_{i=1}^{n} \lambda_i X_k(s_i)\phi_k(t). \end{aligned} \tag{3.8}$$

In Equation (3.8), each term $\sum_{i=1}^{n} \lambda_i X_k(s_i)$ is an scalar corresponding to the predictor $\widehat{X}_k(s_0)$. This is an unbiased and minimum variance predictor if $\lambda_1, \ldots, \lambda_n$ are such that

$$\int_T \mathrm{V}\left(\widehat{X}_t(s_0) - X_t(s_0)\right) \mathrm{d}t,$$

is minimum subject to $\sum_{i=1}^{n} \lambda_i = 1$.

3.4 Relationship between the trace-variogram function and univariate variograms

Note that

$$\begin{aligned}\int_t \Big(X_t(s_i) - X_t(s_j)\Big)^2 \mathrm{d}t &= \int_t \left(\sum_{k=1}^{p} X_k(s_i)\phi_k(t) - \sum_{k=1}^{p} X_k(s_j)\phi_k(t)\right)^2 \mathrm{d}t \\ &= \int_t \left[\sum_{k=1}^{p} \left(X_k(s_i) - X_k(s_j)\right)\phi_k(t)\right]^2 \mathrm{d}t,\end{aligned}$$

with $T = [-\pi, \pi]$. In matrix notation, we get

$$\begin{aligned}\int_t (X_t(s_i) - X_t(s_j))^2 \,\mathrm{d}t &= \int_t \Big[(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j))^\top \,\Phi(t)\Big]^2 \mathrm{d}t \\ &= \int_t (\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j))^\top \left(\Phi(t)\Phi(t)^\top\right)(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j))\,\mathrm{d}t \\ &= (\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j))^\top \int_t \left(\Phi(t)\Phi(t)^\top\right) dt\,(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)) \\ &= (\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j))^\top\, W\,(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j))\,,\end{aligned}$$

with W being the matrix of inner products of $\Phi(t)$. Taking into account that $\Phi(t)$ is an orthonormal basis, we have that $W = I_n$. Thus, we reach

$$\gamma(h) = \frac{1}{2}\mathrm{E}\left[\sum_{k=1}^{p} (X_k(s_i) - X_k(s_j))^2\right].$$

Under second order stationarity, we have that

$$\gamma_k(h) = \frac{1}{2}E\left[(X_k(s_i) - X_k(s_j))^2\right]. \tag{3.9}$$

From Equation (3.9), the trace-variogram can be expressed as

$$\begin{aligned}\gamma(h) &= \frac{1}{2}\mathrm{E}\left[\sum_{k=1}^{p} (X_k(s_i) - X_k(s_j))^2\right] \\ &= \frac{1}{2}\sum_{k=1}^{p} \mathrm{E}\left[(X_k(s_i) - X_k(s_j))^2\right] \\ &= \sum_{k=1}^{p} \gamma_k(h). \qquad (3.10)\end{aligned}$$

Therefore, the theoretical trace-variogram corresponds to the sum of the univariate semivariograms associated to the variables used to define the Andrews curves. This sum can be modeled with a single model once the empirical trace-variogram has been calculated. To carry out the spatial prediction we need to estimate the trace-variogram function $\int \gamma_t\left(||s_i - s_j||\right) dt$, for all $i = 1, \dots, n$. The corresponding estimator is given by

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right)^\top \boldsymbol{W} \left(\boldsymbol{X}(s_i) - \boldsymbol{X}(s_j)\right),$$

where $N(h)$ is the number of pairs (s_i, s_j) such that $h = ||s_i - s_j||$ and $|N(h)|$ is the number of sites separated by a distance h. Hence, the moment estimator of the trace-variogram function is stated as

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \sum_{k=1}^{p} \left(X_k(s_i) - X_k(s_j)\right)^2. \tag{3.11}$$

From Equation (3.10), the total prediction variance can be defined as

$$\sigma^2(s_0) = \sum_{i=1}^{n} \lambda_i \gamma(\|s_i - s_0\|) + \mu = \sum_{i=1}^{n} \lambda_i \sum_{k=1}^{p} \gamma_k(h) + \mu,$$

and its estimation is formulated by

$$\begin{aligned}\widehat{\sigma}^2(s_0) &= \sum_{i=1}^{n} \lambda_i \widehat{\gamma}(\|s_i - s_0\|) + \mu \\ &= \sum_{i=1}^{n} \lambda_i \left(\frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \sum_{k=1}^{p} \Big(x_k(s_i) - x_k(s_j)\Big)^2 \right) + \mu.\end{aligned}$$

4. Numerical Applications

This section initially compares kriging, cokriging and functional kriging using a small simulated dataset. Posteriorly, an application to a real dataset is presented. The computational routines were developed using the R software (R, 2022) version 4.1.3 for Windows platform.

4.1 Simulated Data

Suppose we have data of a stationary bivariate Gaussian random field $\{\boldsymbol{X}(s) = (X_1(s), X_2(s)) : s \in [0,1] \times [0,1]\}$ with means $\mu_1(s) = 2$ and $\mu_2(s) = 90$ and spatial dependence defined by the following variogram models:

$$\begin{aligned}\gamma_{X_1}(h) &= 0.30\gamma_1(h) + 0.26\gamma_2(h) \\ \gamma_{X_2}(h) &= 11\gamma_1(h) + 71\gamma_2(h) \\ \gamma_{X_1X_2}(h) &= 1.2\gamma_1(h) + 3.8\gamma_2(h),\end{aligned}$$

with $\gamma_1(h) = (1 - \exp((-h/0.7))$ and $\gamma_2(h) = (1.5(h/0.95) - 0.5(h/0.95)^2)$. In both models, the parameter ϕ is relatively high ($\phi = 0.7$ for the exponential model and $\phi = 0.95$ in the case of the spherical model), which is an indicator of high spatial simple and cross correlation. Note that ϕ is the parameter that defines the spatial correlation. The values assigned to this parameter correspond respectively to 70% and 95% of the maximum distance between sites of the simulation region.

Table 1. Four simulated data sets of a bivariate Gaussian random field defined on the square $[0,1]\times[0,1]$.

s	Coordinates	$X_1(s)$	$X_2(s)$
s_1	(1.00, 0.22)	1.51	80.83
s_2	(0.00, 0.33)	2.88	80.49
s_3	(0.67, 0.00)	2.94	102.29
s_4	(0.22, 0.78)	1.84	79.22

The corresponding covariance matrix is given by

$$\Sigma = \begin{bmatrix} 0.56 & 0.07 & 0.27 & 0.08 & 5.00 & 0.29 & 2.22 & 0.31 \\ 0.07 & 0.56 & 0.12 & 0.22 & 0.29 & 5.00 & 0.66 & 1.68 \\ 0.27 & 0.12 & 0.56 & 0.08 & 2.22 & 0.66 & 5.00 & 0.35 \\ 0.08 & 0.22 & 0.08 & 0.56 & 0.31 & 1.68 & 0.35 & 5.00 \\ 5.00 & 0.29 & 2.22 & 0.31 & 82.00 & 2.61 & 34.96 & 2.81 \\ 0.29 & 5.00 & 0.66 & 1.68 & 2.61 & 82.00 & 8.38 & 25.78 \\ 2.22 & 0.66 & 5.00 & 0.35 & 34.96 & 8.38 & 82.00 & 3.40 \\ 0.31 & 1.68 & 0.35 & 5.00 & 2.81 & 25.78 & 3.40 & 82.00 \end{bmatrix}.$$

Assume that we want to predict the variables $X_1(s_0)$ and $X_2(s_0)$, $s_0 = (0.22, 0.00)$, using four observations of the process; see Table 1. Based on the covariance matrix and employing univariate ordinary Kriging, ordinary Cokriging, and Functional Kriging predictions are obtained for $X_1(s_0)$ and $X_2(s_0)$.

Table 2. Predictions using the three methods. $\widehat{\sigma}_T^2$ correspond to the total prediction variance (sum of the prediction variances).

Method	$\widehat{X}_1(s_0)$	$\widehat{X}_2(s_0)$	$\widehat{\sigma}_T^2$
Kriging	2.199	93.708	68.269
Cokriging	2.707	93.602	68.163
Functional kriging	2.819	93.789	68.278

Table 2 shows that we obtain reasonable predictions with the three methods (values around the means $\mu_1(s)$ and $\mu_2(s)$ of the processes) with variances of the predictions that only differ slightly. A more intensive simulation study was conducted posteriorly. Considering the same spatial dependence structure defined above by $\gamma_{X_1}(h)$, $\gamma_{X_2}(h)$, and $\gamma_{X_1X_2}(h)$, a realization of size 100 of the bivariate process was generated. A cross-validation analysis was carried out with these data, that is, each simulated datum was partially deleted and predicted based on the remaining 99 observations through the three methods (Kriging, Cokriging, and Functional Kriging). We do not present the results in detail. The means of the prediction errors were in all cases (three methods) close to zero and the means prediction variances were also very similar (around 11.71). A detailed review of the results can be seen in Dueñas (2017). Here, we consider only two processes to show that even when the number of variables is small the methodology based on functional kriging can be applied. If the number of processes increases, there is more significant differences between the methods, but cokriging also is more complex. In these cases, the approach based on functional kriging may be more appropriate.

4.2 Real Data

The lagoon-estuarine system Ciénaga Grande de Santa Marta (CGSM) located at the north coast of Colombia (Figure 1) is of interest for its ecological and hydrological characteristics and its richness in fish, mollusks, and crustaceans (Rodríguez-Rodríguez et al., 2021). Monitoring its physicochemical and biological conditions is essential due to its environmental and economic impact on the region.

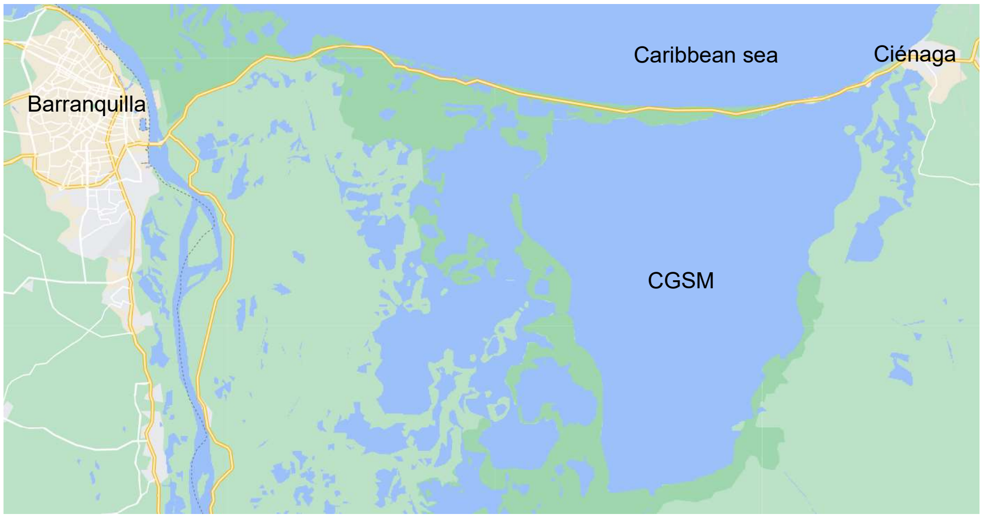

Figure 1. The lagoon-estuarine ecosystem Ciénaga Grande de Santa Marta (CGSM) is located at the north coast of Colombia between the cities of Barranquilla and Ciénaga. A narrow, continuous sandbar borders the entire CGSM complex to the north. Source: Google Maps 2021.

This work shows how to use functional kriging based on Andrews curves to jointly predict the spatial distribution of some of these variables. Specifically, we analyze data of six variables (salinity, dissolved oxygen (mg O_2/L), temperature (°C), chlorophyll-a (μg/l), total suspended solids (mg/l), and depth (cm)) collected in 95 sampling sites of the system. The spatial distribution of these variables according to the quartiles of the recorded values is shown in Figure 2. These plots suggest that is reasonable assuming stationarity, because there is not a defined spatial trend in any case. There are three alternatives for doing prediction in this case. We can apply ordinary kriging (without considering the dependency between the variables), ordinary cokriging which require the estimation of a LMC (a complex procedure in this scenario because we must to take into account data of six random processes simultaneously), or ordinary kriging based on Andrews curves. Below, we show the results considering this last option. We also do a comparison with the results obtained using ordinary kriging.

In Table 3, we report the variation coefficients calculated with the 95 observations from each one of the six variables considered in the study. These values are ordered from highest to lowest. Following Andrews (1972), we employ this order to define the coefficients x_{ij} from Equation (2.2) of the Andrews curves for the dataset of interest (top panel of Figure 3). We note that the curves have a similar behavior. Only two curves have a different pattern (see curves with the lowest values for $t \in (0, 1.7)$). These correspond to places in the north of the Ciénaga that have different conditions of salinity and depth.

Using Equation (3.11), we calculate the empirical trace-variogram function (white circles in bottom panel of Figure 3). An exponential semivariogram model with $\widehat{\phi} = 6460$m y $\widehat{\sigma}^2 = 19224.08$ was fitted to this scatterplot (red curve in bottom panel of Figure 3). The value of $\widehat{\phi}$ indicates that the Andrews curves are correlated up to a distance of about 6.4 km. Using this model, we can estimate the weights λ_i, for $i = 1, \ldots, 95$ in Equation (3.8) and predict the six variables on unsampled sites of the region utilizing functional kriging based on Andres curves. To evaluate the performance of the predictor we do a cross-validation analysis comparing the results with the ones obtained with ordinary kriging. In Table 4 we show the sum of squares of prediction errors for each one of the six variables based on

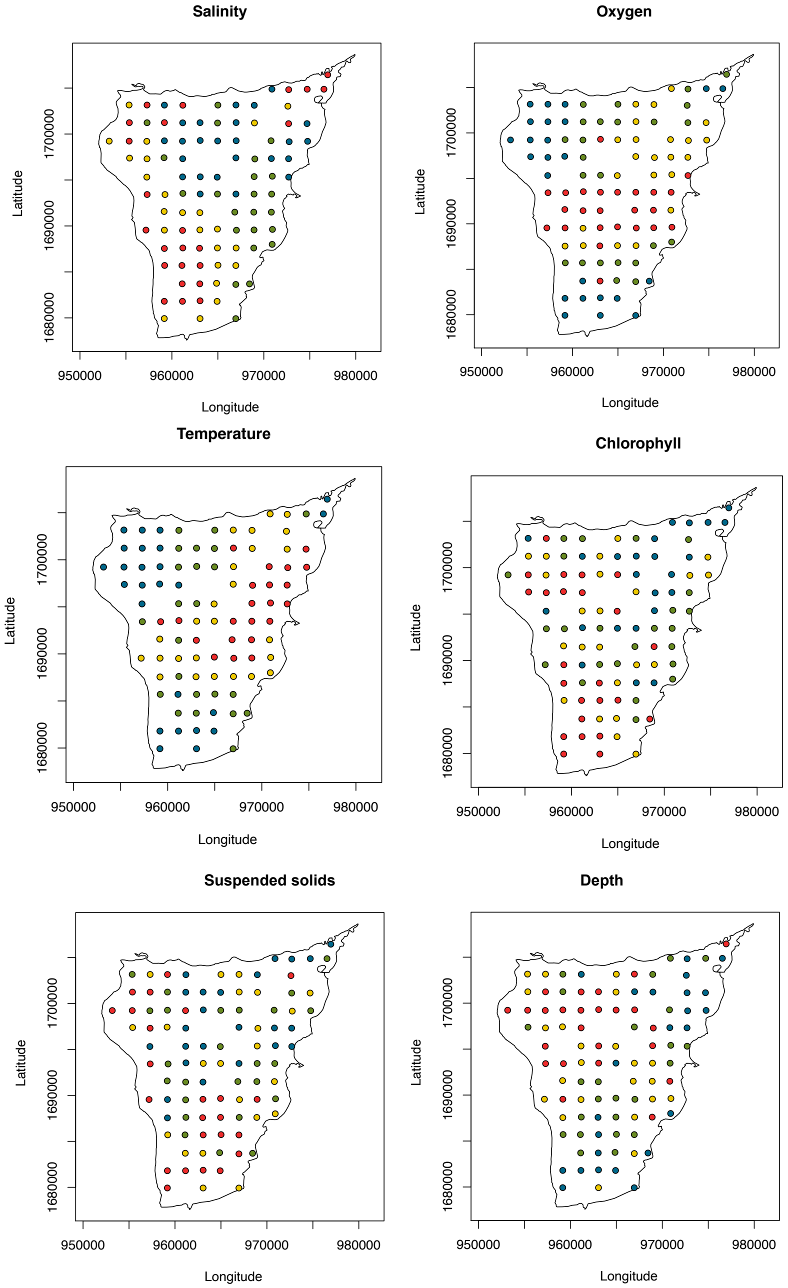

Figure 2. Spatial distribution of data for each one of the six variables considered. The values are, in each case, divided according to the quartiles.

Table 3. Coefficients of variation calculated with data recorded in 95 sites of the lagoon-estuarine system Ciénaga Grande de Santa Marta.

Variable	Coefficient of variation (%)
Oxygen	36.5
Depth	24.5
Chlorophyll	23.8
Suspended solids	19.3
Salinity	17.1
Temperature	7.2

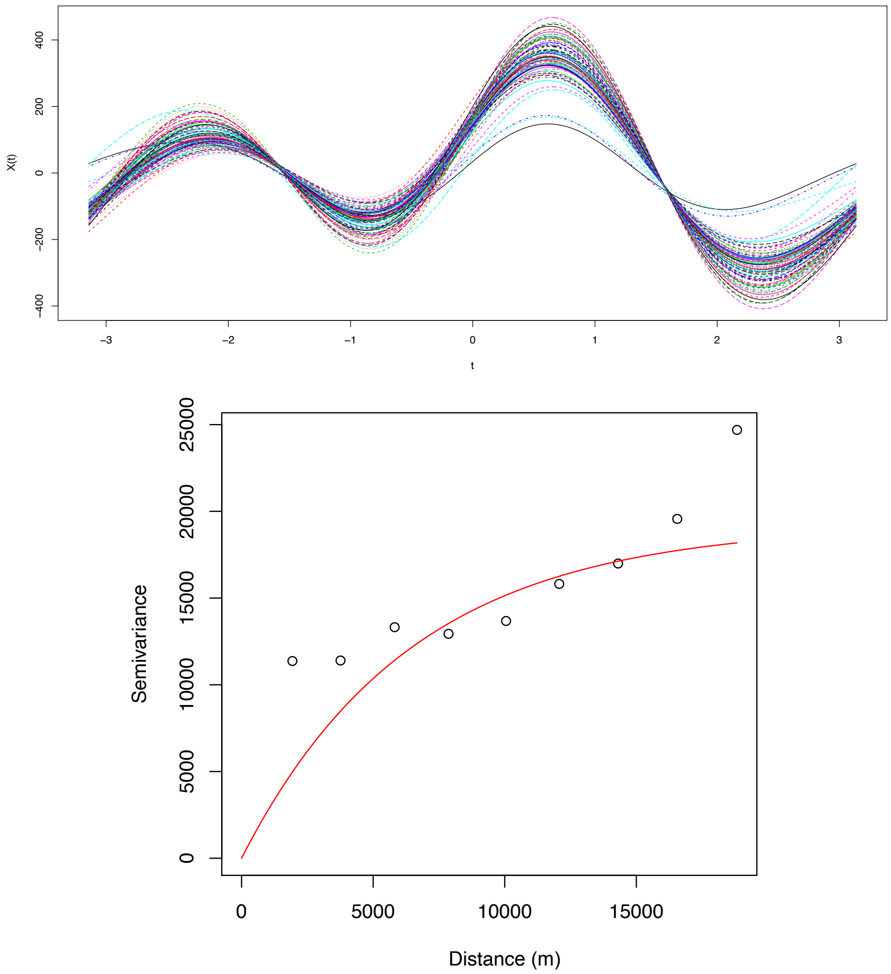

Figure 3. Andrews curve calculated for each one of 95 sites of the lagoon-estuarine system Ciénaga Grande de Santa Marta, based on the values of six physicochemical variables (top); and variogram model (red line) fitted to the empirical trace-variogram function (bottom).

the two approaches considered. In general the results look similar. To test for significant differences we use Wilcoxon tests based on the cross-validation residuals. These indicate that the method based on functional kriging using Andrews curves is better than ordinary kriging in the case of the variables depth and suspended solids. In the other cases there are not significant differences between the two strategies.

Table 4. Sum of squares errors of cross-validation (using the data of 95 sites) obtained by functional kriging based on Andrews curves and ordinary univariate kriging.

	Functional kriging	Univariate kriging
Oxygen	218.8	216.9
Depth	12.3	12.5
Chlorophyll	46335.9	46457.2
Suspended solids	169397.1	170136.2
Salinity	223.2	229.9
Temperature	46.9	46.4

5. Conclusions, limitations, and future research

In the paper, we have proposed the predictor ordinary kriging for functional data (Giraldo et al., 2011) based on Andrews curves (Andrews, 1972) as a method for making spatial prediction in multivariate geostatistics (Smith et al., 1962). The results based on a simulation study and an analysis of real-world data have indicated that this strategy has a good performance. Obviously, if the geostatistical analysis is carried out with two or three variables, it is more convenient to use cokriging, since the prediction variance is reduced. However, when the number of variables increases, this option is limited and the proposed technique emerges as a very appropriate alternative, because it only requires the estimation of just one variogram and does not have the limitations of the linear coregionalization model.

The proposed methodology could be adapted to the case of optimal sampling (Bohorquez et al., 2016), regression, and analysis of variance of multivariate spatial data. Other research alternatives are the extension to the case of non-stationary processes and the treatment of outliers (Borssoi et al., 2011).

Author contributions Conceptualization, M.D., R.G.; methodology, M.D., R.G.; software, M.M.; investigation, M.D., R.G; writing original draft preparation, M.D., R.G., writing review and editing, R.G. Both authors have read and agreed to the published version of the paper.

Acknowledgements The authors would like to thank the Editors and the reviewers for their valuable comments and suggestions which improved the quality of this paper.

Funding Not applicable.

Conflicts of interest The authors declare no conflict of interest.

References

Andrews, D., 1972. Plots of high-dimensional data. Biometrics, 28, 125–136.

Bilodeau, M. and Brenner, D., 1999. Multivariate Regression. Springer, New York, USA.

Bohorquez, M., Giraldo, R., and Mateu, J., 2016. Optimal sampling for spatial prediction of functional data. Statistical Methods Applications, 25, 39–54.

Borssoi, J.A., De Bastiani, F., Uribe-Opazo, M.A., and Galea, M., 2011. Local influence of explanatory variables in Gaussian spatial linear models. Chilean Journal of Statistics, 2(2), 29–38.

Dueñas, M., 2017. Análisis geoestadístico multivariado a través de métodos funcionales y curvas de Andrews. Master thesis. Department of Statistics, Universidad Nacional de Colombia, Colombia.

Embrechts, P., Herzberg, A., and Allen, C., 1986. An investigation of Andrews's plots to detect period and outliers in time series data. Communications in Statistics: Simulation and Computation, 15, 1027–1051.

Fahrmeir, L., Tutz, G., Hennevogl, W. and Salem, E., 1994. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, New York, USA.

Gelfand, A., Diggle, P. Guttorp, P., and Fuentes, M., 2010. Handbook of Spatial Statistics. CRC Press, New York, USA.

Genton, M. and Kleiber, W., 2015. Cross-covariance functions for multivariate geostatistics. Statistical Science, 30, 147–163.

Giraldo, R., Delicado, P., and Mateu, J., 2011. Ordinary kriging for function-valued spatial data. Environmental Ecological Statistics, 18, 411–426.

Giraldo, R., Delicado, P., and Mateu, J., 2017. Spatial prediction of a scalar variable based on data of a functional random field. Comunicaciones en Estadística, 10, 315–344.

Giraldo, R., Herrera, L., and Leiva, V., 2020. Cokriging prediction using as secondary variable a functional random field with application in environmental pollution. Mathematics, 8, 1305.

Mateu, J. and Giraldo, R., 2021. Geostatistical Functional Data Analysis. Wiley, New York, USA.

Mustafa, R., 2011. Andrews curves. Computational Statistics, 3, 373–382.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rodríguez-Rodríguez, J.A., Mancera-Pineda, J.E., and Tavera, H., 2021. Mangrove restoration in Colombia: Trends and lessons learned. Forest Ecology and Management, 496, 119414.

Smith, H., Gnanadesikan, R., and Hughes, J., 1962. Multivariate analysis of variance (MANOVA). Biometrics, 18, 22–41.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M., 2014. The analysis of multivariate longitudinal data: a review. Statistical Methods in Medical Research, 23, 42–59.

Wackernagel, H., 2003. Multivariate Geostatistics: An Introduction with Applications. Springer, New York, USA.

Information for authors

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF electronic version of the manuscript in a free format to the Editors-in-Chief of the ChJS (E-mail: chilean.journal.of.statistics@gmail.com). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a“cover letter” presenting their manuscript and mentioning: “We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere”.

Preparation of accepted manuscripts

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at http://soche.cl/chjs/files/ChJS.zip. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at http://www.ams.org/mathscinet/msc/. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:

Arellano-Valle, R., 1994. Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., 1997. Local influence. In Kotz, S., Read, C.B., and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 1., Wiley, New York, pp. 380-385.

Rukhin, A.L., 2009. Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters, 79, 1004-1007.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Peña, D., and Pankratz, A.E., 2000. Outliers in multivariate time series. Biometrika, 87, 789-804.

References in the text must be given by the author’s name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

Copyright