1

Introduction

Spatial statistics, like all branches of statistics, is the process of learning from data. Many of the questions that arise in spatial analyses are common to all areas of statistics. Namely,

i. What are the phenomena under study.
ii. What are the relevant data and how should it be collected.
iii. How should we analyze the data after it is collected.
iv. How can we draw inferences from the data collected to the phenomena under study.

The way these questions are answered depends on the type of phenomena under study. In the spatial or spatio-temporal setting, these issues are typically addressed in certain ways. We illustrate this from the following study of phosphorus measurements in shrimp ponds.

Figure 1.1 gives the locations of phosphorus measurements in a 300 m × 100 m pond in a Texas shrimp farm.

i. The phenomena under study are:
   a. Are the observed measurements sufficient to measure total phosphorus in the pond? What can be gained in precision by further sampling?
   b. What are the levels of phosphorus at unsampled locations in the pond, and how can we predict them?

Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties Michael Sherman

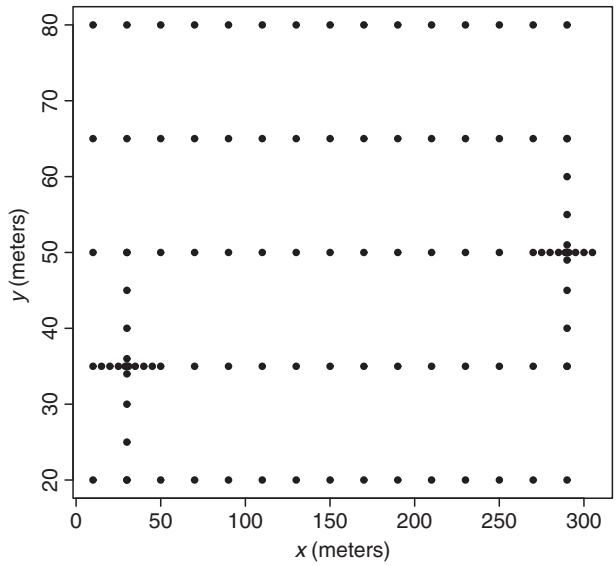

Figure 1.1 Sampling locations of phosphorus measurements.

c. How does the phosphorus level at one location relate to the amount at another location?

d. Does this relationship depend only on distance or also on direction?

ii. The relevant data that are collected are as follows: a total of $n = 103$ samples were collected from the top 10 cm of the soil from each pond by a core sampler with a 2.5 cm diameter. We see 15 equidistant samples on the long edge (300 m), and 5 equidistant samples from the short edge (100 m). Additionally, 14 samples were taken from each of the shallow and deep edges of each pond. The 14 samples were distributed in a cross shape. Two of the sides of the cross consist of samples at distances of 1, 5, 10, and 15 m from the center while the remaining two have samples at 1, 5, and 10 m from the center.

iii. The analysis of the data shows that the 14 samples in each of the two cross patterns turn out to be very important for both the analysis, (iii), and inferences, (iv), drawn from these data. This will be discussed further in Section 3.5.

iv. Inferences show that the answer to (d) helps greatly in answering question (c), which in turn helps in answering question (b) in an informative and efficient manner. Further, the answers to (b), (c), and (d) determine how well we can answer question (a). Also, we will see that increased sampling will not give much better answers to (a); while addressing (c), it is found that phosphorus levels are related but only up to a distance of about 15–20 m. The exact meaning of 'related,' and how these conclusions are reached, are discussed in the next paragraph and in Chapter 2.

We consider all observed values to be the outcome of random variables observed at the given locations. Let $\{Z(\mathbf{s}_i),\ i = 1, \ldots, n\}$ denote the random quantity Z of interest observed at locations $\mathbf{s} \in D \subset \mathbb{R}^d$, where D is the domain where observations are taken, and d is the dimension of the domain. In the phosphorus study, $Z(\mathbf{s}_i)$ denotes the log(phosphorus) measurement at the ith sampling location, $i = 1, \ldots, 103$. The dimension d is 2, and the domain D is the 300 m $\times$ 100 m pond. For usual spatial data, the dimension, d, is 2.

Sometimes the locations themselves will be considered random, but for now we consider them to be fixed by the experimenter (as they are, e.g., in the phosphorus study). A fundamental concept for addressing question (iii) in the first paragraph of the introduction is the covariance function.

For any two variables $Z(\mathbf{s})$ and $Z(\mathbf{t})$ with means $\mu(\mathbf{s})$ and $\mu(\mathbf{t})$, respectively, we define the covariance to be

$$Cov[Z(\mathbf{s}), Z(\mathbf{t})] = E[(Z(\mathbf{s}) - \mu(\mathbf{s}))(Z(\mathbf{t}) - \mu(\mathbf{t}))].$$

The correlation function is then $Cov[Z(\mathbf{s}), Z(\mathbf{t})]/(\sigma_{\mathbf{s}}\sigma_{\mathbf{t}})$, where $\sigma_{\mathbf{s}}$ and $\sigma_{\mathbf{t}}$ denote the standard deviations of the two variables. We see, for example, that if all random observations are *independent*, then the covariance and the correlation are identically zero, for all locations $\mathbf{s}$ and $\mathbf{t}$, such that $\mathbf{s} \neq \mathbf{t}$. In the special case where the mean and variances are constant, that is, $\mu(\mathbf{s}) = \mu$ and $\sigma_{\mathbf{s}} = \sigma$ for all locations $\mathbf{s}$, we have

$$Corr[Z(\mathbf{s}), Z(\mathbf{t})] = Cov[Z(\mathbf{s}), Z(\mathbf{t})]/\sigma^2.$$

The covariance function, which is very important for prediction and inference, typically needs to be estimated. Without any replication this is usually not feasible. We next give a common assumption made in order to obtain replicates.

1.1 Stationarity

A standard method of obtaining replication is through the assumption of second-order stationarity (SOS). This assumption holds that:

i. $E[Z(\mathbf{s})] = \mu$;
ii. $Cov[Z(\mathbf{s}), Z(\mathbf{t})] = Cov[Z(\mathbf{s}+\mathbf{h}), Z(\mathbf{t}+\mathbf{h})]$ for all shifts $\mathbf{h}$.

Figure 1.2 shows the locations for a particular shift vector $\mathbf{h}$. In this case we can write

$$Cov[Z(\mathbf{s}), Z(\mathbf{t})] = Cov[Z(\mathbf{0}), Z(\mathbf{t}-\mathbf{s})] =: C(\mathbf{t}-\mathbf{s}),$$

so that the covariance depends only on the spatial lag between the locations, $\mathbf{t}-\mathbf{s}$, and not on the two locations themselves. Second-order stationarity is often known as 'weak stationarity.' Strong (or strict) stationarity assumes that, for any collection of k variables, $Z(\mathbf{s}_i)$, $i = 1, \ldots, k$, and constants a_i, $i = 1, \ldots, k$, we have

$$\begin{aligned} &P[Z(\mathbf{s}_1) \le a_1, \ldots, Z(\mathbf{s}_k) \le a_k] \\ &\quad = P[Z(\mathbf{s_1}+\mathbf{h}) \le a_1, \ldots, Z(\mathbf{s_k}+\mathbf{h}) \le a_k], \end{aligned}$$

for all shift vectors $\mathbf{h}$.

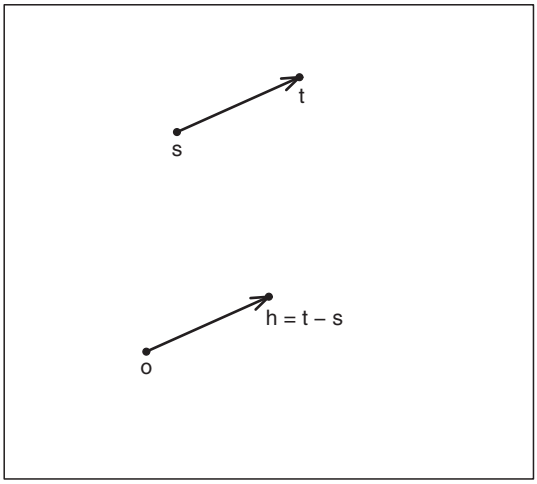

Figure 1.2 A depiction of stationarity: two identical lag vectors.

This says that the entire joint distribution of k variables is invariant under shifts. Taking $k = 1$ and $k = 2$, and observing that covariances are determined by the joint distribution, it is seen that strong stationarity implies SOS. Generally, to answer the phenomenon of interest in the phosphorus study (and many others) only the assumption of weak stationarity is necessary. Still, we will have occasions to use both concepts in what follows.

It turns out that the effects of covariance and correlation in estimation and prediction are entirely different. To illustrate this, the role of covariance in estimation and prediction is considered in the times series setting ($d = 1$). The lessons learned here are more simply derived, but are largely analogous to the situation for spatial observations, and spatio-temporal observations.

1.2 The effect of correlation in estimation and prediction

1.2.1 Estimation

Consider equally spaced observations, Z_i, representing the response variable of interest at time i. Assume that the observations come from an autoregressive time series of order one. This AR(1) model is given by

$$Z_i = \mu + \rho(Z_{i-1} - \mu) + \epsilon_i,$$

where the independent errors, ϵ_i, are such that $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \eta^2$. For the sake of simplicity, take $\mu = 0$ and $\eta^2 = 1$, and then the AR(1) model simplifies to

$$Z_i = \rho Z_{i-1} + \epsilon_i,$$

with $Var(\epsilon_i) = 1$.

For $-1 < \rho < 1$, assume that $Var(Z_i)$ is constant. Then we have $Var(Z_i) = (1 - \rho^2)^{-1}$, and thus direct calculations show that $Cov(Z_{i+1}, Z_i) = \rho/(1 - \rho^2)$. Iteration then shows that, for any time lag k, we have:

$$Cov(Z_{i+k}, Z_i) = \rho^{|k|}/(1 - \rho^2).$$

Noting that the right hand side does not depend on i, it is seen that SOS holds, and we can define $C(k) := \rho^{|k|}/(1 - \rho^2)$. Further, note that the distribution of Z_i conditional on the entire past is the same as the

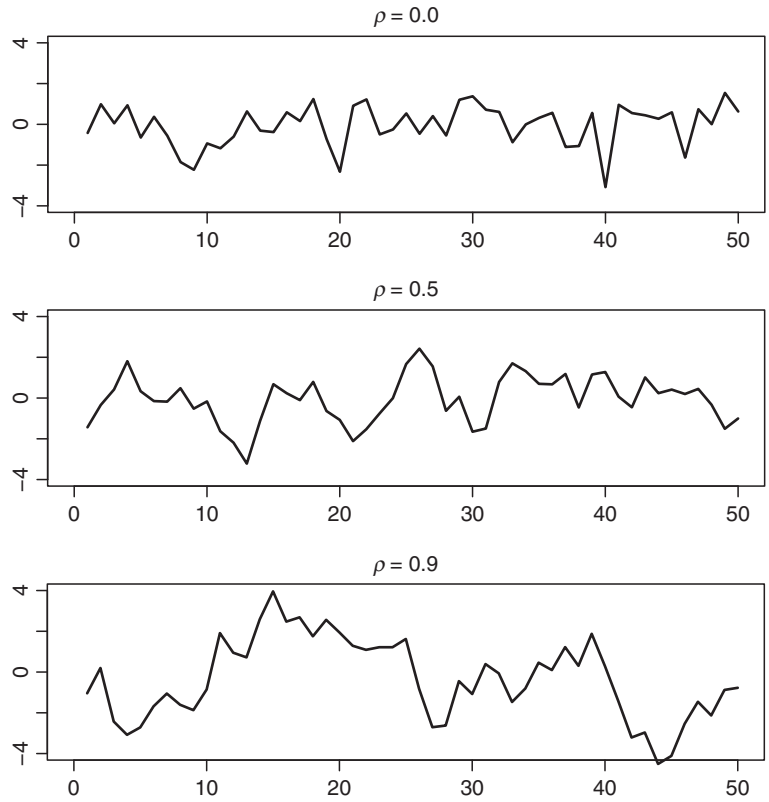

Figure 1.3 Outcomes of three AR(1) time series.

distribution of Z_i given only the immediate past, Z_{i-1}. Any such process is an example of a *Markov process*. We say that the AR(1) process is a Markov process of order one, as the present depends only on the one, immediately previous observation in time.

Figure 1.3 shows the outcomes of three AR(1) time series, the first an uncorrelated series ($\rho = 0.0$), the second with moderate correlation ($\rho = 0.5$), and the third with strong correlation ($\rho = 0.9$). Each time series consists of $n = 50$ temporal observations. Note that as ρ increases the oscillations of the time plots decrease. For example, the number of crossings of the mean ($\mu = 0$) decreases from 22 ($\rho = 0$), to 17 ($\rho = 0.5$), to 8 ($\rho = 0.9$). In other words, the 'smoothness' of the time plots increases. This notion of 'smoothness' and its importance in spatial prediction is discussed in Section 3.3.

To examine the effect of correlation on estimation, assume that SOS holds. From observations $Z_i, i = 1, \ldots, n$, we seek to estimate and draw inferences concerning the mean of the process, μ. To do this we desire a confidence interval for the unknown μ. Under SOS, it holds that each observation has the same variability, $\sigma^2 = Var(Z_i)$, and to simplify we assume that this value is known. The usual large sample 95 percent confidence interval for μ is then given by

$$\left(\overline{Z}_n - 1.96\frac{\sigma}{n^{1/2}}, \overline{Z}_n + 1.96\frac{\sigma}{n^{1/2}}\right),$$

where

$$\overline{Z}_n = \sum_{i=1}^{n} Z_i/n$$

denotes the sample mean of the observations.

We hope that the true coverage of this interval is equal to the nominal coverage of 95%. To see the true coverage of this interval, continue to assume that the data come from a (SOS) time series. Using the fact that, for any constants $a_i, i = 1, \ldots, n$, we have

$$Var\left(\sum_{i=1}^{n} a_i Z_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(Z_i, Z_j),$$

and setting $a_i = 1/n$, for all i, gives:

$$\begin{aligned} Var(\overline{Z}_n) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(Z_i, Z_j) \\ &= \frac{1}{n^2}\left[nC(0) + 2\sum_{j=1}^{n-1}(n-j)C(j)\right], \end{aligned}$$

where the second equality uses SOS and counting. To evaluate this for large n, we need the following result named after the 19th century mathematician, Leopold Kronecker:

Lemma 1.1 *Kronecker's lemma*

For a sequence of numbers $a_i, i = 1, \ldots,$ such that

$$\sum_{i=1}^{\infty} |a_i| < \infty,$$

we have that

$$\frac{\sum_{i=1}^{n} i\ a_i}{n} \to 0, \quad as\ \ n \to \infty.$$

In a direct application taking $a_i = C(i)$ in Kronecker's lemma, it is seen that:

$$(1/n)\sum_{j=1}^{n-1} jC(j) \to 0, \text{ as } n \to \infty \text{ whenever } \sum_{j=-\infty}^{\infty} |C(j)| < \infty,$$

and thus that

$$nVar(\overline{Z}_n) \to \tilde{\sigma}^2 := \sum_{j=-\infty}^{\infty} C(j),$$

whenever $\sum_{j=-\infty}^{\infty} |C(j)| < \infty$. This last condition, known as 'summable covariances' says that the variance of the mean tends to 0 at the same rate as in the case of independent observations. In particular, it holds for the AR(1) process with $-1 < \rho < 1$. The same rate of convergence does not mean, however, that the correlation has no effect in estimation.

To see the effects of correlation, note that in the case of independent observations, it holds that the variance of the standardized mean, $nVar(\overline{Z}_n) = \sigma^2 = C(0)$. It is seen that in the presence of correlation, the true variance of the standardized mean, $\tilde{\sigma}^2$, is quite different from σ^2. In particular, for the stationary AR(1) process (with $\eta = 1$), $\sigma^2 = C(0) = (1-\rho^2)^{-1}$, while arithmetic shows that $\tilde{\sigma}^2 = (1-\rho)^{-2}$, so that the ratio of the large sample variance of the mean under independence to the true variance of the mean is $R = \sigma^2/\tilde{\sigma}^2 = (1-\rho)/(1+\rho)$. In the common situation where correlation is positive, $0 < \rho < 1$, we see that ignoring correlation leads to underestimation of the correct variance.

To determine the practical effect of this, let $\Phi(\cdot)$ denote the cumulative distribution function of a standard normal variable. The coverage of the interval that ignores the correlation is given by

$$\begin{aligned} &P\left[\overline{Z}_n - 1.96\frac{\sigma}{n^{1/2}} \le \mu \le \overline{Z}_n + 1.96\frac{\sigma}{n^{1/2}}\right] \\ &\quad = P\left[-1.96(\sigma/\tilde{\sigma}) \le \frac{n^{1/2}(\overline{Z}_n - \mu)}{\tilde{\sigma}} \le 1.96(\sigma/\tilde{\sigma})\right] \\ &\quad \to \Phi(1.96R^{1/2}) - \Phi(-1.96R^{1/2}). \end{aligned}$$

We have assumed that the Central Limit theorem holds for temporary stationary observations. It does under mild moment conditions on the Z_is and on the strength of correlation. In particular, it holds for the stationary AR(1) model. Some details are given in Chapter 10.

Evaluating the approximate coverage from the last expression, we see that when $\rho = 0.2$, the ratio $R = 0.667$ and the approximate coverage of the usual nominal 95% confidence interval is 89%. When $\rho = 0.5$, $R = 0.333$ and the approximate coverage is 74%. The true coverage has begun to differ from the nominal of 95% so much that the interval is not performing at all as advertised. When $\rho = 0.9$, $R = 0.053$, and the true coverage is approximately 35%. This interval is completely unreliable. It is seen that the undercoverage becomes more severe as temporal correlation increases.

Using the correct interval, with $\tilde{\sigma}$ replacing σ, makes the interval wider, but we now obtain approximately the correct coverage. Note, however, that the estimator, $\overline{Z}_n$, is still (mean square) consistent for its target, μ, as we still have $Var(\overline{Z}_n) \to 0$, as $n \to \infty$, whenever $\sum_{j=-\infty}^{\infty} |C(j)| < \infty$.

To generalize this to the spatial setting, first note that we can write the conditional mean and the conditional variance for the temporal AR(1) model as:

$$E[Z_i | Z_j : j < i] = \mu + \rho(Z_{i-1} - \mu)$$

and

$$Var[Z_i | Z_j : j < i] = \eta^2.$$

A spatial first-order autoregressive model is a direct generalization of these two conditional moments. Specifically, conditioning on the past is replaced by conditioning on all other observations. In the temporal AR(1) case, it is assumed that the conditional distribution of the present given the past depends only on the immediate past. The spatial analogue assumes that the conditional distribution of $Z(\mathbf{s})$ depends only on the nearest neighbors of $\mathbf{s}$. Specifically, with equally spaced observations in two dimensions, assume that:

$$E[Z(\mathbf{s}) | Z(\mathbf{t}), \mathbf{t} \neq \mathbf{s}] = \mu + \gamma \sum_{d(\mathbf{s},\mathbf{t})=1} [Z(\mathbf{t}) - \mu]$$

and

$$Var[Z(\mathbf{s}) | Z(\mathbf{t}), \mathbf{t} \neq \mathbf{s}] = \eta^2.$$

Note how these two conditional moments are a natural spatial analogue to the conditional moments in the temporal AR(1) model. If the observations follow a normal distribution, then we call this spatial model a Gaussian first-order autoregressive model. The first-order Gaussian autoregressive model is an example of a (spatial) Markov process. Figure 1.4 shows sample observations from a first-order Gaussian model on a 100×100 grid with $\gamma = 0.0$ and $\gamma = 0.2$. Note how high values (and low values) tend to accumulate near each other for the $\gamma = 0.2$ data set. In particular, we find that when $\gamma = 0.0$, 2428 of the observations with a positive neighbor sum (of which there are 4891) are also positive (49.6 percent), while when $\gamma = 0.2$, we have that 3166 of the observations with a positive neighbor sum (of which there are 4813) are also positive (65.8 percent). To see the effects of spatial correlation on inference for the mean, we again compare the true variances of the mean with the variances that ignore correlation.

First we need to find the variance of the mean as a function of the strength of correlation. Analogously to the temporal case, we have that

$$nVar(\overline{Z}_n) \to \tilde{\sigma}^2 := \sum_{\mathbf{s} \in \mathbf{Z}^2} Cov[Z(\mathbf{0}), Z(\mathbf{s})]$$

as $n \to \infty$. Unfortunately, unlike in the temporal case of an AR(1), it is not a simple matter to evaluate this sum for the conditionally specified spatial model. Instead, we compute the actual finite sample variance for any given sample size n and correlation parameter γ.

Towards this end, let $\mathbf{Z}$ denote the vector of n spatial observations (in some order). Then $Var(\mathbf{Z})$ is an $n \times n$ matrix, and it can be shown using a factorization theorem of Besag (1974), that $Var(\mathbf{Z}) := \mathbf{\Sigma} = \eta^2(\mathbf{I} - \mathbf{\Gamma})^{-1}$, where $\mathbf{\Gamma}$ is an $n \times n$ matrix with elements $\gamma_{st} = \gamma$ whenever locations $\mathbf{s}$ and $\mathbf{t}$ are neighbors, that is, $d(\mathbf{s}, \mathbf{t}) = 1$. This model is discussed further in Chapter 4.

Using this and the fact that $nVar(\overline{Z}_n) = (1/n)\mathbf{1}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{1}$, for any sample size n, we can compute the variance for any value of γ (this is simply the sum of all elements in $\mathbf{\Sigma}$ divided by the sample size n). Take this value to be $\tilde{\sigma}^2$. In the time series AR(1) setting, we were able to find the stationary variance explicitly. In this spatial model this is not simply done. Nevertheless, observations from the center of the spatial field are close to the stationary distribution. From the diagonal elements of $Var(\mathbf{Z})$, for observations near the center of the field we can see the (unconditional) variance of a single observation, σ^2, for various values of γ.

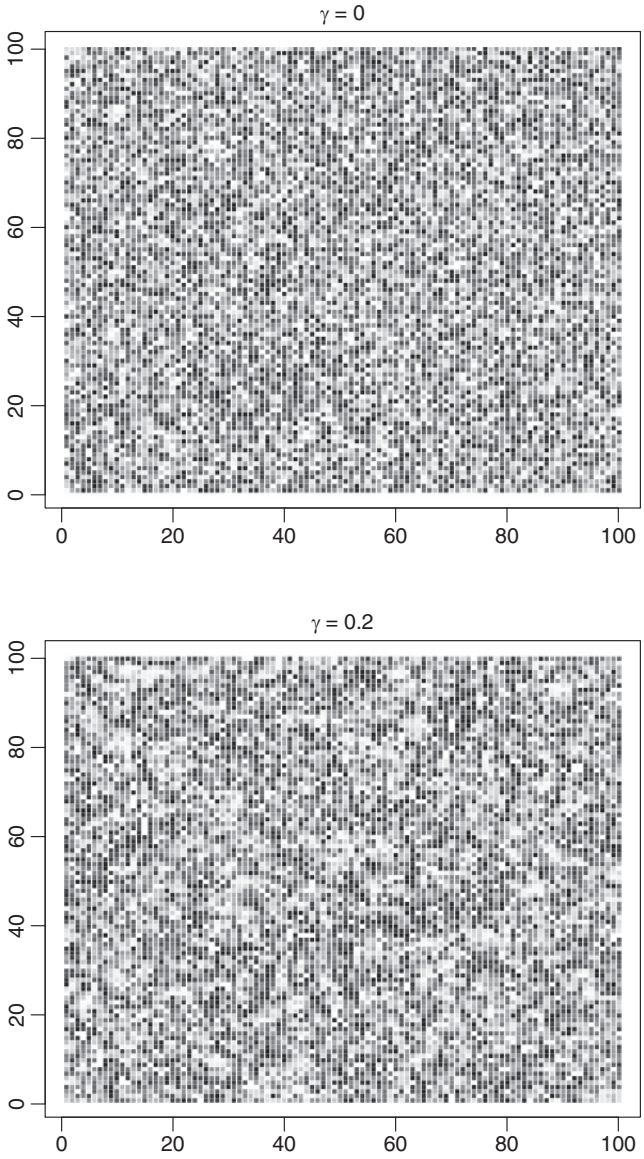

Figure 1.4 Output from two 100×100 *first-order spatial Gaussian models.*

For a 30×30 grid of observations (with $\eta^2 = 1.0$), direct calculation shows that $\tilde{\sigma}^2 = 1.24$ for $\gamma = 0.05$ ($\sigma^2 = 1.01$), $\tilde{\sigma}^2 = 1.63$ for $\gamma = 0.10$ ($\sigma^2 = 1.03$), and $\tilde{\sigma}^2 = 4.60$ for $\gamma = 0.20$ ($\sigma^2 = 1.17$). It is seen that, as in the temporal setting, the variance of the mean increases as spatial correlation increases, and that the ratio $R = \sigma^2/\tilde{\sigma}^2 = 0.813, 0.632, 0.254$ for $\gamma = 0.05, 0.10, 0.20$, respectively. This leads to approximate coverages of the usual 95% nominal confidence interval for μ of 92%, 88%, and 68%, respectively. We have seen that, as in the temporal setting, accounting for spatial correlation is necessary to obtain accurate inferences. Further, it is seen that when correlations are positive, ignoring the correlation leads to undercoverage of the incorrect intervals. This corresponds to an increased type-I error in hypothesis testing, and thus the errors are often of the most serious kind. Further, to obtain accurate inferences, we need to account for the spatial correlation and use the correct $\tilde{\sigma}^2$, or a good estimate of it, in place of the incorrect σ^2.

1.2.2 Prediction

To see the effects of correlation on prediction, consider again the temporal AR(1) process. In this situation, we observe the first n observations in time, Z_i, $i = 1, \ldots, n$, and seek to predict the unobserved Z_{n+1}. If we entirely ignore the temporal correlation, then each observation is an equally good predictor, and this leads to the predictor $\widehat{Z}_{n+1} := \overline{Z}_n$. Direct calculation shows that the true expected square prediction error for this estimator, $E[(\widehat{Z}_{n+1} - Z_{n+1})^2]$, is approximately given by

$$MSE(\widehat{Z}_{n+1}) \simeq \sigma^2 \left[1 + \frac{1}{n} \left(1 - \frac{2\rho}{n(1-\rho)} - \frac{2\rho^2}{n(1-\rho)^2} \right) \right],$$

where the error is in terms of order smaller than $1/n^2$. From this equation we see that, as $n \to \infty$, $MSE(\widehat{Z}_{n+1}) \to \sigma^2 \neq 0$. This is in stark contrast to the situation in Section 1.2.1, where the sample mean estimator has asymptotic MSE equal to 0. This is generally true in prediction. No amount of data will make the prediction error approach zero. The reason is that the future observation Z_{n+1} is random for any sample size n. Additionally, unlike in Section 1.2.1, we see that as ρ increases, $MSE(\widehat{Z}_{n+1})$ decreases. So, although strong correlation is hurtful when the goal is estimation (i.e., estimation becomes more difficult), strong correlation is helpful when the goal is prediction (prediction becomes easier).

Consider the unbiased linear estimator, $\widetilde{Z}_{n+1} = \sum_{i=1}^{n} a_i Z_i$, with $\sum_{i=1}^{n} a_i = 1$, that minimizes MSE over $a_1, \ldots, a_n$. Then, it can be shown using the methods in Chapter 2, Section 2.2, that

$$\widetilde{Z}_{n+1} = \rho Z_n + \frac{(1-\rho)^2 \sum_{i=2}^{n-1} Z_i}{n(1-\rho)+2\rho} + (1-\rho)\frac{Z_1 + Z_n}{n(1-\rho)+2\rho}.$$

Note that this predictor is approximately the weighted average of Z_n and the average of all previous time points. Also, the weight on Z_n increases as correlation increases. The methods in Section 2.2 further show that, for this predictor,

$$MSE(\widetilde{Z}_{n+1}) \simeq \sigma^2 \left[1 - \rho^2 + \frac{(1-\rho)(1+\rho)}{n}\right],$$

where the error is in terms of order smaller than $1/n$.

Imagine that we ignore the correlation and use the predictor $\widehat{Z}_{n+1}$ (i.e., we assume $\rho = 0$). Then we would approximately report

$$MSE^*(\widehat{Z}_{n+1}) = \sigma^2 \left[1 + \frac{1}{n}\right],$$

which is approximately equal to $MSE(\widehat{Z}_{n+1})$ (for large n and/or moderate ρ, the error is of order $1/n^2$), and thus $\widehat{Z}_{n+1}$ is approximately *accurate*. Accurate means that the inferences drawn using this predictor and the assumed MSE would be approximately right *for this predictor*. In particular, prediction intervals will have approximately the correct coverage for the predictand Z_{n+1}. This is in stark contrast to the estimation setting in Section 1.2.1, where ignoring the correlation led to completely inaccurate inferences (confidence intervals with coverage far from nominal). It seems that, in prediction, ignoring the correlation is not as serious as in estimation. It holds on the other hand, that

$$\frac{MSE(\widetilde{Z}_{n+1})}{MSE(\widehat{Z}_{n+1})} \simeq 1 - \rho^2.$$

This shows that $\widehat{Z}_{n+1}$ is not the *correct* predictor under correlation. Correct means that the inferences drawn using this predictor are the 'best' possible. Here 'best' means the linear unbiased predictor with minimal variance. The predictor $\widetilde{Z}_{n+1}$ is both accurate and correct for the AR(1) model with known AR(1) parameter ρ. In estimation, the estimator

$\widehat{Z}_{n+1} := \overline{Z}_n$ is approximately correct for all but extremely large $|\rho|$, but is only approximately accurate when we use the correct variance, $\tilde{\sigma}^2$.

The conclusions just drawn concerning prediction in the temporal setting are qualitatively similar in the spatial setting. Ignoring spatial correlation leads to predictions which are approximately accurate, but are not correct. The correct predictor is formed by accounting for the spatial correlation that is present. This is done using the kriging methodology discussed in Chapter 2.

In summary, it is seen that, when estimation is the goal, we need to account for correlation to draw accurate inferences. Specifically, when positive correlation is present, ignoring the correlation leads to confidence intervals which are too narrow. In other words, in hypothesis testing there is an inflated type-I error. When prediction is the goal, we can obtain approximately accurate inferences when ignoring correlations, but we need to account for the temporal or spatial correlation in order to obtain correct (i.e., efficient) predictions of unobserved variables.

We now discuss in the temporal setting a situation where ignoring correlation leads to inaccurate and surprising conclusions in the estimation setting.

1.3 Texas tidal data

A court case tried to decide a very fundamental question: where is the coastline, that is, the division between land and water. In many places of the world, most people would agree to within a few meters as to where the coastline is. However, near Port Mansfield, TX (south of Corpus Christi, TX), there is an area of approximately six miles between the intercoastal canal and a place where almost all people would agree land begins. Within this six-mile gap, it could be water or land depending on the season of the year and on the observer. To help determine a coastline it is informative to consider the history of this question.

In the 1300s, the Spanish 'Las Siete Partidas,' Law 4 of Title 28, stated that the '... sea shore is that space of ground ... covered by water in their ... highest annual swells.' This suggests that the furthest reach of the water in a typical year determines the coastline. This coastline is approximately six miles away from the intercoastal canal. In 1935, the US Supreme Court, in *Borax* v. *Los Angeles*, established MHW – 'Mean High Water' as the definition of coastal boundary. This states that the coastline is the average of the daily high-tide reaches of the water. In 1956, in

Rudder s. *Ponder*, Texas adopted MHW as the definition of coastal boundaries. This coastline is relatively close to the intercoastal canal. The two definitions of coastline do not agree in this case and we seek to understand which is more appropriate. The development here follows that in Sherman *et al.* (1997).

The hourly data in a typical year are given by Y_t, $t = 1, \ldots, 8760$, where Y_t denotes the height of the water at hour t at a station at the intercoastal canal. The horizontal projection from this height determines the coastal boundary. The regression model dictated by NOAA (National Oceanographic and Atmospheric Administration) is

$$Y_t = a_0 + \sum_{i=1}^{37} a_i \cos(\pi t S_i)/180) + \sum_{i=1}^{37} b_i \sin(\pi t S_i)/180) + \epsilon_t,$$

where a_i and b_i are amplitudes associated with S_i, the speed of the i^{th} constituent, $i = 1, \ldots, 37$, and ϵ_ts are random errors. The speeds are assumed to be known, while the amplitudes are unknown and need to be estimated. This model is similar to that in classical harmonic analysis and periodogram analysis as discussed in, for example, Hartley (1949).

The basic question in the coastal controversy is: which constituents best explain the variability in water levels? If annual or semiannual constituents explain a large proportion of the overall variability in tidal levels, this suggests that the flooded regions between the intercoastal canal and land are an important feature in the data, and suggests that the contested area cannot be called land. If, however, daily and twice-daily constituents explain most of the variability in tidal levels, then the contested area should be considered land. Note that the regression model is an example of a general linear model, and the amplitudes can be estimated using least squares estimation. In an effort to assess goodness of fit, consider the residuals from this fitted model. Figure 1.5 shows (a) the first 200 residuals, e_t, $t = 1, \ldots, 200$, and (b) residuals e_t, $t = 1001, \ldots, 1200$, from the least squares fit. One typical assumption in multiple regression is one of independent errors, that is, $Cov[\epsilon_s, \epsilon_t] = 0$ whenever $s \neq t$.

Notice that the plot of the first 200 residuals shows a stretch of approximately 60 consecutive negative residuals. This suggests that the errors are (strongly) positively correlated. The second residual plot similarly suggests a clear lack of independence in the errors, as do most stretches of residuals. From the results in estimation in Section 1.2.1, we know that ignoring the correlation would likely be a serious error if our

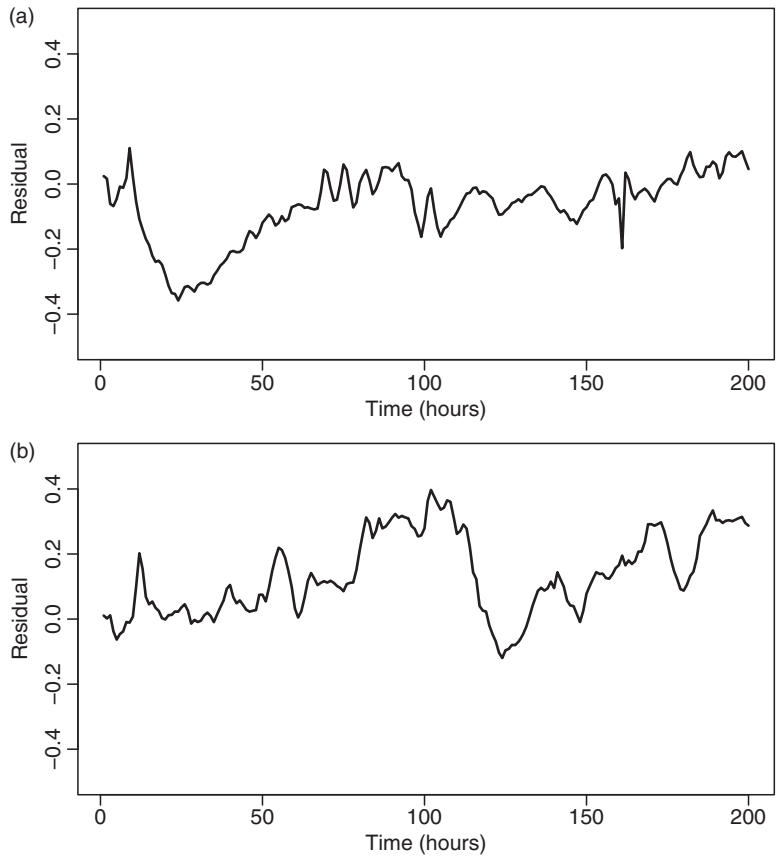

Figure 1.5 Two sets of residuals from OLS in the tidal data. (a) Residuals 1–200; (b) residuals 1001–1200.

goal is to estimate the mean of the process. The goal here, however, is to estimate the regression parameters in the harmonic analysis, and it is not clear in the regression setting what the effect of ignoring the correlation would be. To explore this, consider the setting of a simple linear regression model:

$$Y_t = \alpha + \beta x_t + \epsilon_t, \quad t = 1, \ldots, T,$$

where Y_t is the response, x_t denotes a covariate, and ϵ_t are stationary errors. The ordinary least squares (OLS) estimator of β is

$$\widehat{\beta} := \frac{\sum_{t=1}^{T}(x_t - \overline{x})Y_t}{\sum_{t=1}^{T}(x_t - \overline{x})^2},$$

with an associated variance of

$$Var(\widehat{\beta}) = \frac{\sum_{t=1}^{T}\sum_{u=1}^{T}(x_t - \overline{x})(x_u - \overline{x})E(\epsilon_t \epsilon_u)}{\left[\sum_{t=1}^{T}(x_t - \overline{x})^2\right]^2}.$$

It is seen that the variance of the estimated slope depends on the correlations between errors *and* on the structure of the design. To see the effect of the latter, consider AR(1) errors, $\epsilon_t = \rho\epsilon_{t-1} + \eta_t$ [with $Var(\eta_t) = 1.0$], under the following two designs:

Design 1: Monotone

$$x_i = i, \quad i = 1, \ldots, T$$

Design 2: Alternating

$$x_i = \begin{cases} (i+1)/2, & \text{if } i \text{ is odd;} \\ T - (i/2) + 1, & \text{if } i \text{ is even.} \end{cases}$$

To numerically compare the variance under these two scenarios, consider $T = 10$ and $\rho = 0.5$. In this case we have

$$Var(\widehat{\beta}) = \begin{cases} 0.0306 \text{ in Design 1;} \\ 0.0031 \text{ in Design 2.} \end{cases}$$

If we ignore the correlation, we then report the variance to be:

$$Var_{ind}(\widehat{\beta}) := \frac{Var(\epsilon_i)}{\sum_{t=1}^{T}(x_t - \overline{x})^2} = 0.0162$$

for *both* designs.

The conclusion is that, in contrast to the stationary case in Section 1.2.1, OLS variance estimates that ignore the positive correlation can under *or* over estimate the correct variance, depending on the structure of the design. In the tidal data, constituents of fast speed (small periods) correspond to the alternating design, while constituents of low speed (long periods) correspond to the monotone design.

Table 1.1 P-value comparison between ignoring and accounting for correlation in tidal data.

Period	Correlation	OLS
8765 (annual)	0.550	$\leq$0.001
4382 (semiannual)	$\leq$0.001	$\leq$0.001
327	0.690	0.002
25.8 (day)	$\leq$0.001	$\leq$0.001
12	$\leq$0.001	0.145

OLS are the p-values for parameter estimates that ignore correlation. Correlation p-values account for the correlation.

Table 1.1 gives the p-values for the test that the given constituent is not present in the model, based on the usual t-statistic for a few selected constituents for data from 1993. One set of p-values is computed under an assumption of independent errors, while the second set of p-values is based on standard errors which account for correlation. Variances in these cases are constructed using the block bootstrap in the regression setting.

We discuss the block bootstrap in more detail in Chapter 10. The block bootstrap appears to be a reliable tool in this case, as the residual process is well approximated by a low-level autoregressive moving-average (ARMA) process.

From the table we see that the OLS p-values that ignore correlation cannot be trusted. Further, the errors in ignoring the correlation are as predicted from the simple linear regression example. Namely that, for long periods, OLS variance estimates underestimate the correct variance, and thus lead to large t-statistics and hence p-values which are too small. For short periods, however, this is reversed and the OLS variances are too large, leading to small t-statistics and overly large p-values. The block bootstrap accounts for the temporal correlation and gives reliable variance estimates and thus reliable p-values. A careful parametric approach that estimates the correlation from within the ARMA(p, q) class of models gives results similar to those using the block bootstrap.

Finally, the semiannual period (Period = 4382) is very significant. This suggests that the flooding of the contested area is a significant feature of the data and thus this area cannot reasonably be considered as land. This outcome is qualitatively similar for other years of data as well. Although the Mean High Water criterion may be reasonable for tides in Los Angeles,

CA (on which the original Supreme Court decision was based), it does not appear to be reasonable for tides in Port Manfield, TX.

Much of the discussion in this chapter has focused on the role of correlation and how the effects of correlation are similar in the time series and spatial settings. There are, however, several fundamental differences between time series and spatial observations. Some of these will become clear as we develop spatial methodology. For now, note that in time there is a natural ordering, while this is not the case in the spatial setting. One effect of this became clear when considering the marginal variance of time series and spatial fields in Section 1.2.1. A second major difference between the time series and spatial settings is the effect of edge sites, observations on the domain boundary. For a time series of length n, there are only two observations on the boundary. For spatial observations on a $n \times n$ grid, there are approximately $4n$ observations on the boundary. The effects of this, and methods to account for a large proportion of edge sites are discussed in Section 4.2.1. A third fundamental difference is that, in time series, observations are typically equally spaced and predictions are typically made for future observations. In the spatial setting, observations are often not equally spaced and predictions are typically made for unobserved variables 'between' existing observations. The effects of this will be discussed throughout the text, but especially in Chapters 3 and 5. Other differences in increased computational burden, complicated parameter estimation, and unwieldy likelihoods in the spatial setting will be discussed, particularly in Chapters 4 and 9.