Ronny Vallejos
Felipe Osorio
Moreno Bevilacqua

Spatial Relationships Between Two Georeferenced Variables

With Applications in R

Springer

Spatial Relationships Between Two Georeferenced Variables

Ronny Vallejos · Felipe Osorio ·
Moreno Bevilacqua

Spatial Relationships Between Two Georeferenced Variables

With Applications in R

Springer

Ronny Vallejos
Department of Mathematics
Federico Santa María Technical University
Valparaíso, Chile

Felipe Osorio
Department of Mathematics
Federico Santa María Technical University
Valparaíso, Chile

Moreno Bevilacqua
Faculty of Engineering and Sciences
Universidad Adolfo Ibañez
Viña del Mar, Chile

ISBN 978-3-030-56680-7 ISBN 978-3-030-56681-4 (eBook)
https://doi.org/10.1007/978-3-030-56681-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my lovely wife and son,
Carmen and Ronny Javier.

Ronny Vallejos

To my children Vicente and Florencia,
for their love.

Felipe Osorio

To Gianni and Agostina.

Moreno Bevilacqua

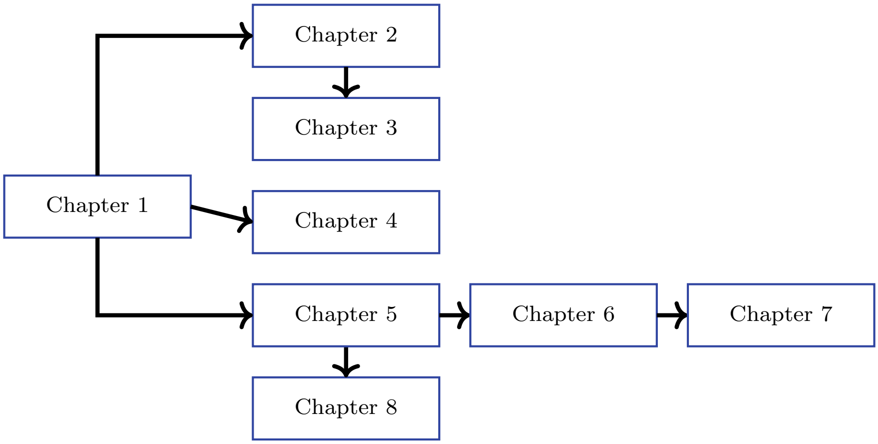
Preface

In this book we cover a wide range of topics that currently are available only as a material included in many research papers. The material we cover is related to 35 years of research in spatial statistics and image processing. Our approach includes an exposition of the techniques keeping the mathematical and statistical background at a minimum so that the technical aspects are placed in an appendix in order to facilitate the readability. Each chapter contains a section with applications and R computations where real datasets in different contexts (Fisheries Research, Forest Sciences, and Agricultural Sciences) are analyzed.

We trust that the book will be of interest to those who are familiar with spatial statistics and to scientific researchers whose work involves the analysis of geostatistical data. For the first group, we recommend a fast reading of Chap. 1 and then the chapters of interest. For the second group, the preliminaries given in Chap. 1 are recommended as a prerequisite, especially because of the language and notation used further in the book. The interdependence of the chapters is depicted below, where arrow lines indicate prerequisites.

Extensive effort was invested in the composition of the reference list for each chapter, which should guide readers to a wealth of available materials. Although our reference lists are extensive, many important papers that do not fit our presentation have been omitted. Other omissions and discrepancies are inevitable. We apologize for their occurrence.

Many colleagues, students, and friends have been of great help to our work in this book in several ways: by having discussions that improve our understanding of specific subjects; by doing research with us in a number of collaborative projects; by providing constructive criticism on earlier versions of the manuscript; and by supporting us with enthusiasm to finish this project. In particular, we would like to thank Aaron Ellison, Daniel Griffith, Andrew Rukhin, Wilfredo Palma, Manuel Galea, Emilio Porcu, Pedro Gajardo, Jonathan Acosta, Silvia Ojeda, Javier Pérez, Francisco Alfaro, Rogelio Arancibia, Carlos Schwarzenberg, Angelo Gárate, and Macarena O’Ryan.



Ronny Vallejos' research was supported by Fondecyt grant 1120048, by the Advanced Center for Electrical and Electronics Engineering, FB-0008 and by internal grants at Universidad Técnica Federico Santa María, Chile. Felipe Osorio's work was supported by Fondecyt grant 1140580. Moreno Bevilacqua's research was partially supported by Fondecyt grant 1160280, and by Millennium Science Initiative of the Ministry of Economy, Development, and Tourism, grant Millenium Nucleus Center for the Discovery of Structures in Complex Data, Chile.

During the period in which the book was written Ronny Vallejos and Felipe Osorio were in the Departamento de Matemática at Universidad Técnica Federico Santa María, Valparaíso, Chile. Moreno Bevilacqua was in the Departamento de Estadística at Universidad de Valparaíso, Chile.

We wish to end this preface by thanking our families, friends, and others who helped make us what we are today and thereby contributed to this book. In particular, we would like to thank Eva Hiripi of Springer for her constant support.

Valparaíso, Chile
May 2020

Ronny Vallejos
Felipe Osorio
Moreno Bevilacqua

Contents

Chapter 1
Introduction

1.1 Motivating Examples

The types of spatial data we describe in the following examples have been widely discussed in Cressie (1993) and Schabenberger and Gotway (2005) in the context of a single realization of a stochastic sequence. Since the book addresses spatial association between two stochastic sequences, we consider some basic assumptions that do not vary throughout the book. For example, we denote the two random sequences as $X(\boldsymbol{s})$ and $Y(\boldsymbol{s})$ for $\boldsymbol{s} \in D \subset \mathbb{R}^2$, and the available information is the observations $X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_n)$ and $Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)$. That is, both variables have been measured at the same locations in space.

1.1.1 The Pinus Radiata Dataset

Pinus radiata, which is one of the most widely planted species in Chile, is planted in a wide array of soil types and regional climates. Two important measures of plantation development are dominant tree height and basal area. Research shows that these measures are correlated with the regional climate and local growing conditions (see Snowdon 2001). The study site is located in the *Escuadrón* sector, south of Concepción, in the southern portion of Chile (36° 54′ S, 73° 54′ O) and has an area of 1244.43 hectares. In addition to mature stands, there is also interest in areas that contain young (i.e., four years old) stands of Pinus radiata. These areas have an average density of 1600 trees per hectare. The basal area and dominant tree height in the year of the plantation's establishment (1993, 1994, 1995, and 1996) were used to represent stand attributes. These three variables were obtained from 200 m^2 circular sample plots and point-plant sample plots. For the latter, four quadrants were established around the sample point, and the four closest trees in each quadrant (16 trees in total) were then selected and measured. The samples were located systematically using a mean distance of 150 m between samples. The total number of plots available

R. Vallejos et al., *Spatial Relationships Between Two Georeferenced Variables*,
https://doi.org/10.1007/978-3-030-56681-4_1

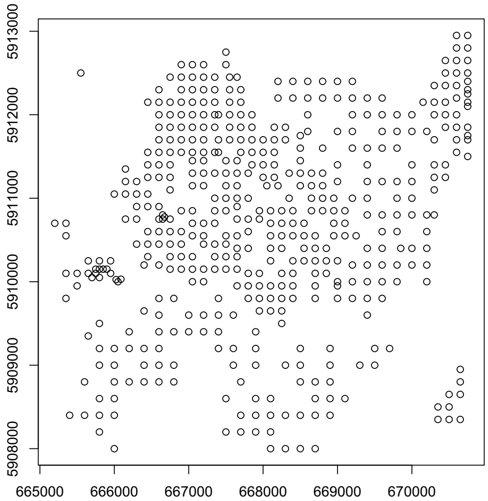


Fig. 1.1 Locations where the samples were taken

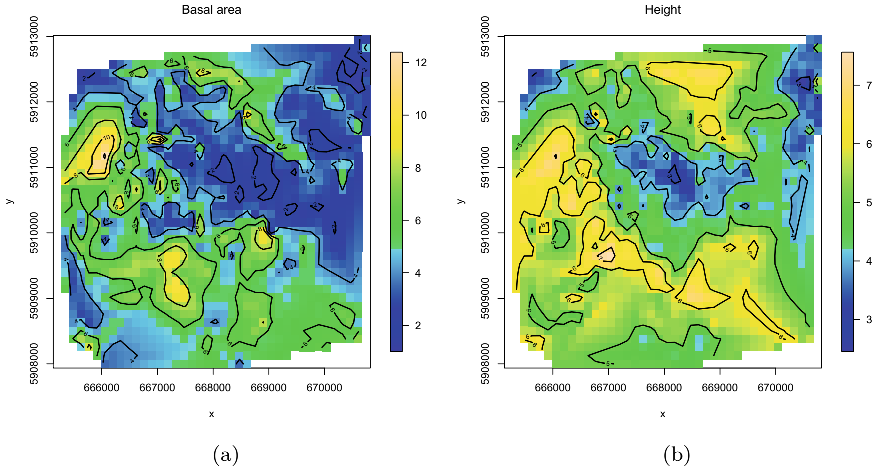


Fig. 1.2 **a** Bilinear interpolation of the three basal areas; **b** Bilinear interpolation of the three heights

for this study was 468 (Fig. 1.1). Figure 1.2 shows a simple bilinear interpolation and the corresponding contours for the two variables. The original georeferenced data do not enable estimation of the sample correlation coefficient because it is challenging to train the human eye to capture two-dimensional patterns.

The objective of analyzing these data is to construct a suitable measure that takes into account the spatial association between the two variables. One could be tempted to compute the Pearson correlation coefficient for the two sequences by considering these variables to be two simple columns. Then, the construction of a scatterplot could

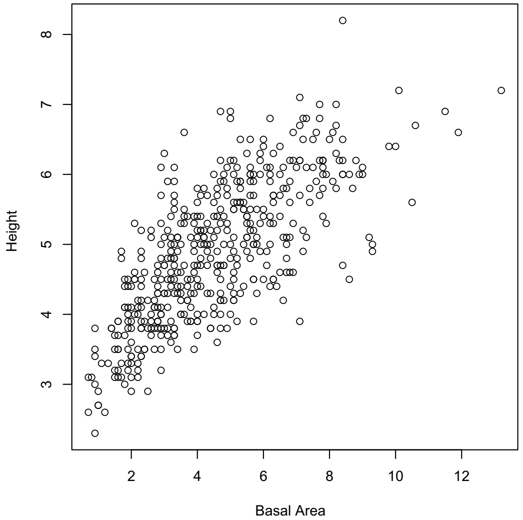


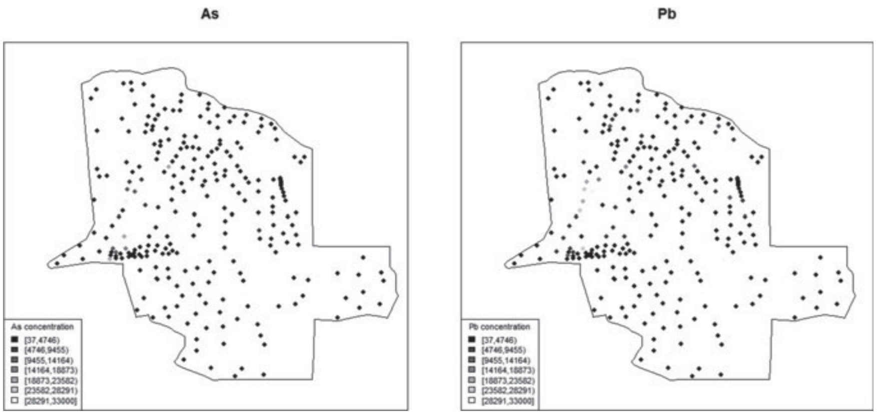
Fig. 1.3 Height versus Basal area (468 observations)

help to determine whether there is a linear trend between the basal area and height. It is interesting to emphasize that the human eye can usually be trained to estimate the value of the correlation coefficient from the information provided by a scatterplot between the variables of interest. However, when the data have been georeferenced on two-dimensional space, it is difficult to estimate a reasonable association between the variables. For the forest variables, a scatterplot between the basal area and height is displayed in Fig. 1.3, which shows a clear linear correlation between the basal area and height. The correlation coefficient confirms the linear pattern (0.7021).

Although the exploratory data analysis provides good initial insight into the real problem, the issue of how to take into account the possible spatial association for each variable has not yet been addressed. Thus, a primary objective in analyzing these data is to develop coefficients for the spatial association between two georeferenced variables that take into account the existing spatial association within and between the variables.

1.1.2 The Murray Smelter Site Dataset

The dataset consists of soil samples collected in and around the vacant, industrially contaminated, Murray smelter site (Utah, USA). This area was polluted by airborne emissions and the disposal of waste slag from the smelting process. A total of 253 locations were included in the study, and soil samples were taken from each location. Each georeferenced sample point is a pool composite of four closely adjacent soil samples in which the concentration of the heavy metals arsenic (As) and lead (Pb) was determined. A detailed description of this dataset can be found in Griffith (2003) and



(a) (b)

Fig. 1.4 Locations of 253 geocoded aggregated surface soil samples collected in a 0.5 square mile area in Murray, Utah and their measured concentrations of As and Pb. Of these 173 were collected in a facility Superfund site, and 80 were collected in two of its adjacent residential neighborhoods located along the western and southern borders of the smelter site. **a** As measurements; **b** Pb measurements

Griffith and Paelinck (2011). For each location, the As and Pb attributes are shown in Fig. 1.4a, b.

The objective for this data is to assess the spatial association between As and Pb. Figure 1.4 shows that, in this case, the observations are clearly located in a nonrectangular grid in two-dimensional space. Again, the goal can be achieved by quantifying the coefficients of the spatial association or by constructing a suitable hypothesis test for the Pearson correlation coefficient ρ between As and Pb.

A hypothesis test of the form

$$H_0 : \rho = 0 \quad \text{against} \quad H_1 : \rho \neq 0$$

can be stated under the assumption of normality for both variables (As and Pb). Then, the test statistic is

$$t = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} = 11.5548, \tag{1.1}$$

where $n = 253$ and $r = 0.5892$. The p-value associated with the test is 2.2×10^{-16}; thus, there is sufficient evidence to reject H_0 for a significance $\alpha > p$.

In the previous analysis, we assumed that the correlation between the variables is constant, i.e., $\text{cor}[X(\boldsymbol{s}), Y(\boldsymbol{s})] = \rho$, for all $\boldsymbol{s} \in D$. However, as we will see in the following chapters, this dataset and several others do not support this restriction. Instead, they exhibit a clear spatial association between the variables of interest.

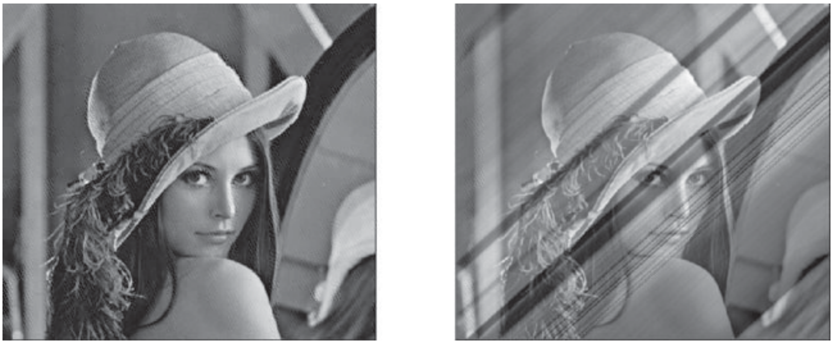
The objective in analyzing these data is to develop suitable hypothesis testing methods to assess the significance of the spatial association between two spatial variables by considering the existing spatial association between them.

1.1.3 Similarity Between Images

With the rapid proliferation of digital imaging, image similarity assessment has become a fundamental issue in many applications in various fields of knowledge (Martens and Meesters 1998). Many proposals of indices that capture the similarity or dissimilarity between two digital images have received attention during the past decade. One important feature to consider is the capability of some coefficients to provide a better interpretation of human visual perception than is provided by the widely used mean square error (MSE) (Wang et al. 2004).

Here, we introduce an example that uses real data to illustrate the dependence of the spatial association on a particular direction in space, noting that the correlation coefficient (a crude measure of spatial association between two processes) cannot account for the directional association between two images. To accomplish this goal, an original image (Lenna) of size 512 × 512 was taken from the USC-SIPI image database http://sipi.usc.edu/database/ (See Fig. 1.5a). The image shown in Fig. 1.5a was processed by Algorithm 4.1 in Vallejos et al. (2015) to transform the original image into an image with a clear pattern in the direction $\boldsymbol{h} = (1, 1)$. The processed image is displayed in Fig. 1.5b.

The correlation coefficient between the images shown in Fig. 1.5 is $r = 0.6909$. Clearly, the correlation coefficient does not capture the evident pattern observed by the human eye between the original and transformed images. In fact, the trend in the off-diagonal of the image in Fig. 1.5b is sufficient to decrease the correlation



(a) (b)

Fig. 1.5 **a** Original image (Lenna); **b** Image transformed into the direction $\boldsymbol{h} = (1, 1)$

coefficient to 0.6909 even though the features of the original image are still present and detectable by the human eye.

The objective in analyzing these data is to construct image similarity coefficients that can detect patterns in different directions in space and to appropriately represent the human visual system.

1.2 Objective of the Book

The aim of this book is to gather the published material that is spread throughout the literature. The book may be of interest to two types of users. First, researchers from applied areas, such as agriculture, soil sciences, forest sciences, environmental sciences, and engineering. For these and other users who possibly are more interested in the applications, the book is organized in such a way that the mathematical foundations in each chapter can be skipped. Second, for investigators who are interested in the development of new techniques and methods to assess the significance of the correlation between two or more spatial processes that are well defined on a two-dimensional plane, at the end of the book, we include an appendix with the proofs of the results presented in the book and some mathematical details that support the expressions and equations that are briefly explained in the main text. Although the book contains methods that were discovered and proposed approximately thirty years ago and are very well known to readers working in spatial statistics and geostatistics, other methods in this book have recently been developed and are not yet available in a publication like this.

1.3 Layout of the Book

This book is divided into three parts. The first part considers the association between two random fields from an hypothesis testing perspective (Chaps. 2 and 3). The second part is devoted to point estimation coefficients of association. These perspectives are developed in Chaps. 4–7. The third part considers the spatial association between two images (Chap. 8). Several applications are presented throughout the book. Each chapter ends with a set of theoretical and applied exercises. Most of the applied problems are related to real datasets, and it is expected that the reader will use R software to solve them.

1.4 Computation

To illustrate the applicability of the methods exposed in this book, each chapter contains a section on R computations with practical applications. In most of the examples, we show how R software and the contributed packages SpatialPack and

GeoModels can be used to implement the techniques described in the corresponding chapter. The SpatialPack package is freely available from the R website www.r-project.org and the package GeoModels can be downloaded from the website https://vmoprojs.github.io/GeoModels-page/.

1.5 Preliminaries and Notation

In this section, we provide the necessary material that will be used in subsequent chapters. Readers interested in practical applications with real datasets can skip the rest of this chapter and move ahead.

1.5.1 Spatial Processes

In this section, we introduce the basic notion of stochastic processes. Our goal is to define the mean, variance, and covariance functions of spatial processes. These concepts will be used in the subsequent chapters.

A stochastic process is a family or collection of random variables in a probability space. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, and let D be an arbitrary index set. A stochastic process is a function

$$X : (\Omega, \mathcal{F}, \mathcal{P}) \times D \longrightarrow \mathbb{R},$$

such that for all $\boldsymbol{s} \in D$, $X(w, \boldsymbol{s})$ is a random variable. In the sequel, we will denote a stochastic process as $X(\boldsymbol{s})$, for $\boldsymbol{s} \in D$, or $\{X(\boldsymbol{s}) : \boldsymbol{s} \in D\}$.

The above definition enables us to define a variety of processes. For example, if $D = \mathbb{Z}$, $X(s)$ is a discrete time series. Similarly, a spatial process is a collection of random variables that are indexed by the set $D \subset \mathbb{R}^d$. In the time series case, the realizations of the process are observations indexed by time, while in the spatial case, the realizations of the process are regions on the subspace $\mathbb{R}^d$. Additionally, in the first case, the index set is a totally ordered set; however, in the spatial case, it is possible to define a partially ordered set. We denote the coordinates of a spatial process defined on a space of dimension d as $\boldsymbol{s} = (s_1, \ldots, s_d)^\top$.

As an example, consider the process $\{X(s) : s \in \mathbb{Z}\}$ defined by the equation

$$X(s) = A\cos(\eta s + \phi), \tag{1.2}$$

where A is a random variable independent of $\phi \sim \mathcal{U}(0, 2\pi)$, and η is a fixed constant. For the particular case when $A \sim \mathcal{N}(0, 1)$ and $\eta = 1$, 1000 observations from process (1.2) were generated. This realization of process $\{X(s)\}$ is displayed in Fig. 1.6.

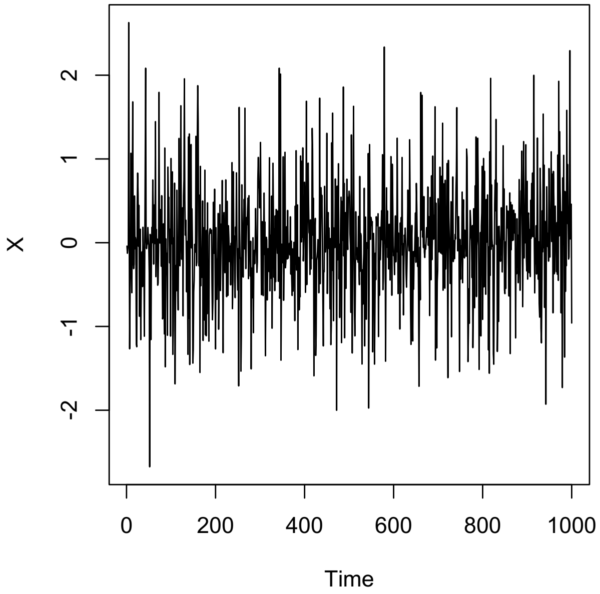


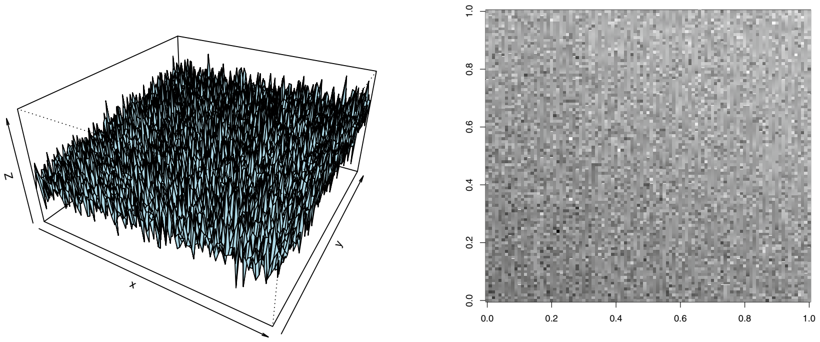
Fig. 1.6 A realization from process (1.2)

As another example, consider the process $\{Z(x, y) : (x, y) \in \mathbb{Z}^2\}$ defined as follows

$$Z(x, y) = \beta_1 x + \beta_2 y + \epsilon(x, y), \tag{1.3}$$

where $\{\epsilon(x, y); (x, y) \in \mathbb{Z}^2\}$ is a collection of independent and identically distributed random variables with zero-mean and variance σ^2. For $\beta_1 = 1, \beta_2 = 1, \epsilon(x, y) \sim \mathcal{N}(0, 1)$, and for a grid of size 512 × 512, a realization of this process is shown in Fig. 1.7.

The class of all random functions or stochastic processes is too large to enable methods that are suitable for all types of processes to be designed. In fact, most of the developments have been proposed for special cases. One important class of processes are characterized by the feature that their distributions do not change over time/space. We can summarize this property by saying that for any set of locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, the joint probability of $\{X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_n)\}$ must remain the same if we shift each location by the same amount. In other words, the distribution of the process does not change if the spatial distribution is invariant under translation of the coordinates. A process $\{X(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ with this property is called strictly stationary. Indeed, a spatial process $\{X(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is said to be strictly stationary if for any set of locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$ and for any $\boldsymbol{h} \in D$, the joint distribution of $\{X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_n)\}$ is identical to that of $\{X(\boldsymbol{s}_1 + \boldsymbol{h}), \ldots, X(\boldsymbol{s}_n + \boldsymbol{h})\}$. Strict stationarity is a severe requirement and can be relaxed by introducing a milder version that imposes conditions on only the first two moments of the process. The second-order condition, $\mathbb{E}[X^2(\boldsymbol{s})] < \infty$, for all



(a) (b)

Fig. 1.7 **a** A realization from (1.3). **b** Image associated with (**a**)

$\boldsymbol{s} \in D$, guarantees the existence of the mean, variance, and covariance functions (see Exercise 1.2), which we define, respectively, as $\mu(\boldsymbol{s}) = \mathbb{E}[X(\boldsymbol{s})], \sigma^2(\boldsymbol{s}) = \text{var}[Z(\boldsymbol{s})]$, and $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \text{cov}[X(\boldsymbol{s}_1), X(\boldsymbol{s}_2)]$. Then, a second-order process $\{X(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is said to be weakly stationary if the mean function is constant and the covariance function between $X(\boldsymbol{s}_i)$ and $X(\boldsymbol{s}_j)$ depends on only the difference $\boldsymbol{s}_i - \boldsymbol{s}_j$, i.e.,

(i) $\mathbb{E}[X(\boldsymbol{s})] = \mu$, for all $\boldsymbol{s} \in D$.
(ii) $\text{cov}[X(\boldsymbol{s}_i), X(\boldsymbol{s}_j)] = g(\boldsymbol{s}_i - \boldsymbol{s}_j)$, for all $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$ and for some function g. More information about the function $g(\cdot)$ will be given later.

From condition (ii), we see that the variance of a weakly stationary process is also constant (does not depend on $\boldsymbol{s}$). The covariance function can then be written as

$$C(\boldsymbol{h}) = \text{cov}[X(\boldsymbol{s}), X(\boldsymbol{s} + \boldsymbol{h})].$$

It should be emphasized that if a second-order process is strictly stationary, then it is also weakly stationary. The reciprocal is not true (see Exercise 1.3). However, the normality assumption for the process guarantees that both stationary notions are equivalent.

Example 1.1 Consider the process given by Eq. (1.2) with $A \sim \mathcal{N}(0, 1)$. Clearly, $\mathbb{E}[X(s)] = 0$. Moreover,

$$\begin{aligned} \text{cov}[X(s_1), X(s_2)] &= \mathbb{E}[A^2/2]\,\mathbb{E}[\cos(\eta(s_1 - s_2)) + \cos(\eta(s_1 + s_2) + 2\phi)] \\ &= \tfrac{1}{2}\cos(\eta(s_1 - s_2)). \end{aligned}$$

Equivalently, $C(h) = \frac{1}{2}\cos(\eta h)$. Thus, $\{X(s) : s \in \mathbb{Z}\}$ is a weakly stationary process.

The second-order property required for weak stationarity is crucial. There are examples of processes that are strictly stationary for which the second-order property does not hold (see Exercise 1.4).