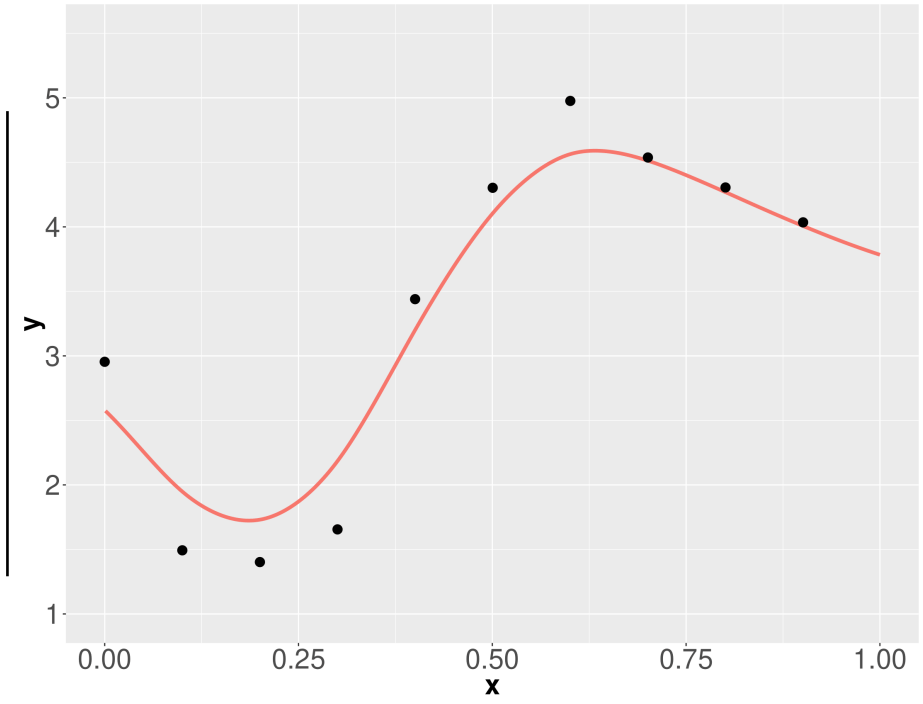
An Introduction to Kriging

Nikola Surjanovic



1 Introduction

Danie G. Krige, in his master's thesis from 1951, described an initial version of a technique that is now referred to as kriging (see Krige, 1951). The goal that he had was simple: to estimate the distribution of gold at the Witwatersrand reef complex in South Africa, using data from only several boreholes. From its inception, kriging has played an important role in geostatistics. In this context, kriging generally aims to estimate the quantity of a resource at an unobserved location, based on observed quantities at several (potentially) nearby locations. Kriging is, in essence, an interpolation method. We will see that interpolation of observed points is not necessary, however, and that we can obtain kriging prediction curves that do not pass directly through the observed data. We begin our discussion with Gausian processes in Section 2. Then, we discuss various aspects of kriging in Section 3. Several special kriging techniques are dealt with in Section 4, and Section 5 illustrates kriging on a real, two-dimensional dataset.

2 Gaussian Processes

It is possible to approach kriging from an angle that does not necessitate the use of Gaussian processes (GPs). However, we find that making use of GPs allows for certain visualizations that can serve as an aid in understanding how kriging works. Nevertheless, we will still illustrate how the fundamental kriging results can be obtained from the viewpoint of searching for a best linear unbiased estimator (BLUE), and making use of properties of the multivariate normal distribution. We begin with an informal definition of a stochastic process.

Definition 1. *A* ***stochastic process*** *is a collection of random variables,* $\{X_t : t \in T\}$*, where* T *is referred to as the index set.*

The title page is based on a template by Harish Kumar and "azetina" on StackExchange.

A more formal definition would acknowledge, among other things, that the random variables are defined on a common probability space. However, we will avoid such technical details. With this definition, we can now define a Gaussian process.

Definition 2. *A **Gaussian process** (GP) is a stochastic process that satisfies the condition that for all finite k, and for any $t_1, t_2, \ldots, t_k \in T$, $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ has a multivariate normal distribution.*

Note that for a random variable, X, that takes on values in $\mathbb{R}$, a realization of X yields a value $x \in \mathbb{R}$. However, the realization of a stochastic process can yield a collection of values in $\mathbb{R}$, for example. In the remainder of this overview, we will denote Gaussian (spatial) processes as $\{S(x) : x \in \mathbb{R}^d\}$, where $d \in \{1, 2\}$, generally, and the random variables, $S(x)$, satisfy the condition of Definition 2. For simplicity, we will sometimes refer to the process as $S(\cdot)$. It can be shown that these Gaussian spatial processes are completely specified by the mean function, $\mu : \mathbb{R}^d \to \mathbb{R}$, and the covariance function, $\gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, where

$$\mu(x) := \mathrm{E}(S(x)) \tag{1}$$

$$\gamma(x, x') := \mathrm{Cov}(S(x), S(x')) \tag{2}$$

$$\rho(x, x') := \mathrm{Corr}(S(x), S(x')) \tag{3}$$

for $x, x' \in \mathbb{R}^d$. Note that replacing the covariance function, $\gamma(x, x')$, with the correlation function, $\rho(x, x')$, would not completely specify the model. Stationary (and isotropic) Gaussian processes form a special class of GPs that have certain desirable properties.

Definition 3. *A Gaussian spatial process, $\{S(x) : x \in \mathbb{R}^d\}$, is said to be **stationary and isotropic** if:*

1. *The mean function is constant, i.e. $\mu(x) = \mu$ for all $x \in \mathbb{R}^d$, for some $\mu \in \mathbb{R}$.*

2. *The covariance between two points depends only on the (Euclidean) distance between those two points, i.e.* $\gamma(x, x') = \gamma(||x - x'||)$ *for all* $x, x' \in \mathbb{R}^d$*, for some function* $\gamma : \mathbb{R} \to \mathbb{R}$*, where* $||\cdot||$ *is the Euclidean norm.*

Similar to Diggle and Ribeiro (2007), we will use the term "stationary GP" to refer to stationary and isotropic GPs. Figure 1 provides two realizations of a stationary GP with $\mu = 3$. Note that the second realization appears to be centered around the value 2. However, $\mu = 3$ for this GP, since for any given $x_0 \in [0, 1]$, the average of many realizations of $S(x_0)$ would converge to μ. The same "long-run" analogy can be used to understand the definition of the covariance of a Gaussian process. We also note that with an appropriate choice of covariance function, the realization of each Gaussian process can be path differentiable. Differentiability is discussed in more detail in Section 3.3.

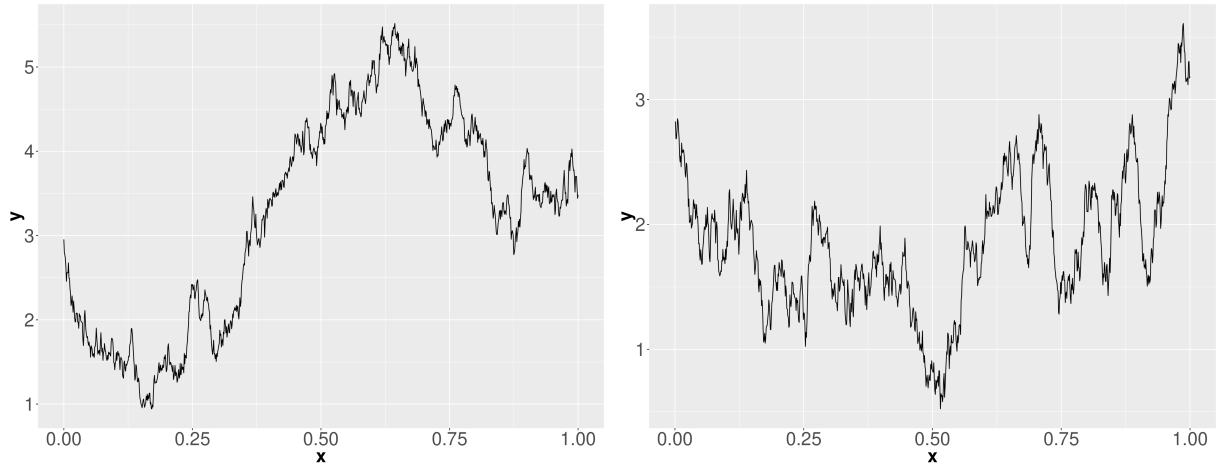
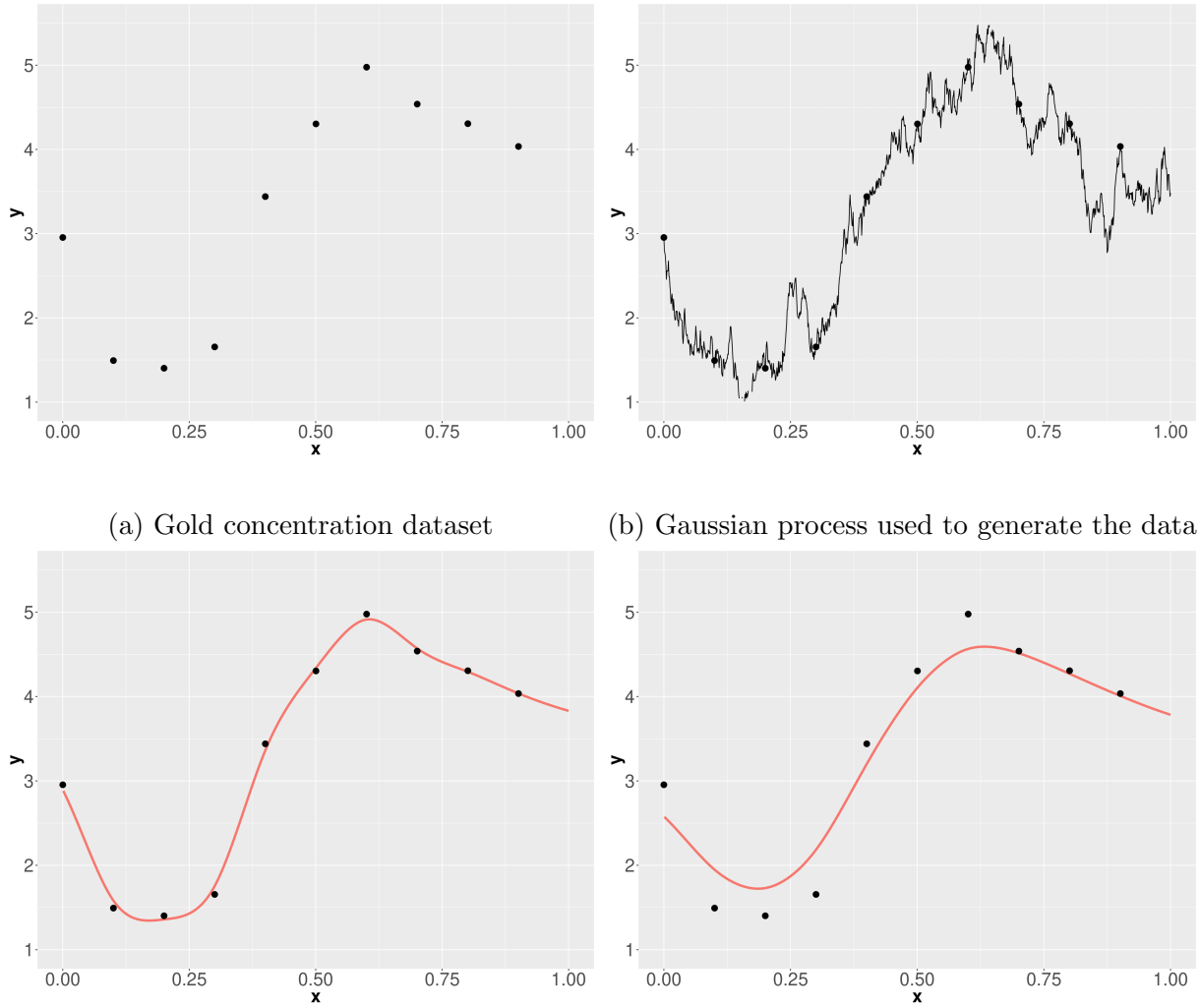


Figure 1: Two realizations of a GP with $\mu = 3$

3 Kriging

Suppose, for the moment, that we observe gold concentrations $y_1, y_2, \ldots, y_n$ at one-dimensional locations $x_1, x_2, \ldots, x_n$, with $x_i \in [0, 1]$, as in Figure 2a. A natural question to ask is: "what is the concentration of gold at an unobserved location?" Let us denote

the location of interest by x_0. In order to make a prediction of the concentration of gold at x_0, we first consider a model for the distribution of gold.



(a) Gold concentration dataset

(b) Gaussian process used to generate the data

(c) Kriging prediction curve (version 1)

(d) Kriging prediction curve (version 2)

Figure 2: Hypothetical observed gold concentration dataset

3.1 Model Assumptions

We will make use of Gaussian processes in order to model the distribution of gold in our hypothetical example from the beginning of the section. In fact, GPs are commonly used in geostatistics to model various phenomena. Diggle and Ribeiro (2007) note that these

models "rarely have any physical justification", but are instead "convenient empirical models which can capture a wide range of spatial behaviour". Specifically, we assume that

$$Y_i = S(x_i) + Z_i, \quad i = 1; \ldots, n, \tag{4}$$

where $Z_i \overset{iid}{\sim} N(0, \tau^2)$ and $S(x_i)$ is independent of Z_i. In other words, the observed gold concentrations, y_i, are a result of the realization of a GP, along with an added term, z_i, representing (realized) measurement error or spatial variation on a very small scale. The addition of the term Z_i with $\tau^2 > 0$ is sometimes referred to as the "nugget effect". This term introduces a discontinuity in the covariance function for the measurement process, $Y(\cdot)$. We can see in Figure 2b that the gold concentration dataset was generated from a realization of a Gaussian process.

3.2 Simple and Ordinary Kriging

We will start with the assumption that the underlying Gaussian process, $S(\cdot)$, is stationary (and isotropic). We also assume that we know μ and $\gamma(\cdot)$. If our goal is to predict $T = S(x_0)$, the value of the underlying process at an unobserved location, x_0, then we can construct an estimator, $\hat{T}$, that makes use of $x_1, \ldots, x_n$, and $Y_1, \ldots, Y_n$. A natural criterion for the selection of an estimator is to find one that minimizes the mean squared error (MSE). In this case, we seek to minimize the MSE between T and $\hat{T}$, when both are treated as random. That is, we seek to minimize

$$\text{MSE}(T, \hat{T}) = \text{E}\left((T - \hat{T})^2\right), \tag{5}$$

taking the expectation over both T and $\hat{T}$. It is a fundamental result that the MSE is minimized by

$$\hat{T} = \text{E}(T|Y). \tag{6}$$

For a proof, one can consult Diggle and Ribeiro (2007), for example. From equation (4), we see that $Y = (Y_1, \ldots, Y_n)^\top$ has a multivariate normal distribution. More concretely,

$$Y \sim N(\mu \mathbf{1}_n, \sigma^2 V),$$

where $\mathbf{1}_n$ is an $n \times 1$ vector of ones, and

$$\sigma^2 V = \sigma^2 R + \tau^2 I_n,$$

with I_n the $n \times n$ identity matrix, $R_{ij} = \rho(||x_i - x_j||)$, and $\sigma^2 = \gamma(0)$. Finally, since $T = S(x_0)$ is normally distributed, (T, Y) is also multivariate normal, with

$$(T, Y) \sim N(\mu \mathbf{1}_{n+1}, \Sigma),$$

where

$$\Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right] = \left[\begin{array}{c|c} \sigma^2 & \sigma^2 r^\top \\ \hline \sigma^2 r & \sigma^2 V \end{array}\right],$$

with $r^\top = (\rho(||x_0 - x_1||), \ldots, \rho(||x_0 - x_n||))$, $\rho(\cdot)$ being the correlation function. Finally, making use of properties of the multivariate normal distribution (see Johnson et al., 2002),

$$\mathrm{E}(T|Y) = \mu + \Sigma_{21}^\top \Sigma_{22}^{-1}(Y - \mu \mathbf{1}_n) = \mu + r^\top V^{-1}(Y - \mu \mathbf{1}_n). \tag{7}$$

Then, **simple kriging** refers to either the use of

$$\hat{T}_{\text{simple},1} = \mu + r' V^{-1}(Y - \mu \mathbf{1}_n), \tag{8}$$

or

$$\hat{T}_{\text{simple},2} = \hat{\mu} + r' V^{-1}(Y - \hat{\mu} \mathbf{1}_n), \tag{9}$$

with $\hat{\mu} = \bar{Y}$. On the other hand, **ordinary kriging** uses the estimate

$$\hat{T}_{\text{ordinary}} = \hat{\mu} + r'V^{-1}(Y - \hat{\mu}\mathbf{1}), \tag{10}$$

with $\hat{\mu} = (\mathbf{1}_n^\top V^{-1}\mathbf{1}_n)^{-1}\mathbf{1}_n^\top V^{-1}Y$, the generalized least squares estimate of μ. Assuming that σ^2, τ^2, and $\rho(\cdot)$ are known, then V is known and $\hat{T}_{\text{simple},1}, \hat{T}_{\text{simple},2}, \hat{T}_{\text{ordinary}}$ are all linear in Y (and are unbiased for μ). Assuming μ and V are known, we see that (7) yields a best linear unbiased estimator (BLUE) for T. As promised at the beginning of Section 2, we see that obtaining kriging estimates is possible through the use of BLUEs, and that GPs are not necessary. However, we will see in Section 3.5 that a GP approach allows for a pleasant visualization.

We make a brief comment on the interpretation of equation (7). Suppose that one would like to make a prediction at a point x_0, far away from all of the observed $x_1, \ldots, x_n$. Hopefully, $\rho(\cdot)$ has the property that $\rho(u) \to 0$ as $u \to \infty$, i.e. distant points have low correlation. In this case, we would expect $r \approx \mathbf{0}_n$, an $n \times 1$ vector of zeroes, and then (7) would reduce to μ. In other words, for distant points, our best guess is simply the mean of the underlying process, since we do not have nearby points from which we can gather information. On the other hand, in the presence of measurement error, it would make sense to make a prediction even at an *observed* location using a weighted average of nearby points. In fact, if we were to make a prediction at x_1, for example, then the first entry of r would be 1, with other entries being possibly non-zero. We can see that in (7) this would yield a weighted average of nearby points, with weights decreasing for distant points, provided that our correlation function decays.

3.3 Choosing a Correlation Function

We have assumed that we know the correlation function, $\rho(\cdot)$. If we are interested in estimating $\rho(\cdot)$, it would be nice to first try to do this using a parametric approach.

Fortunately, there are flexible families of functions that have two desirable properties:

1. $\rho(u) \to 0$ as $u \to \infty$

2. $\rho(\cdot)$ is a positive definite function

The latter property ensures that $\sum_{i=1}^{k} a_i S(x_i^*)$ has non-negative variance for all $a_i \in \mathbb{R}$ and points $x_i^* \in \mathbb{R}^d$. We will focus on the **Matérn family** of correlation functions (see Matérn, 1986). The **powered exponential family** and the **spherical family** are also other options. The spherical family has the property that $\rho(u) = 0$ for large enough u. It is also possible to generate a correlation function that is not monotonically decreasing. Unless if there is a belief that the correlation function is not monotonically decreasing, or that past a certain threshold points are completely uncorrelated, we recommend the Matérn family, which is fairly flexible. It is defined as

$$\rho_{\phi,\kappa}(u) := \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}K_{\kappa}(u/\phi), \tag{11}$$

where $K_\kappa(\cdot)$ is an order κ modified Bessel function, and $\phi, \kappa > 0$ are scale and shape parameters, respectively. In fact, the choice of κ affects the smoothness of the underlying process, $S(\cdot)$. A plot of various correlation functions from the Matérn family is displayed in Figure 3.

Once a parametric family of correlation functions has been chosen, assuming model (4), we can fairly easily use maximum likelihood estimation, described in Section 3.4, to estimate all of the relevant parameters in our model. It is a good habit to inspect the goodness-of-fit of the correlation function. One such tool is the **theoretical variogram**, defined as

$$V(x, x') := \frac{1}{2}\operatorname{Var}(S(x) - S(x')), \tag{12}$$

for two given points, x, x'. It can then be seen that $V(x, x') = \sigma^2(1 - \rho(u))$, where

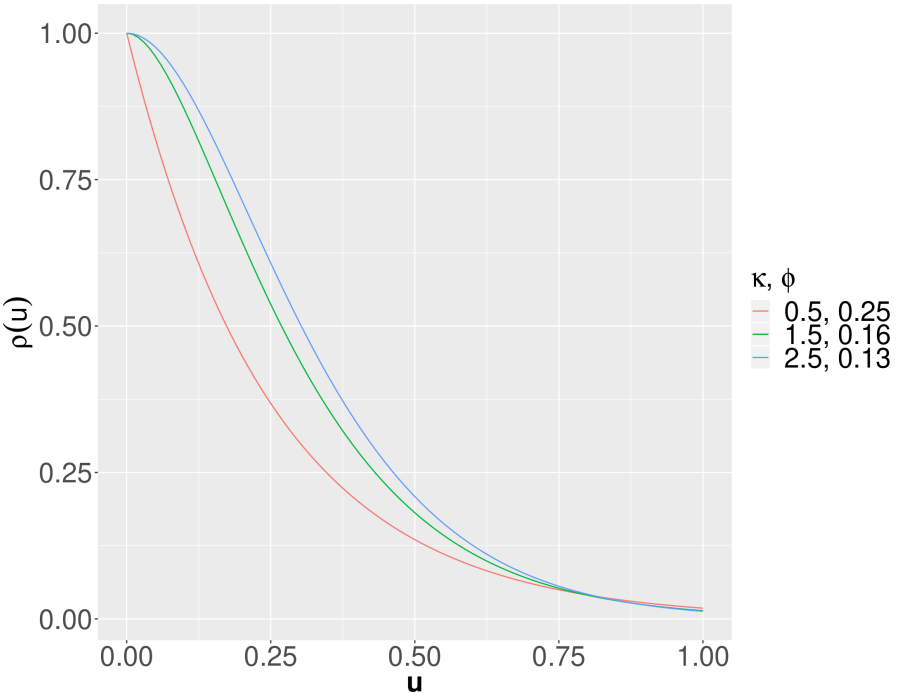


Figure 3: Several correlation functions from the Matérn family

$u = ||x - x'||$. The theoretical variogram of the Y observation process is

$$V_Y(x, x') := \frac{1}{2} \mathrm{E}\left((Y(x) - Y(x'))^2\right), \tag{13}$$

and then $V_Y(x, x') = \tau^2 + \sigma^2(1 - \rho(u))$. The **empirical variogram** for data (x_i, Y_i) is the collection of points (u_{ij}, v_{ij}), where $u_{ij} = ||x_i - x_j||$, and $v_{ij} = \frac{1}{2}(Y_i - Y_j)^2$, for $i, j = 1, \ldots, n$. Essentially, a scatterplot is obtained from which we can visually assess the adequacy of our correlation function. For more details on theoretical and empirical variograms, we refer the reader to Diggle and Ribeiro (2007).

Upon defining the Matérn correlation function family, it was mentioned that the choice of κ has an affect on the smoothness of $S(\cdot)$. There are two main notions of the differentiability of a stochastic process: path-differentiability and mean-square differentiability. The former essentially means that the "path" of each *realization* is differentiable. $S(x)$ is

said to have mean-square derivative $S'(x)$ if

$$\mathrm{E}\left(\left(\frac{S(x+h)-S(x)}{h}-S'(x)\right)^2\right)\to 0 \tag{14}$$

as $h\to 0$. Then, $S(x)$ is said to be twice mean-square differentiable if $S'(x)$ is mean-square differentiable. The idea is the same for higher derivatives. The following theorem, with more detail provided in Bartlett (1955), provides some more information on mean-square differentiability of stochastic processes.

Theorem 1. *A stochastic process that is stationary and has correlation function $\rho(u)$ is k times mean-square differentiable iff there exist $2k$ derivatives of $\rho(u)$ at $u=0$.*

With the Matérn correlation function family, $S(x)$ is $\lceil\kappa\rceil-1$ times mean-square differentiable, with $\lceil x\rceil$ denoting the smallest integer that is greater than or equal to x. As a final comment, it is important to be aware that mean-square differentiability of a stochastic process does not necessarily imply path-differentiability.

3.4 Maximum Likelihood Estimation

Assuming the model given by (4), it should be relatively straightforward to obtain maximum likelihood estimates of the model parameters. However, depending on the computer, the computation can still take a noticeable amount of time. With the Matérn correlation function family, we are interested in estimating $\sigma^2,\tau^2,\kappa,\phi$. From model (4), the log-likelihood function is given by

$$\begin{aligned} l(\sigma^2,\tau^2,\kappa,\phi) = -\frac{1}{2}\{n\log(2\pi)+\log(|\sigma^2R(\kappa,\phi)+\tau^2I_n|) \\ +(y-\mu)^\top(\sigma^2R(\kappa,\phi)+\tau^2I_n)^{-1}(y-\mu)\} \end{aligned} \tag{15}$$

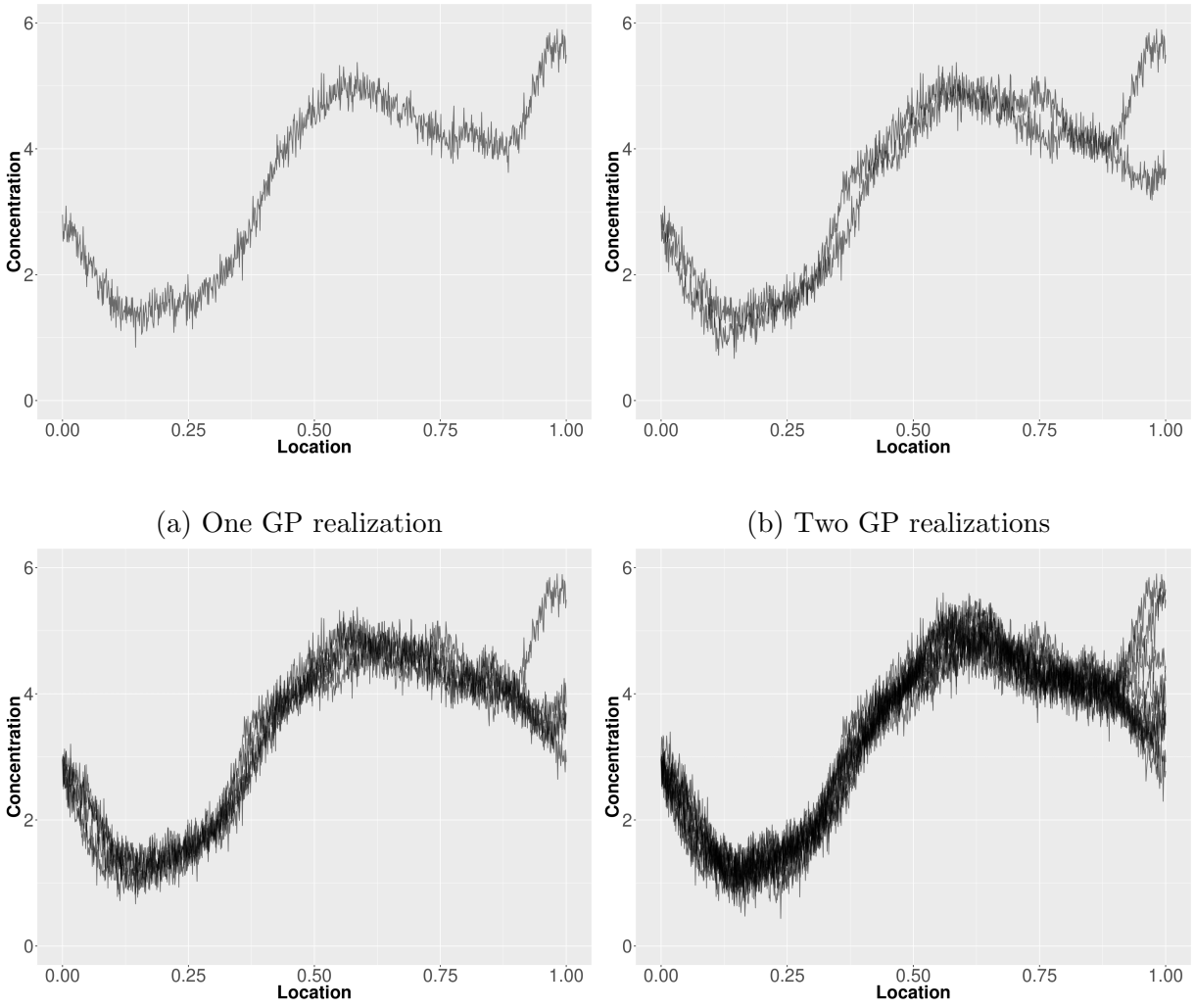
Diggle and Ribeiro (2007) comment that κ is often identified poorly. Their suggestion is to choose values of κ from a set such as $\{0.5, 1.5, 2.5\}$, thereby also simplifying the parameter space.

3.5 Visual Interpretation of Kriging

To conclude this section, we give an illustrative example of how kriging prediction curves can be interpreted as the average of realizations of a GP conditioned on Y. Consider the hypothetical gold concentration problem of Section 3 and Figure 2a. Having observed $y = (y_1, \ldots, y_n)^\top$, we would like to obtain $\hat{T} = \mathrm{E}(T|y)$, where $T = S(x_0)$, for some point x_0. Let us suppose for the moment that $\tau^2 = 0$, i.e. that there is no measurement error. Then, we can imagine taking the average of all GP realizations that pass directly through all of the observed y_i. This idea is illustrated in Figure 4. Of course, a similar analogy works for the cases in which $\tau^2 > 0$, except that the GPs do not directly interpolate the observed y_i. Instead, we can think of hypothetically taking the average of all $S(\cdot)$ and $Z = (Z_1, \ldots, Z_n)^\top$ realizations for which $s(x_i) + z_i$ interpolate the y_i. Fixing τ^2 at various values in (15), we can see the resulting kriging prediction curves in Figures 2c and 2d. The latter curve results from a larger value of τ^2, and we see that this parameter seems to act as a penalty on non-smoothness. In fact, as $\tau \to \infty$, (7) reveals that the kriging curve approaches a constant function.

4 Special Techniques

In this section we provide a brief overview of some special techniques that can be used if we are led to believe that certain assumptions of Section 3 have been violated, such as stationarity or normality, or if our goals extend beyond point prediction.



(a) One GP realization

(b) Two GP realizations

(c) Five GP realizations

(d) Ten GP realizations

Figure 4: Realizations of GPs that interpolate the y_i of the gold concentration dataset

4.1 Dealing with Trends

It is natural to believe that there might be a systematic trend in the underlying process, $S(\cdot)$. In this subsection we deal with the case where $\mu(\cdot)$ is not constant. One simple solution is to model $\mu(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$, where $x^\top = (x_1, \ldots, x_d) \in \mathbb{R}^d$. We can also add interactions or express $\mu(x)$ as a higher degree polynomial. However, we side with Diggle and Ribeiro (2007), who claim that "higher-degree surfaces should be avoided because complicated trends are better described through the stochastic component of the

model". We also advocate the use of trend surface models that *explain* response variable variation. One example might be the model $\mu(x) = \beta_0 + \beta_1 h(x)$, where $h(x)$ represents a property of interest at the location x. Then, upon choosing a model, the appropriate question to ask is how we can estimate the parameters and perform inference. Estimating the parameters should not be too complicated, since we can replace μ with $\mu(x)$ in the log-likelihood equation (15). The subtraction of the *true* trend would then yield a stationary process. We can instead subtract the *estimated* trend and then make use of the techniques for stationary processes described in previous sections. More details can be found in Cressie (1993) and Diggle and Ribeiro (2007).

4.2 Trans-Gaussian Kriging

Up until this point, we have worked with the assumption that Y is multivariate normal. However, it is easy to imagine scenarios where one might be led to believe that this assumption is violated. Fortunately, there exists a fairly powerful family of transformations that can be indexed by a single parameter, λ. We make use of the Box-Cox family of transformations (see Box and Cox, 1964):

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases}, \tag{16}$$

where operations are defined component-wise on the vector Y. We might have a reason to believe that the Gaussian model holds for $Y^{(\lambda)}$, for some value of λ. This parameter can be estimated using MLEs, using a modification of (15). The MLE of λ provides some insight into the transformation that should be used, since values of λ that are in $\{0, \pm 0.5, \pm 1, \ldots\}$ might have desirable interpretations. Care should be taken when making MSE predictions that are on the non-transformed scale, however. To achieve this, we can refer to the approaches for non-linear target predictions, as outlined in the next section.

4.3 Prediction of Complex Targets

We have focused on the prediction of quantities such as $T = S(x_0)$, for some point x_0 in the relevant space. With the reasoning of Section 3, it is straightforward to obtain predictors for quantities such as $T = \sum_{i=1}^{k} a_i S(x_i^*)$, for some $a_i \in \mathbb{R}$ and x_i^* in the relevant space. In fact, if we wish to predict a weighted average over a region $\mathcal{R} \subset \mathbb{R}^d$,

$$T_1 = \int_{\mathcal{R}} w(x) S(x)\, dx,$$

where $w(\cdot)$ is a given weight function, then, with some care, it follows that

$$\mathrm{E}(T_1|Y) = \int_{\mathcal{R}} w(x)\, \mathrm{E}(S(x)|Y)\, dx,$$

which can be computed using results of Section 3, although the integral might be intractable.

Finally, the target, T, might be a non-linear property of the underlying process, $S(\cdot)$. Letting $\mathbb{1}(\cdot)$ denote the indicator function, we might wish to predict

$$T_2 = \mathbb{1}(S(x_1^*) + S(x_2^*) > c),$$

for given points, x_1^*, x_2^*, and a threshold value, c. For example, this might be interpreted as whether or not the total amount of rainfall in two locations exceeds a certain threshold. However, in this case,

$$\mathrm{E}(T_2|Y) = \mathrm{P}(S(x_1^*) + S(x_2^*) > c \,|\, Y),$$

and it is not clear how to use the predicted surface, $\hat{S}(x)$, to obtain $\hat{T}$. One might also

try to predict

$$T_3 = \int_{\mathcal{R}} w(x) f(S(x))\, dx,$$

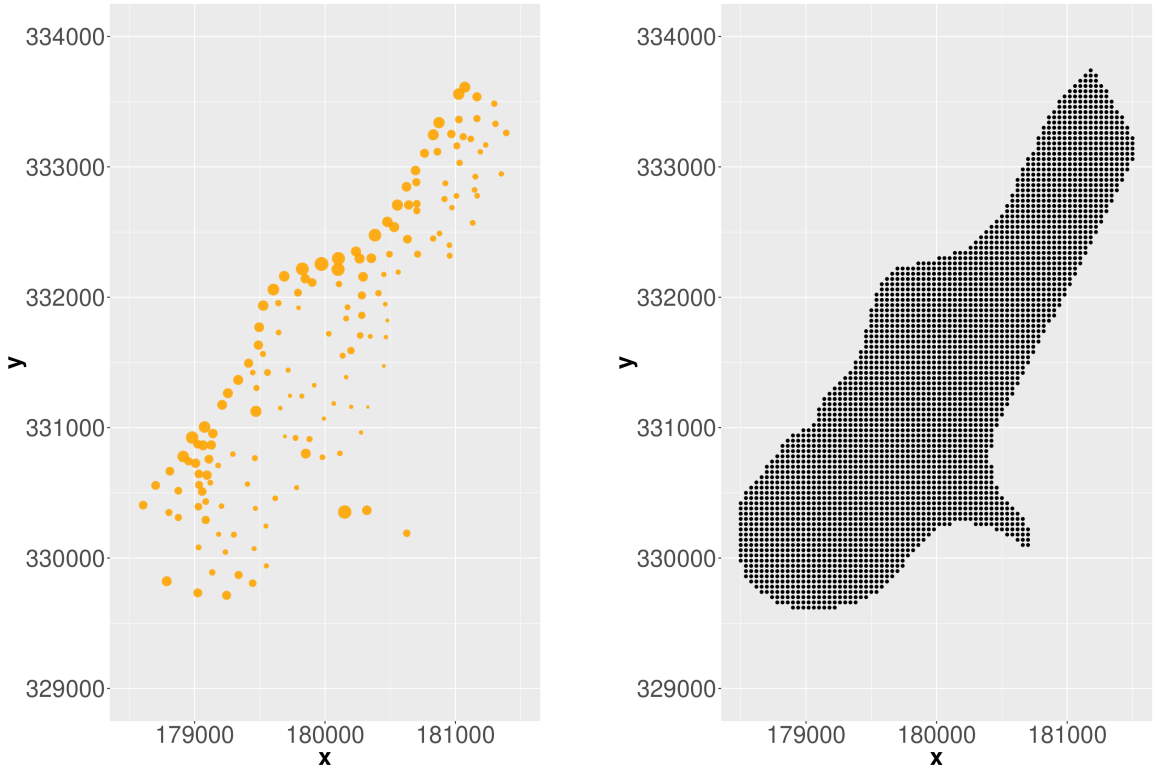
for some non-linear function $f(\cdot)$. As a solution, one can simulate realizations of $S(\cdot)$ conditioned on Y, using properties of the multivariate normal distribution, if we are assuming model (4). More specifically, we can form a grid of points, $\{x_1^*, \ldots, x_k^*\} \subset \mathcal{R}$, for a sufficiently large value of k, and draw N realizations of $S^* = \{S(x_1^*), \ldots, S(x_k^*)\}$ conditioned on Y, obtaining $S_1^*, \ldots, S_N^*$. For each S_i^*, $i = 1, \ldots, N$, we can approximate T_3 using a numerical integration method, obtaining $\hat{T}_3^{(1)}, \ldots, \hat{T}_3^{(N)}$. Then, we can use a final estimate of $\hat{T}_3 = 1/N \sum_{i=1}^{N} \hat{T}_3^{(i)}$. We would hope that, with a proper selection of the grid of points and for sufficiently large N and k, $\hat{T}_3$ would serve as a good estimate of T_3. Whether or not this is truly a good estimate depends on the selection of the grid of points and on the behaviour of the integrand to be estimated.

4.4 Relationship with Splines

Although there is much to be said about the relationship between kriging and splines, Handcock et al. (1994) comment that "a kriging estimate will be identical to a spline function if an appropriate (generalized) covariance is used". The plots of Figure 2 seem to suggest that a relationship between the two methods should exist. More details can be found in Wahba (1990).

5 Application

In this section we analyze a dataset containing the concentration of zinc in milligrams per kilogram of soil, i.e. "parts per million" or "ppm", along the Meuse river. The Meuse river is a major European river that passes through France, Belgium, and the Netherlands. The dataset to be analyzed focuses on data from the Netherlands. Figure 5 displays



(a) Observed concentrations of zinc

(b) Points of interest for prediction

Figure 5: Concentrations of zinc (ppm) along the Meuse river in the Netherlands. Circle sizes in (a) vary from 500 to 1500 ppm.

the observed concentrations of zinc, as well as the points that are of interest for spatial prediction. Using the packages "geoR", "gstat", and "sp" in **R**, a log transformation of the response is used, due to previous indication of the benefit of using this transformation on this particular dataset. The Matérn correlation function family is used, and relevant model parameters are estimated. Since the log-transformed process is assumed to be stationary, kriging predictions for this transformed process can be obtained in a straightforward way using built-in functions from the mentioned packages. Figure 6 displays the predicted zinc log-concentrations at each location. Obtaining MSE predictions on the original scale is not as straightforward as exponentiating, but is nevertheless possible, using properties of the log-normal distribution, or using techniques described in Section 4.3. For those interested

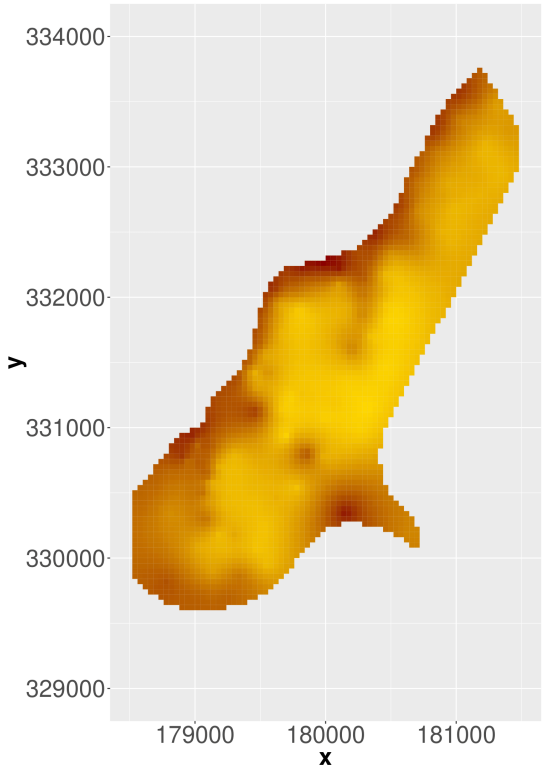


Figure 6: Kriging results for the estimated log-concentrations of zinc. Gold represents 5.0 and dark red represents 7.5 log ppm.

in the implementation, our code is a modification of the code found within the RPubs document "An Introduction to Kriging in R" at `https://rpubs.com/nabilabd/118172`. Diggle and Ribeiro (2007) also include **R** code for kriging in their book.

6 Discussion

There are still quite a few things to be said about kriging. Notably, a Bayesian approach could have been taken. Throughout, having taken a non-Bayesian approach, we have treated model parameters as fixed, plugging in parameter estimates at the end. However, in this case, our estimates for the predictive accuracy of our model might be optimistic. A Bayesian approach does not treat prediction and the estimation of parameters as two distinct problems, and can yield more conservative estimates of prediction accuracy. Al-

though we have provided an explanation of kriging through the use of Gaussian processes, defined in Section 2, we have also seen that the estimators obtained in Section 3 appear naturally from the perspective of a search for a best linear unbiased estimator. We briefly discussed how to deal with trends and transformations of the response in Section 4, when either stationarity or normality assumptions are violated. For the prediction of more complex targets, we saw how a simulation approach can serve as an aid. We also saw in this section that a relationship between kriging prediction curves and splines exists. Finally, the dataset used in Section 5 once again illustrated both the practicality and simplicity of kriging. Perhaps most importantly in this overview of kriging, we have learned that kriging is a powerful tool that offers results that can be easily interpreted by scientists from a wide range of fields.

References

Nabil A. An Introduction to Kriging in R. `https://rpubs.com/nabilabd/118172`. Accessed: 2019-11-13.

Maurice S. Bartlett. *Stochastic Processes*. Cambridge University Press, 1955.

George E.P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

Noel Cressie. *Statistics for Spatial Data, Revised Edition*. Wiley, 1993.

Peter Diggle and Paulo J. Ribeiro. *Model-based Geostatistics*. Springer, 2007.

Mark S. Handcock, Kristen Meier, and Douglas Nychka. Kriging and splines: An empirical comparison of their predictive performance in some applications: Comment. *Journal of the American Statistical Association*, 89(426):401–403, 1994.

Richard A. Johnson, Dean W. Wichern, et al. *Applied multivariate statistical analysis*, volume 5. Prentice Hall, 2002.

Danie G. Krige. A statistical approach to some mine valuations and allied problems at the Witwatersrand. Master's thesis, University of Witwatersrand, 1951.

Bertil Matérn. Spatial variation, volume 36 of Lecture Notes in Statistics, 1986.

Grace Wahba. *Spline models for observational data*, volume 59. SIAM, 1990.